Contents

Multiple Disease Prediction System

Presented by :- Neha Mahajan

ABSTRACT:-

- Predicts diseases like diabetes and heart disease.
- Uses Support Vector Classifier (SVC) and Logistic Regression
- Analyzes health data for predictions.
- Built using the Pickel framework for easy access.
- Users can input health data for immediate risk insights.
- Offers prevention tips after risk prediction.
- Facilitates early diagnosis and improves patient outcomes.
- Keywords:
  Multiple Disease Prediction System, Machine Learning, Support Vector Classifier (SVC), Logistic Regression, Pickel Framework, Data Preprocessing, Early Diagnosis, Health Management,

OBJECTIVES:-

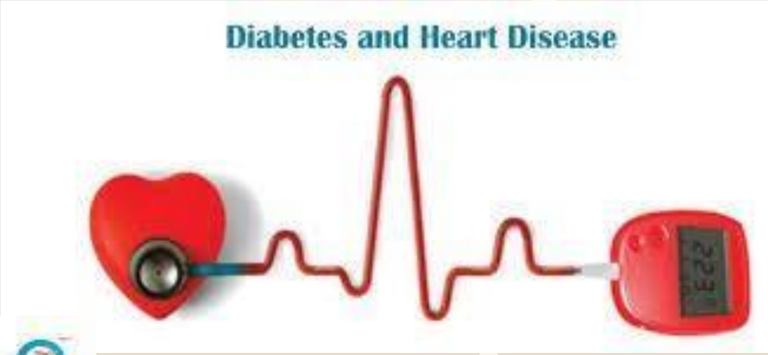
1. Accurate Disease Prediction: Develop a machine learning-based system to predict diseases like heart disease and diabetes using user-reported symptoms and health data.

2. Real-Time Insights: Utilize NLP for symptom analysis, providing users and healthcare professionals with real-time predictions to support early diagnosis.

3. User-Friendly Web Application: Create an accessible, privacy-focused web platform with an intuitive interface for easy symptom input and result visualization.

Problem Statement :-

DEVELOP A MACHINE LEARNING MODEL IN PYTHON TO PREDICT MULTIPLE DISEASES LIKE DIABETES AND HEART DISEASE



Introduction :-

- the "Disease Prediction Web Application" project tackles the important issue of getting timely diagnoses as various health conditions become more common.

- This project uses Python and machine learning to build a web app that predicts multiple diseases based on symptoms reported by users.

- By using tools like Scikit-Learn, TensorFlow, and NLTK, the app aims to provide quick diagnostic predictions and personalized advice, making health information more accessible.

- Through smart use of machine learning and natural language processing, the system evaluates user inputs to offer precise and useful health insights, speeding up the diagnosis process and improving healthcare efficiency.

- the project stands out by integrating a user-friendly interface that allows easy symptom entry and results interpretation.

Motivation

- Early Detection: By predicting the risk of heart disease and diabetes at an early stage, the system allows users to take preventative actions and seek timely medical care.

- AI-Driven Precision: The use of machine learning models such as Support Vector Classifier (SVC) and Logistic Regression ensures that predictions are based on data-driven insights. This increases the accuracy and reliability of the system

- User-Friendly Interface: Streamlit-based platform for easy health data input and instant results.This accessibility makes it easier for non-technical users to monitor their health and take proactive measures.

- Health Awareness: Provides prevention tips, promoting proactive health management.

Proposed Methodology :- Logistic Regression Model (for Heart)

- Overview of the Logistic Regression Model:-

Logistic regression is a supervised learning algorithm used for binary classification problems. It is a statistical model that estimates probabilities using a logistic function, also known as the sigmoid function. The logistic function maps any real-valued number to a value between 0 and 1, which can be interpreted as a probability.

- Logistic Regression Operations and Workings:-

1. Data Preparation: The dataset is prepared by splitting it into features (X) and the target variable (y).
2. Model Training: The logistic regression model is trained using the training data. The model learns the relationship between the features and the target variable.
3. Logistic Function: The logistic function is used to estimate the probability of the target variable being 1 (positive class) given the features.
4. Optimization: The model's parameters are optimized using an optimization algorithm, such as gradient descent.
5. Prediction: The trained model is used to make predictions on new, unseen data.

- Uses

1. Handle Binary Classification: Heart disease prediction is a binary classification problem, where the target variable is either 0 (no heart disease) or 1 (heart disease).
2. Estimate Probabilities: Logistic regression can estimate the probability of heart disease given the features, which can be useful for risk assessment.
3. Handle Non-Linear Relationships: Logistic regression can handle non-linear relationships between the features and the target variable, Interpret Results: The results of logistic regression can be easily interpreted, which can be useful for clinicians and researchers

SVS (Support Vector Classifier) :- Diabetes

Overview of the Support Vector Classifier (SVC)

Support Vector Classifier (SVC) is a supervised learning algorithm used for binary classification problems. It works by finding the optimal hyperplane that separates the data into different classes. It's especially effective in high-dimensional spaces and can handle complex, non-linear data.

SVC Operations:

-Data Preparation: The dataset is split into features (X) and the target variable (y), representing whether a person has diabetes (1) or not (0).

