Health Assistant: Multiple Disease Prediction System

Project Guide: Prof. Ankit Javeri

Project By: Devika S. Jonjale
College: Nagindas Khandwala College
Programme: TYBSc CS (hons.) - AIML
Roll No: 22

Table of Contents

Project Timeline

01

Literature Review

Research Papers,
Existing Systems

02

System Workflow

Model Architecture,
Tools and Techniques

03

Data Preprocessing

Data Acquisition,
Analysis and Preprocessing

04

Model Development

Different ML models
integration and Evaluation

05

User Interface

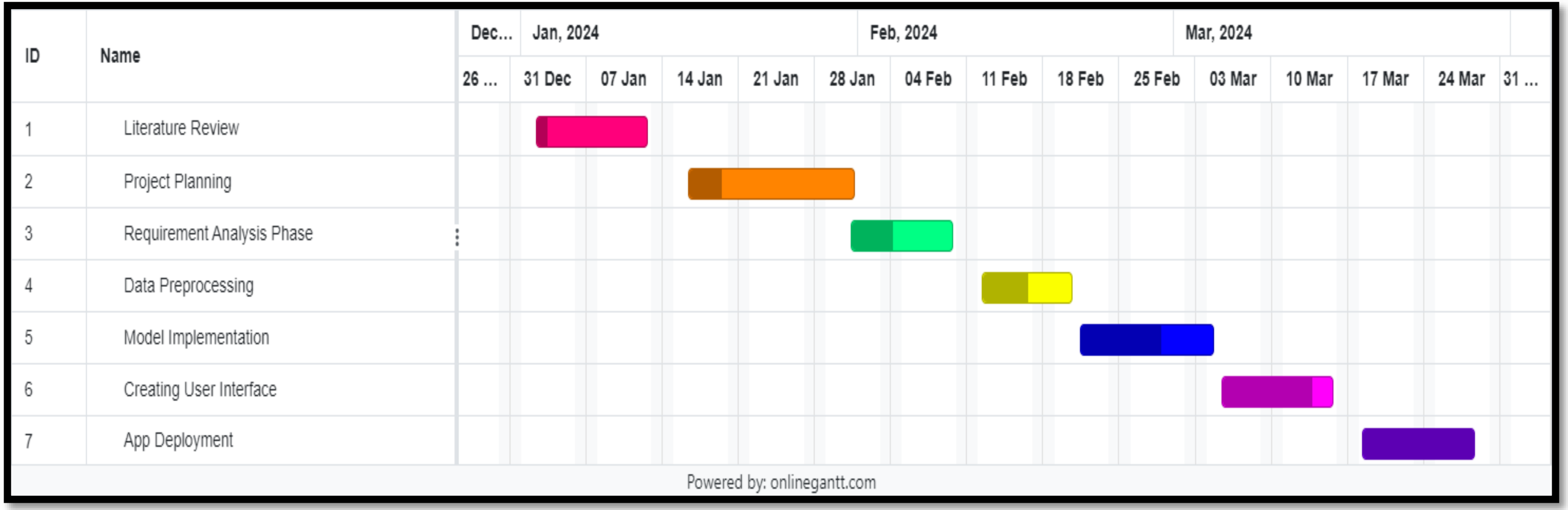
Creating User Interface
using Spyder IDE

06

Project Deployment

System Deployment using
Streamlit library

Schedule



ID
Name
Dec...
Jan, 2024
Feb, 2024
Mar, 2024
26 ...
31 Dec
07 Jan
14 Jan
21 Jan
28 Jan
04 Feb
11 Feb
18 Feb
25 Feb
03 Mar
10 Mar
17 Mar
24 Mar
31 ...
1 Literature Review
2 Project Planning
3 Requirement Analysis Phase
4 Data Preprocessing
5 Model Implementation
6 Creating User Interface
7 App Deployment
Powered by: onlinegantt.com

01

Statement

Problem definition, Project Objectives/Goals

Statement

Project Statement?

This project proposes "**Health Assistant**," a web application using machine learning to predict chronic diseases early and promote preventive healthcare.

Project Objective?

The rise of chronic diseases poses a significant global health challenge. Hence, "Health Assistant" aspires to contribute by **providing early disease detection.**

Project Goal?