-Model Training: SVC learns to separate the classes by finding the best hyperplane that divides the data into diabetic and non-diabetic groups.

Kernel Function: Different kernel functions (linear, RBF, etc.) are applied to transform the data, enabling better separation of non-linear relationships.

-Optimization: The model maximizes the margin between the closest data points (support vectors) and the hyperplane, ensuring optimal classification.

-Prediction: After training, SVC predicts whether new data corresponds to diabetes or not.

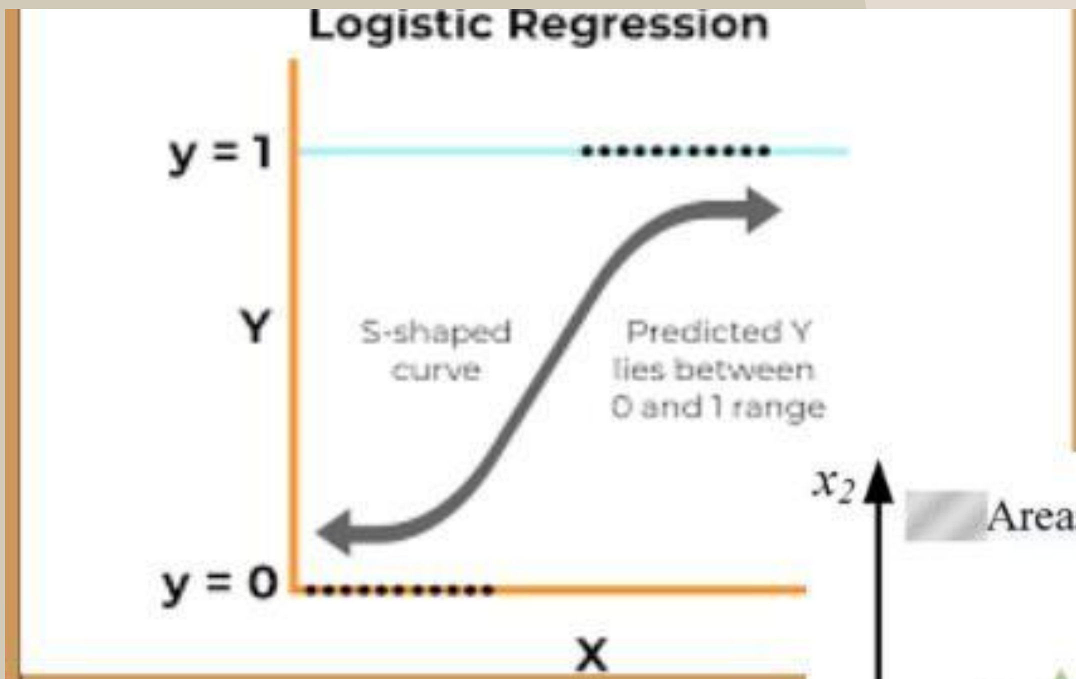
Uses of SVC:

-Binary Classification: Ideal for predicting diabetes (0 or 1).

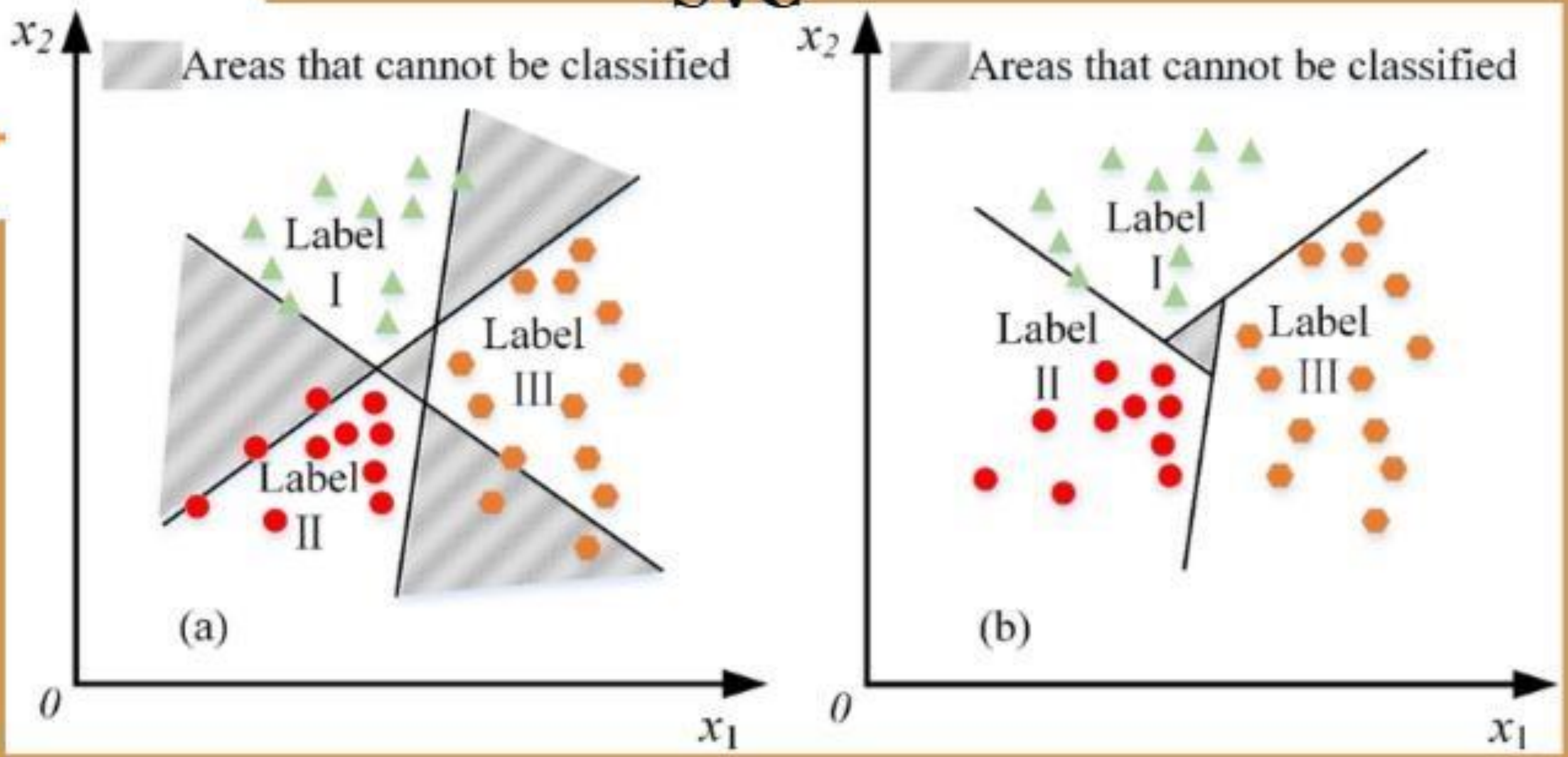
-Handles High Dimensions: Works well with health data features like glucose, BMI, etc.

-Non-Linear Boundaries: Can model complex relationships using kernels like RBF.

-Robust Generalization: Effective with unseen data, important for real-world diabetes prediction.


Logistic Regression
y = 1
Y
S-shaped curve
Predicted Y lies between 0 and 1 range
y = 0
X

SVC

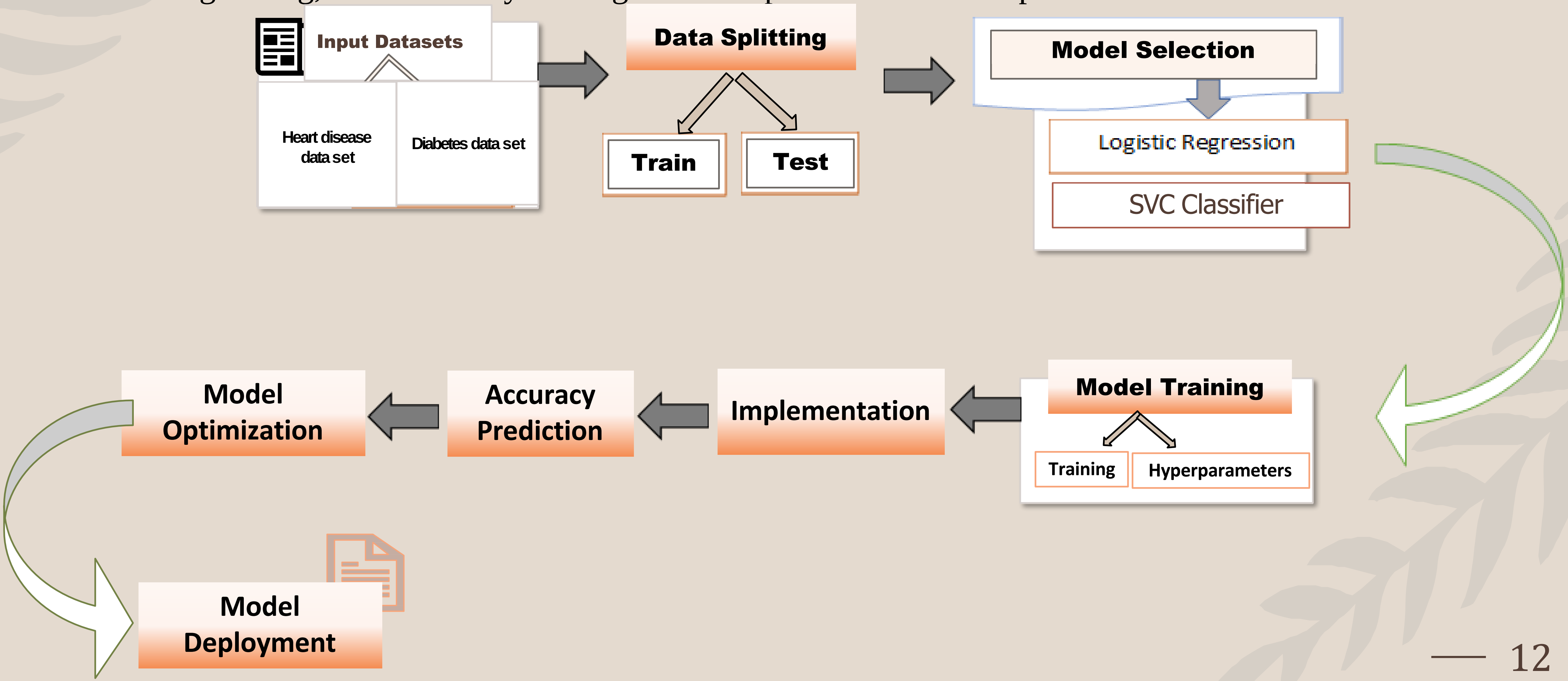

x_2
Areas that cannot be classified
Label I
Label II
Label III
(a)
0
x_1
x_2
Areas that cannot be classified
Label I
Label II
Label III
(b)
0
x_1

Algorithm

Step 1:- Start.
Step 2:- Collect health-related datasets for diseases like diabetes, and heart disease from sources such as Kaggle.
Step 3:- Clean the data by handling missing values, normalizing numerical features, and encoding categorical variables for consistent analysis.
Step 4:- Identify and select important features from the dataset that significantly influence disease predictions, improving model performance.
Step 5:- Divide the dataset into training and testing sets, typically in an 80/20 ratio, to validate the model's performance on new data.
Step 6:- Choose machine learning algorithms such as Support Vector Classifier (SVC) and Logistic Regression for disease classification.