To develop a **user-friendly platform for multi-disease prediction**, specifically focusing on three prevalent and concerning conditions: **Diabetes, Heart disease, and Parkinson's disease.**

02

Literature Review

Research Papers, Existing Systems

Research Papers

Diabetes

Mitushi Soni and Dr. Sunita Varma (2020): **Diabetes Prediction using Machine Learning Techniques**, this research explores various ML techniques (KNN, Logistic Regression, Decision Tree, etc.) to build models for predicting diabetes from patient data. Their findings suggest that Random Forest outperforms other techniques in achieving the most accurate predictions.

Heart Disease

Mohammed Khalid Hossen (2022): **Heart Disease Prediction Using Machine Learning Techniques**, this paper compares different ML algorithms on a dataset of patient information. Logistic regression achieved the highest accuracy (95%) compared to other algorithms tested (Support Vector Machine, KNN, Random Forest, Gradient Boosting Classifier).

Parkinson's Disease

Aditi Govindua and Sushila Palweb (2023): **Early detection of Parkinson's disease using machine learning**, this research explores using ML in telemedicine to remotely detect PD early on. By analyzing voice data from patients, they found a Random Forest machine learning model achieved the highest accuracy (91.83%) in detecting PD.

Existing Systems

Diabetes Prediction:

Risk questionnaires: There are risk questionnaire that can help you assess your risk of developing type 2 diabetes.

Heart Disease Prediction:

Cardiovascular disease risk calculators: These are online tools that use risk factors to estimate your risk of developing heart disease.

Parkinson’s Diseease Prediction:

MDS-PD risk score: The MDS-PD risk score is a tool that can be used to help identify people who are at high risk of developing Parkinson's disease.

03

Datasets

Dataset description, Tools and Techniques

Datasets:

Diabetes

- **Pima Indian females**
- **Number of Instances:** 768
- **Features:**
  - Number of times pregnant
  - Plasma glucose concentration (oral glucose tolerance test)
  - Diastolic blood pressure
  - Triceps skin fold thickness (mm)
  - Insulin level (mu U/ml)
  - Body mass index (BMI)
  - Diabetes pedigree function
  - Age (years)
- **Outcome:** 0 = negative test
  1 = positive test for diabetes

Heart Disease

- **Number of Attributes:** 76
- **Features:**
  - Age
  - Sex
  - chest pain type
  - resting blood pressure
  - serum cholesterol
  - fasting blood sugar
  - resting electrocardiographic results
  - maximum heart rate
  - exercise induced angina
  - Oldpeak
  - slope of the peak exercise ST
  - number of major vessels
  - Thal
- **Class:**0 = no disease,1 = disease

Parkinson's Disease

- **Source:** Voice recordings
- **Number of Individuals:** 31 (23 with PD, 8 healthy)
- **Number of Recordings:** 195 (around 6 recordings per person)
- **Format:** ASCII CSV
- **Features:**
  - Name
  - MDVP (14)
  - Average, Maximum, Minimum Frequency (Hz)
  - Shimmer (6)
  - NHR & HNR
  - RPDE & D2
  - DFA
  - spread1, spread2, PPE
- **Status:** 0 = healthy, 1 = diseased

Tools and Techniques

Development and Model Training:

Google Colab using Python: NumPy, Pandas, Matplotlib, Seaborn and Scikit-learn
ML Models: Linear Regression, SVM, KNN, Random Forest and XGBoost.

IDE (Integrated Development Environment):

Spyder on Anaconda: A user-friendly IDE specifically designed for scientific computing with Python, it offers features like code completion, debugging tools, and variable inspection, streamlining the development process.

Web Application Deployment:

Streamlit: A Python framework for creating web applications. It leverages existing Python code and data analysis results to build data apps with minimal coding. However, customization options are limited compared to full-fledged frameworks.