Step 7:- Train the selected machine learning models using the training dataset to allow the model to learn patterns and relationships in the data.
Step 8:- Evaluate the performance of the models on the testing set using metrics like accuracy, precision, recall, and F1 score to ensure robustness.
Step 9:- Build a front-end interface using the Streamlit framework, enabling users to input health data for disease predictions.
Step 10:- Capture user inputs like age, glucose levels, and blood pressure through the Streamlit interface for processing and prediction.
Step 11:- Process the user input, feed it into the trained models, and predict the probability of diseases such as diabetes, and heart disease.
Step 12:- Display the predicted disease likelihood and personalized health insights to the user on the interface.
Step 13:- End

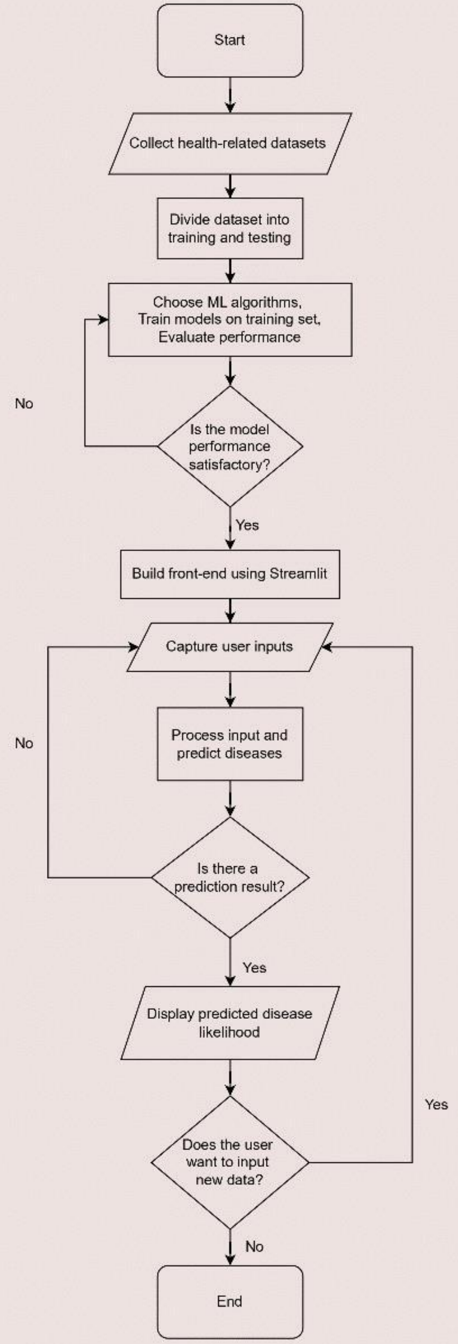
System Architecture: -

This flow demonstrates how the project processes data from raw collection through cleaning, feature extraction, model training/testing, and ultimately making real-time predictions with explanations for users.

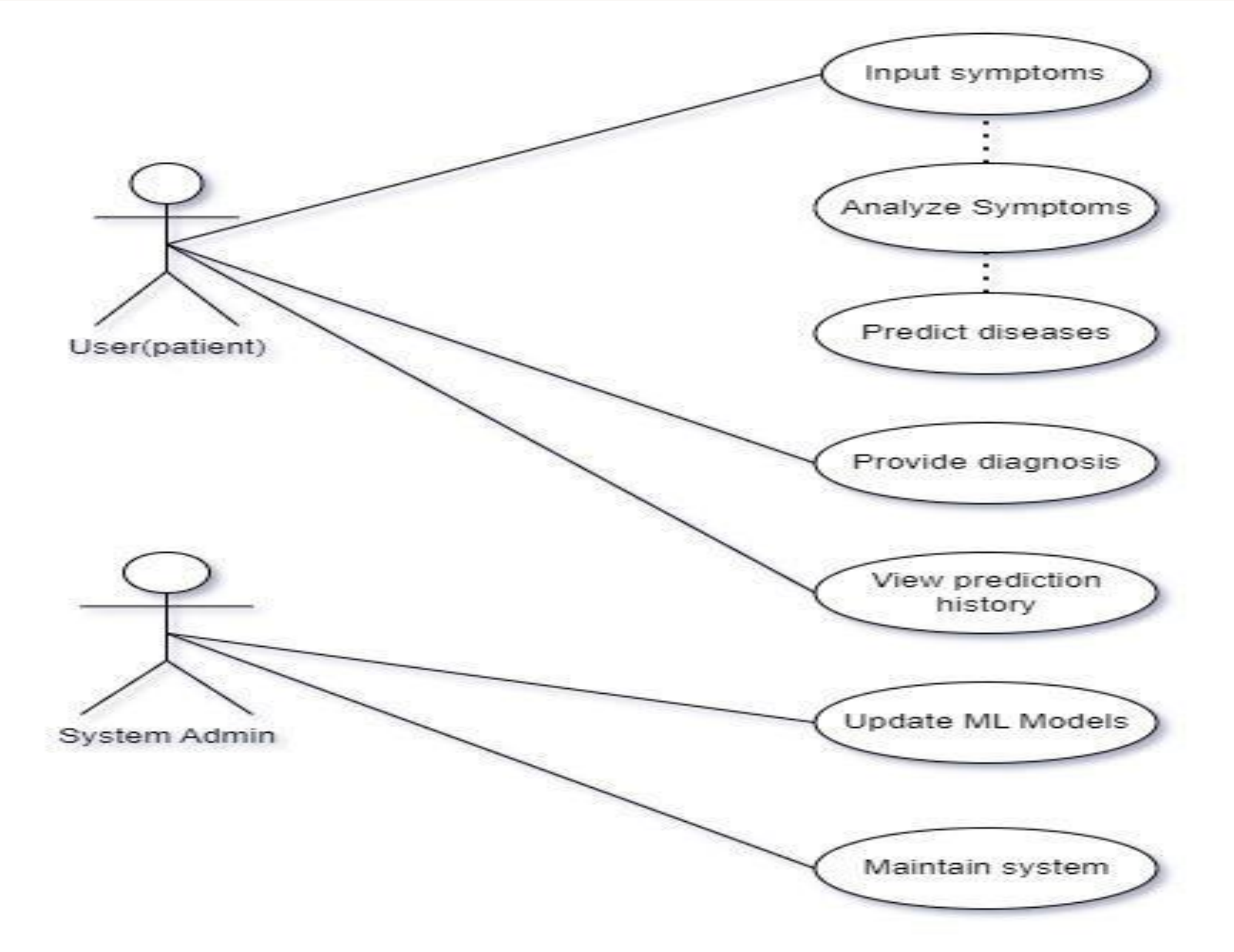


FLOWCHART:-

Dr. D. Y. PATIL VIDYAPEETH, PUNE
(DEEMED UNIVRSITY)