04

Methodology

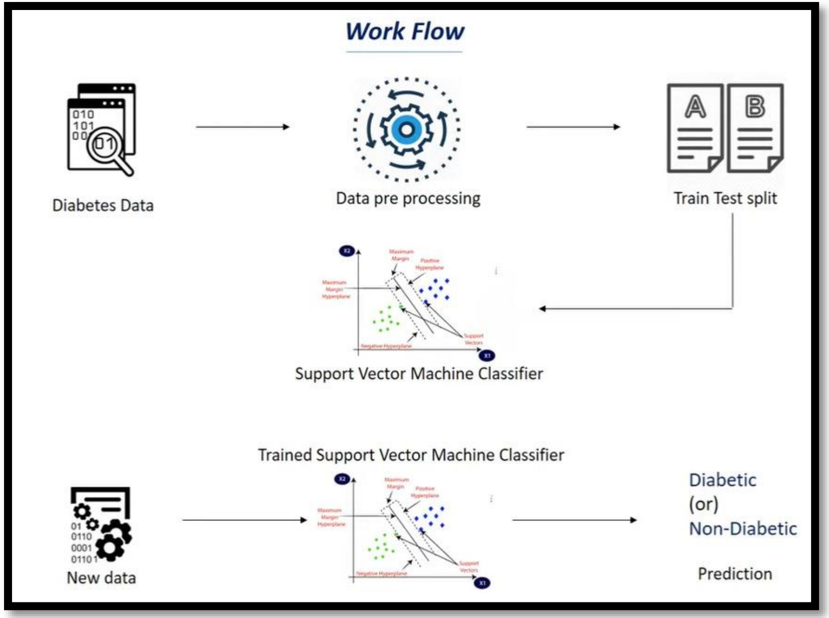
Project Workflow, Model Development

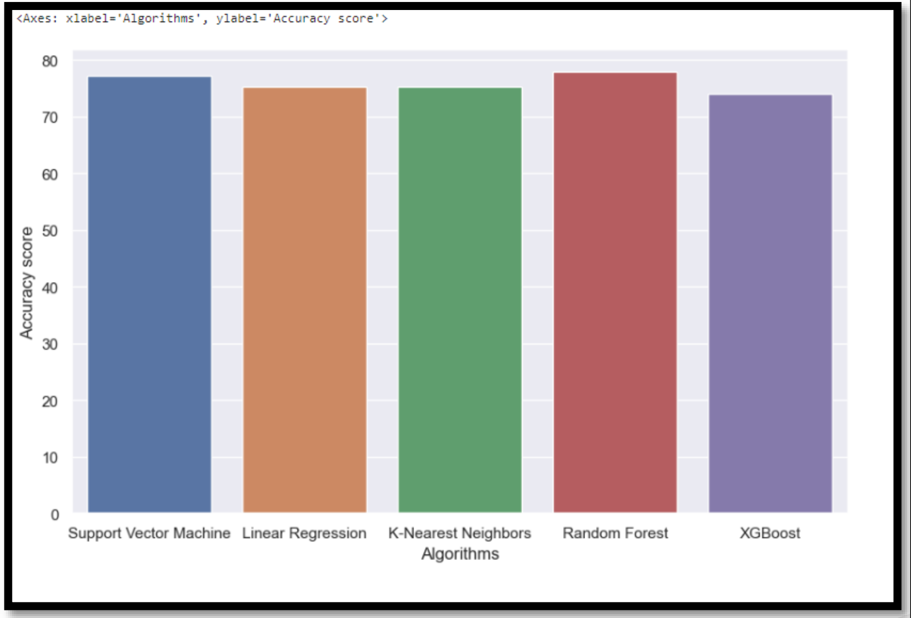
Methodology

Workflow

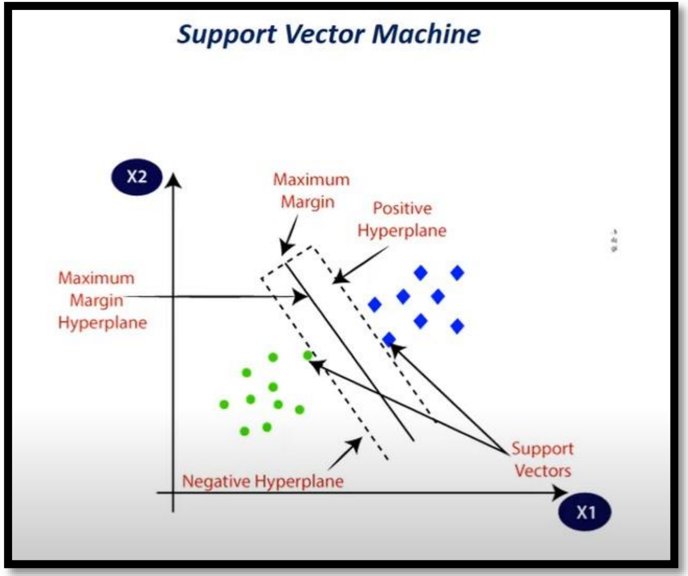
- **Data Analysis** — Dataset acquisition, data preprocessing and visualizing
- **Model Development** — Training and Testing with different ML models and evaluation metrics
- **User Interface** — Integrating the 3 systems on Spyder in Anaconda environment using Python
- **System Deployment** — Deploying the final model using Streamlit library