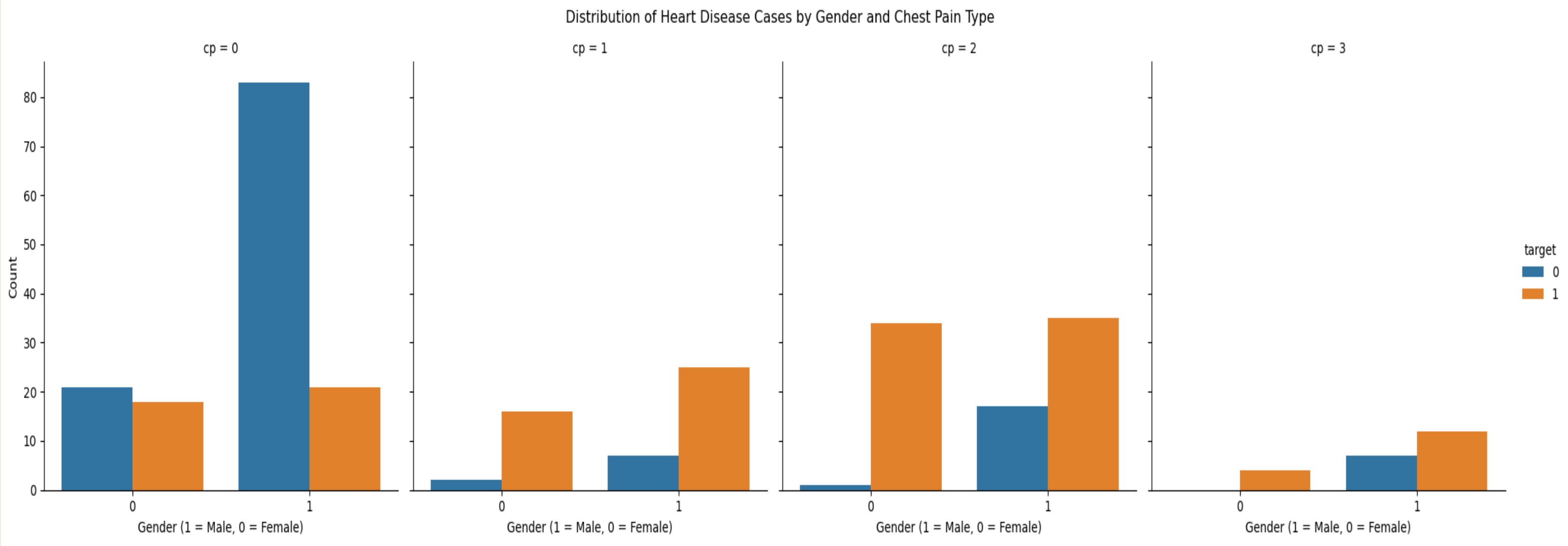
USE CASE DIAGRAM:-

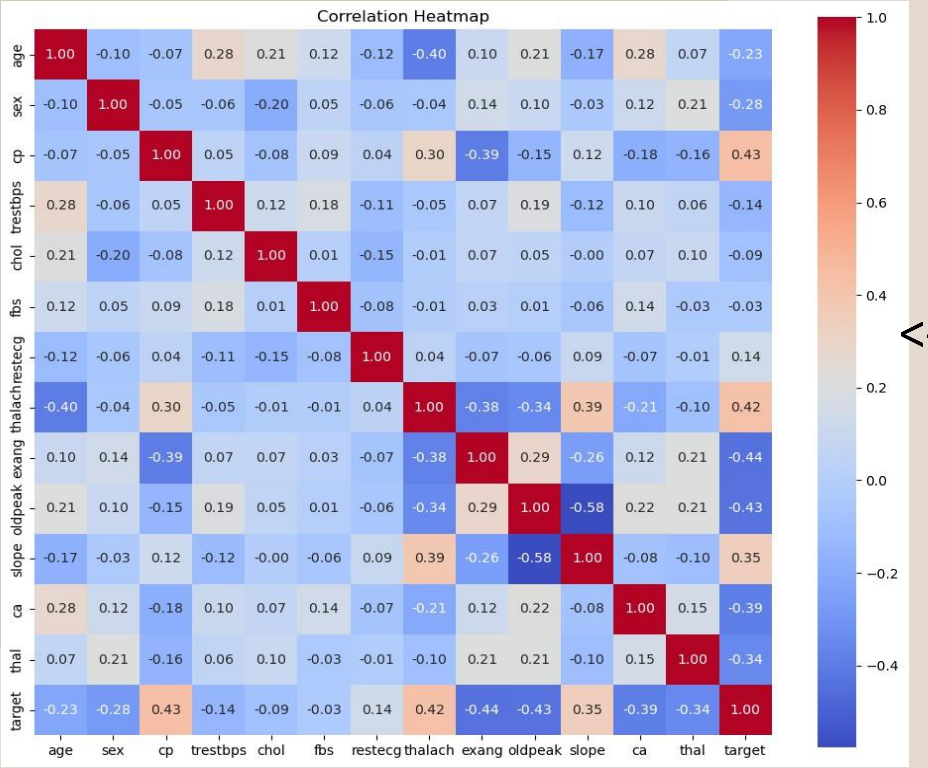


Result and Discussion:-

Logistic Regression(Heart disease prediction)

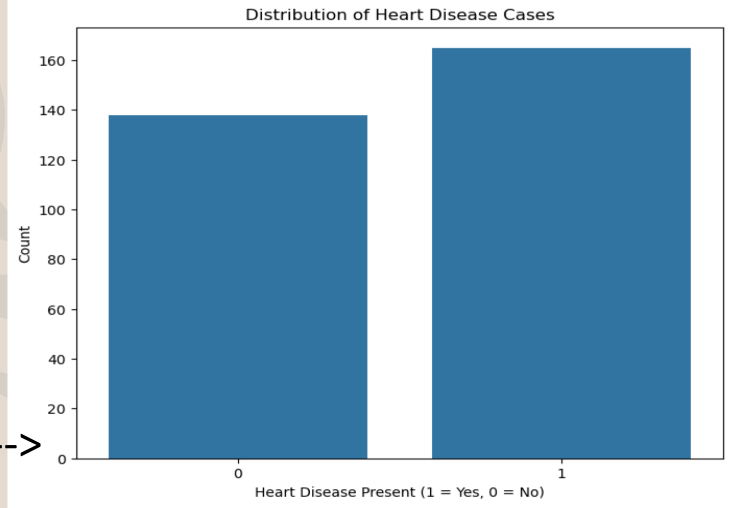
Catplot to visualize multiple categorical variables -->