Diabetes Prediction Model





Support Vector Machines (SVMs) are a powerful tool used regression and classification tasks in machine learning. It creates a decision boundary that can divide N dimensional spaces into classes; this decision boundary is called a hyperplane. It selects excess points to create hyperplanes and is called support vectors. It creates multiple decision boundaries to segregate data but we need to find out the best decision boundary to classify our data The best decision boundary is known as the hyperplane, supporting vectors are the data points that are in the most proximity to the hyperplane



1] Support Vector Machine

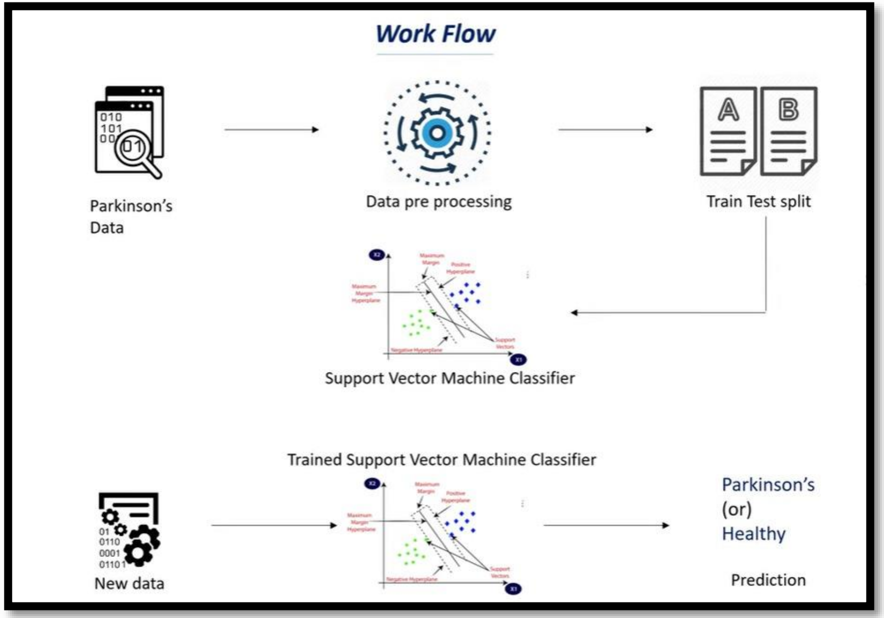
In [26]:
```
score_svm = round(accuracy_score(Y_pred_svm,Y_test)*100,2)

print("The accuracy score achieved using Linear SVM is: "+str(score_svm)+" %")
```
```
The accuracy score achieved using Linear SVM is: 77.27 %
```

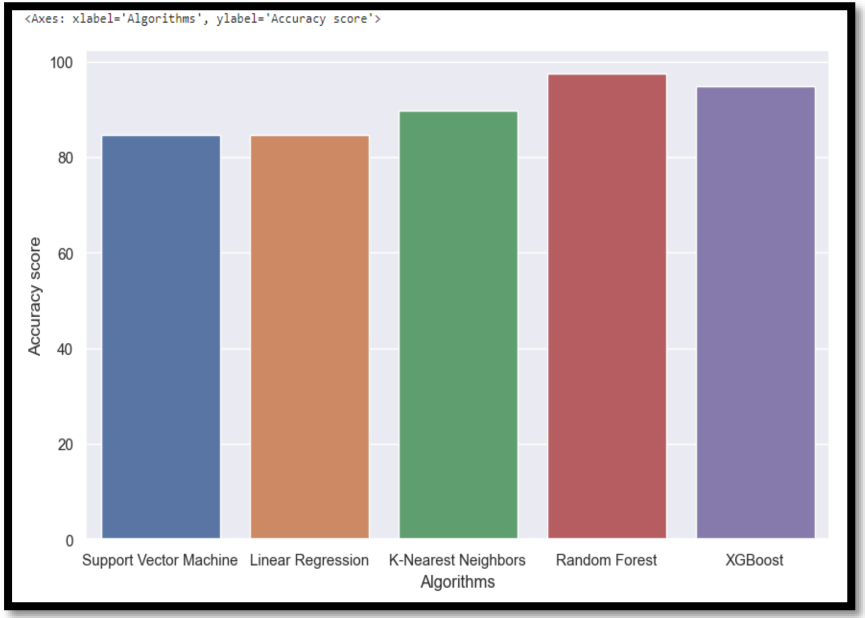
In [27]:
```
# accuracy on train data
sv_train_prediction = sv.predict(X_train)
training_data_accuracy = accuracy_score(sv_train_prediction, Y_train)
print('Accuracy on Training data : ', training_data_accuracy)
```
```
Accuracy on Training data :  0.7833876221498371
```