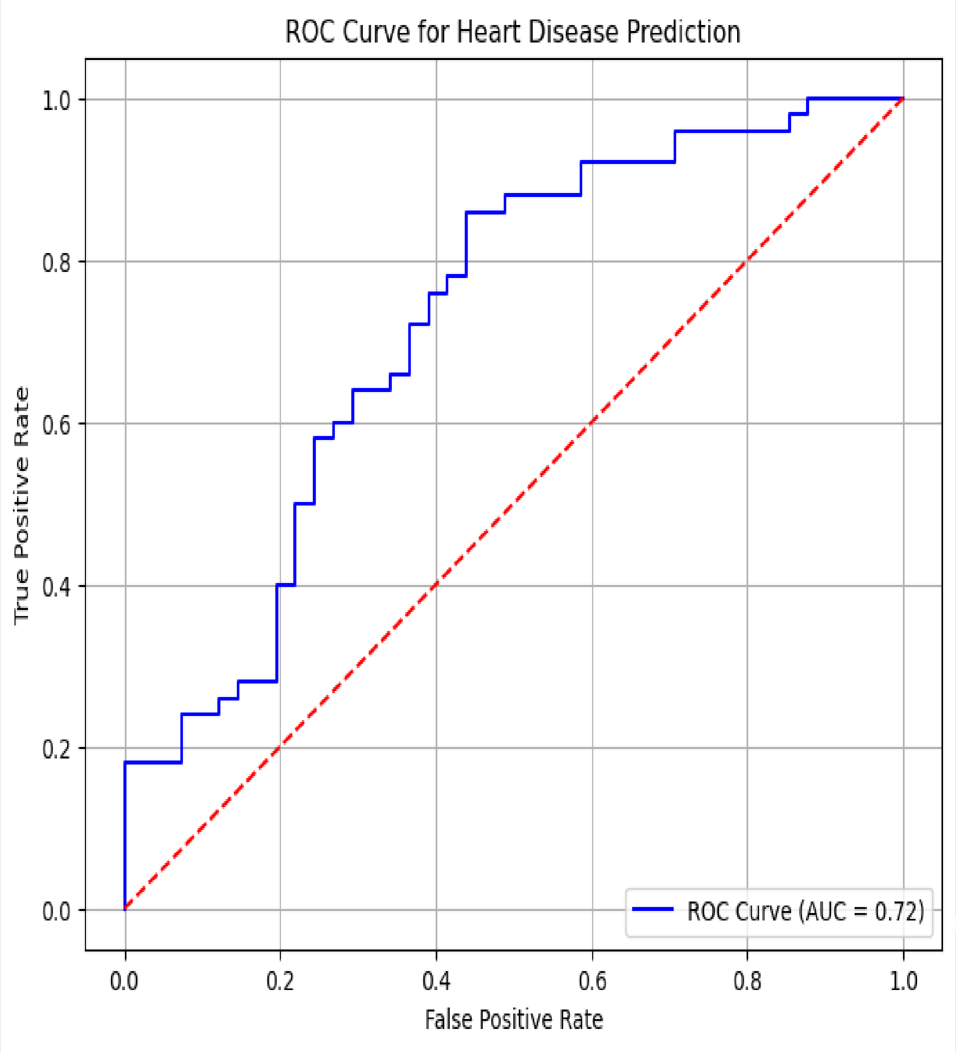
<--Correlation Heatmap

Count Plot
(Heart Disease cases)-->



ROC Curve (Receiver Operating Characteristics)---->

-Performance Assessment: ROC curves visualize logistic regression model performance for heart disease prediction.
-AUC Metric: AUC(Area Under the Curve.) measures model discrimination, with values closer to 1 indicating better accuracy.
-Threshold Optimization: ROC analysis helps choose the best classification threshold to balance sensitivity and specificity