In [28]:
```
# accuracy on test data
sv_test_prediction = sv.predict(X_test)
test_data_accuracy = accuracy_score(sv_test_prediction, Y_test)
print('Accuracy on Test data : ', test_data_accuracy)
```
```
Accuracy on Test data :  0.7727272727272727
```

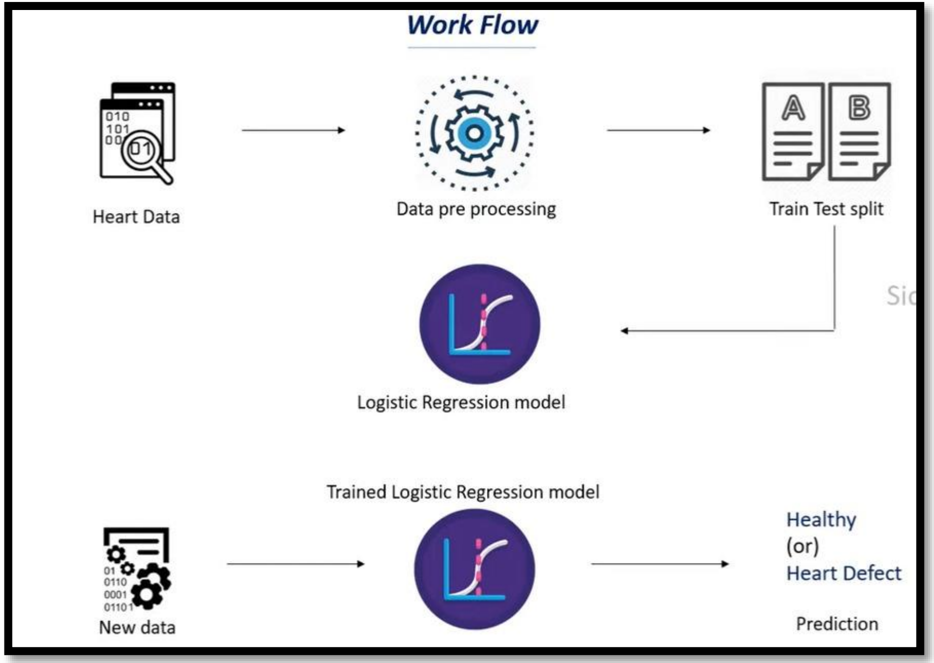
In [29]:
```
# performance evaluation metrics
print(classification_report(sv_test_prediction, Y_test))
```
```
              precision    recall  f1-score   support

           0       0.91      0.78      0.84       117
           1       0.52      0.76      0.62        37

    accuracy                           0.77       154
   macro avg       0.71      0.77      0.73       154
weighted avg       0.82      0.77      0.79       154
```

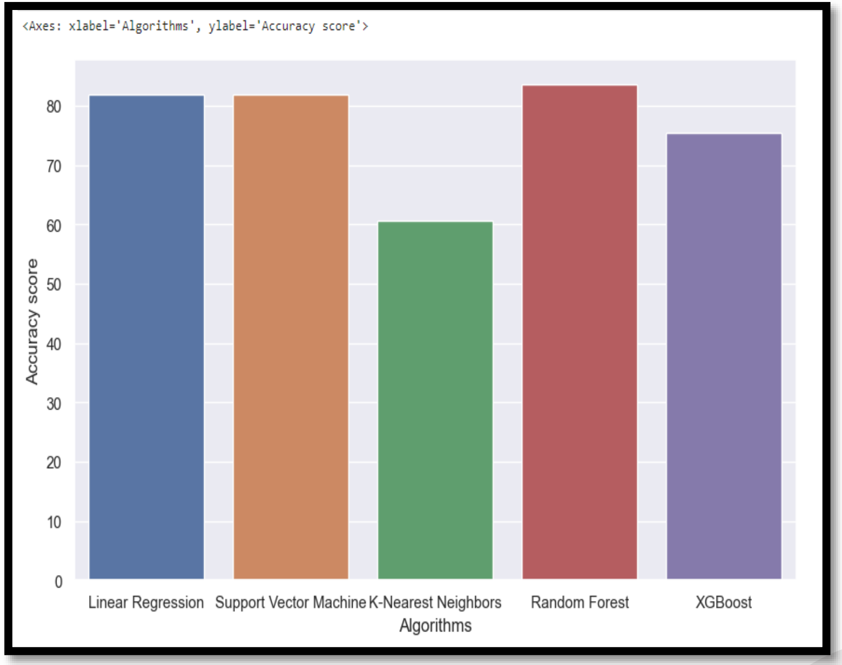
Parkinson’s Disease Prediction Model





Heart Disease Prediction Model





Logistic Regression is a supervised classification algorithm. It is a predictive analysis algorithm based on the concept of probability. It measures the relationship between the dependent variable and the one or more independent variables (risk factors) by estimating probabilities using underlying logistic function (sigmoid function). Sigmoid function is used as a cost function to limit the hypothesis of logistic regression between 0 and 1 (squashing) i.e. $0 \leq h\theta\ (x) \leq 1$

In logistic regression cost function is defined as:

$$Cost(h\theta(x),y) = \begin{cases} -\log(h\theta(x)) & if\ y = 1 \\ -\log(1 - h\theta(x)) & if\ y = 0 \end{cases}$$

1] Logistic Regression

```
In [25]: # accuracy on training data
X_train_prediction = model.predict(X_train)
training_data_accuracy = accuracy_score(X_train_prediction, Y_train)

print('Accuracy on Training data : ', training_data_accuracy)
```

```
Accuracy on Training data :  0.8512396694214877
```

```
In [26]: # accuracy on test data
X_test_prediction = model.predict(X_test)
test_data_accuracy = accuracy_score(X_test_prediction, Y_test)
print('Accuracy on Test data : ', test_data_accuracy)
```

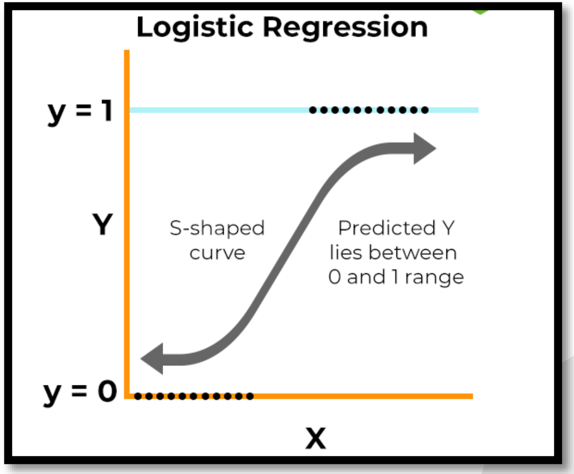
```
Accuracy on Test data :  0.819672131147541
```

```
In [27]: score_lr = round(accuracy_score(X_test_prediction, Y_test)*100,2)
print("The accuracy score achieved using Linear Regression is: "+str(score_lr)+" %")
```

```
The accuracy score achieved using Linear Regression is: 81.97 %
```

```
In [28]: # performance evaluation metrics
print(classification_report(X_test_prediction, Y_test))
```

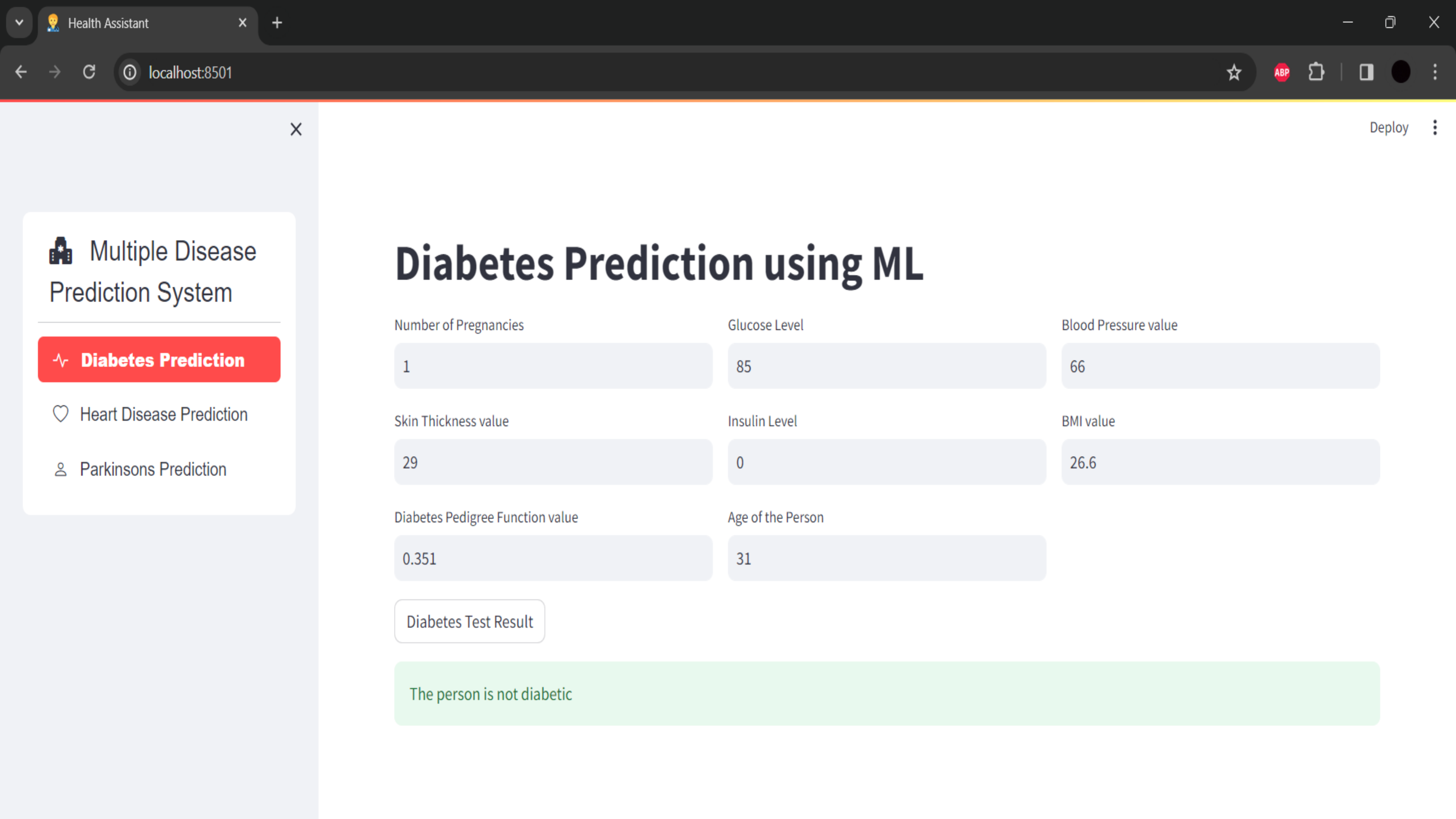
```
              precision    recall  f1-score   support

           0       0.82      0.79      0.81        29
           1       0.82      0.84      0.83        32

    accuracy                           0.82        61
   macro avg       0.82      0.82      0.82        61
weighted avg       0.82      0.82      0.82        61
```



05

Deployment

User Interface, Project Demonstration


Health Assistant
localhost:8501
Deploy
Multiple Disease Prediction System
Diabetes Prediction
Heart Disease Prediction
Parkinsons Prediction
Diabetes Prediction using ML
Number of Pregnancies
1
Glucose Level
85
Blood Pressure value
66
Skin Thickness value
29
Insulin Level
0
BMI value
26.6
Diabetes Pedigree Function value
0.351
Age of the Person
31
Diabetes Test Result
The person is not diabetic

Health Assistant

localhost:8501

Deploy

Heart Disease Prediction using ML

Multiple Disease Prediction System

- Diabetes Prediction
- **Heart Disease Prediction**
- Parkinsons Prediction

Age	Sex	Chest Pain types
63	1	3
Resting Blood Pressure	Serum Cholestoral in mg/dl	Fasting Blood Sugar > 120 mg/dl
145	233	1
Resting Electrocardiographic results	Maximum Heart Rate achieved	Exercise Induced Angina
0	150	0
ST depression induced by exercise	Slope of the peak exercise ST segment	Major vessels colored by flourosopy
2.3	0	0

thal: 0 = normal; 1 = fixed defect; 2 = reversable defect

1

Heart Disease Test Result

The person is having heart disease

Health Assistant

localhost:8501

Deploy

Multiple Disease Prediction System

- Diabetes Prediction
- Heart Disease Prediction
- **Parkinsons Prediction**

Parkinson's Disease Prediction using ML

MDVP (Hz): 197.076	MDVP (Hz): 206.896	MDVP (Hz): 192.055	MDVP (%): 0.00289	MDVP (Abs): 0.00001
MDVP: 0.00166	MDVP: 0.00168	Jitter: 0.00498	MDVP: 0.01098	MDVP (dB): 0.097
Shimmer: 0.00563	Shimmer: 0.0068	MDVP: 0.00802	Shimmer: 0.01689	NHR: 0.00339
HNR: 26.775	RPDE: 0	DFA: 0.422229	spread1: 0.741367	spread2: -7.3483
D2: 0.177551	PPE: 1.743867			

Parkinson's Test Result

The person has Parkinson's disease

06

Conclusion

Future Scope, References

Conclusion

The **Health Assistant** web application has the potential to be a valuable asset in promoting preventative healthcare and early disease detection.

- **Key Strengths:**

- **Early Disease Detection:** Machine learning models can analyze user data to predict potential risks for diabetes, heart disease, and Parkinson's disease, prompting users to seek professional medical evaluation.

- **Accessibility and Convenience:** Web-based accessibility eliminates geographical and mobility barriers, allowing users to conveniently assess their health risks.

- **Cost-Effectiveness:** Development and deployment can utilize open-source tools and cloud platforms, making it a financially sustainable solution.

Future Scope

- **Improved Model Accuracy**: Techniques like data augmentation and hyperparameter tuning can be employed to refine models and enhance prediction accuracy.
- **Integration with Wearable Devices**: Connecting with wearable devices for real-time data collection (e.g., blood pressure, heart rate) could provide more comprehensive insights.
- **AI-powered Chatbot Integration**: A chatbot assistant can guide users through the app, answer questions, and provide educational resources about preventative healthcare.
- **Integration with Electronic Health Records (EHR):** Potential future integration with EHR systems could offer a more holistic view of user health data, with user consent of course

References

[1] A. Govindu and S. P. , "Early detection of Parkinson's disease using machine learning," in Procedia Computer Science, Pune, 2023.
[2] F. S. M. A.-S. M. A.-M. A. E. W. B. M. A. and F. G. , "Enhancing Parkinson’s Disease Prediction Using Machine Learning and," Tech Science Press, 2021.
[3] K. K. S. . D. D. D. A. P. G. K. and . D. , "AN EFFECTIVE PARKINSON'S DISEASE PREDICTION USING LOGISTIC DECISION REGRESSION AND MACHINE LEARNING WITH BIG DATA," Turkish Journal of Physiotherapy and Rehabilitation, 2019.
[4] A. M. and D. V. V. , "Diabetes Prediction using Machine Learning Algorithms," in Procedia Computer Science, 2019.
[5] M. S. and D. S. V. , "Diabetes Prediction using Machine Learning Techniques," International Journal of Engineering Research & Technology (IJERT), vol. 9, no. 09, 2020.
[6] K. J. Rani, "Diabetes Prediction Using Machine Learning," International Journal of Scientific Research in Computer Science Engineering and Information Technology, vol. 6, no. 4, pp. 294-305, 2020.
[7] M. K. Hossen, "Heart Disease Prediction Using Machine Learning Techniques," American Journal of Computer Science and Technology, vol. 5, no. 3, pp. 146-154, 2022.
[8] V. R. . A. D. and M. K. R. , "Heart disease prediction using machine learning techniques: a survey," International Journal of Engineering & Technology, vol. 7, no. 2.8, pp. 684-687, 2018.

Thank You!

Do you have any questions?

devikajonjale007@gmail.com