Accuracy :- Logistic regression model :- 80 %

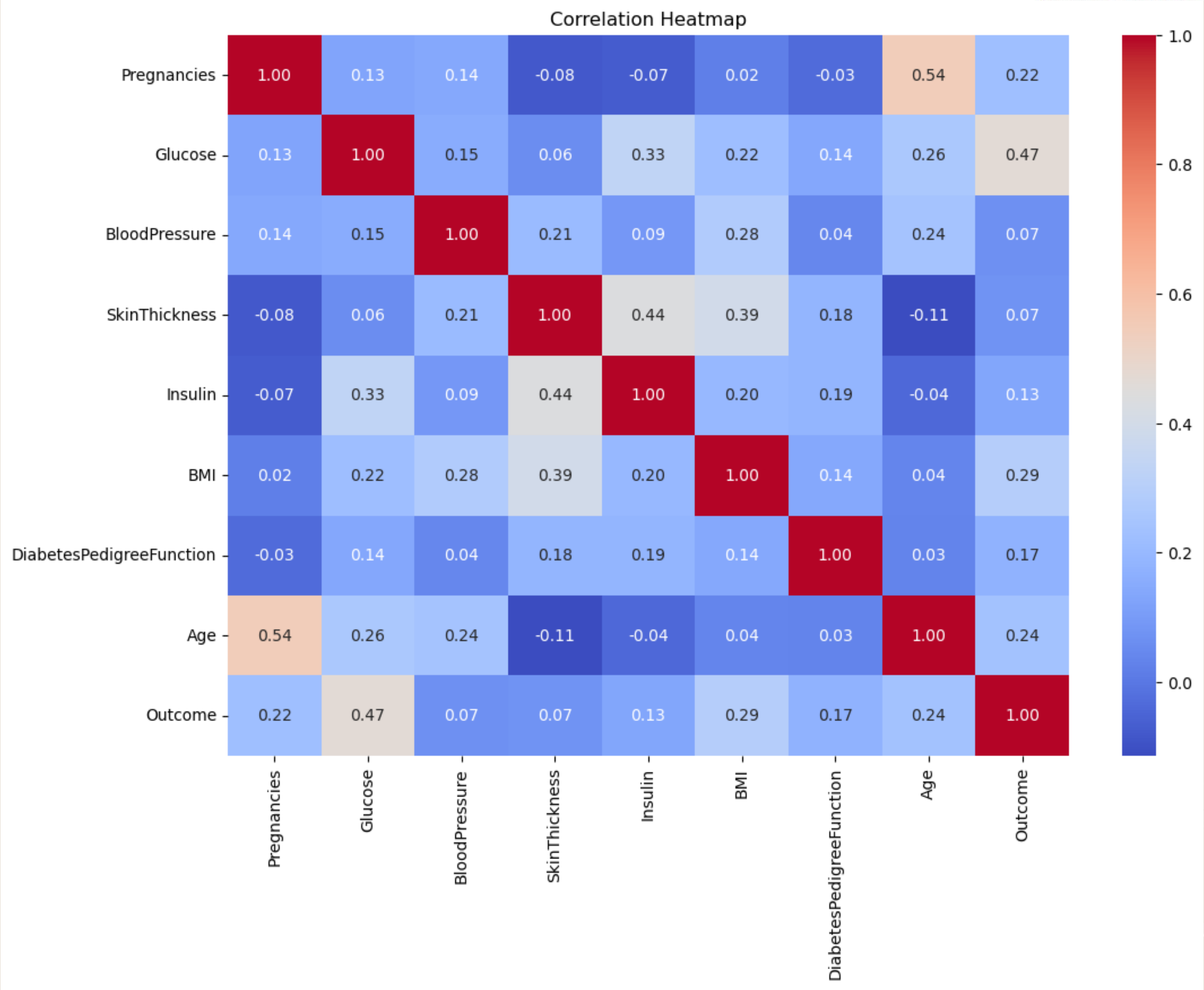
Accuracy of Logistic Regression on Test Data"

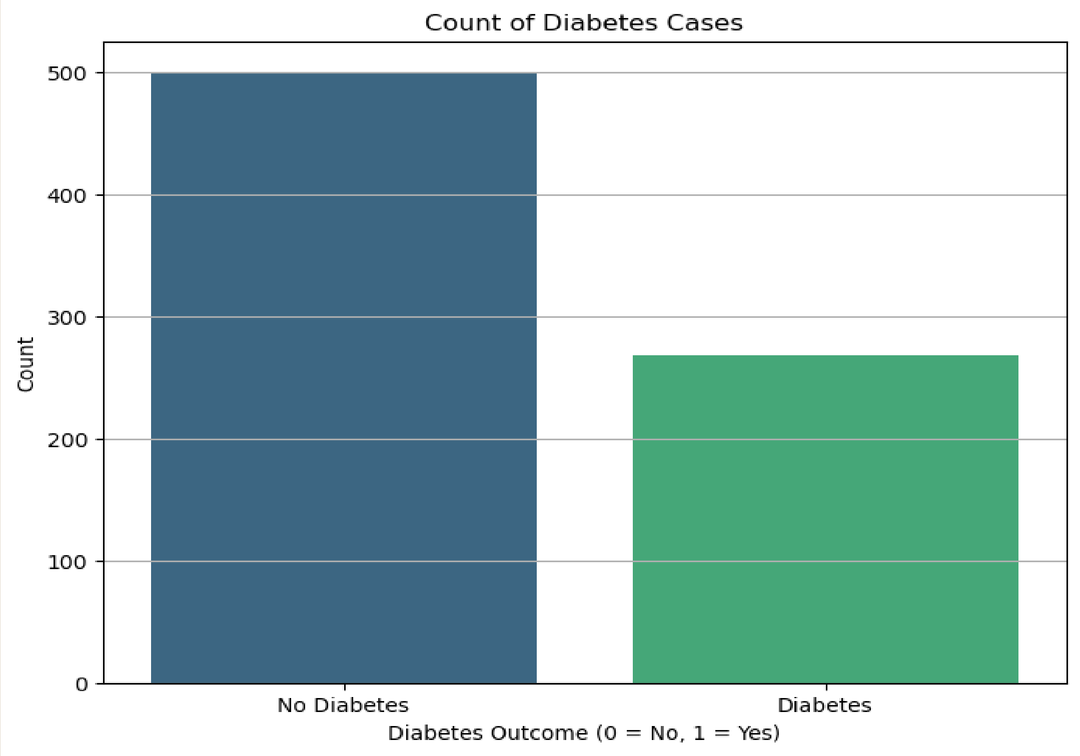
```
print("Accuracy on Test Data: ", test_data_accuracy)
```

```
Accuracy on Test Data:  0.8032786885245902
```

SVC (Diabetes Prediction)

Heatmap for Correlation matrix-->

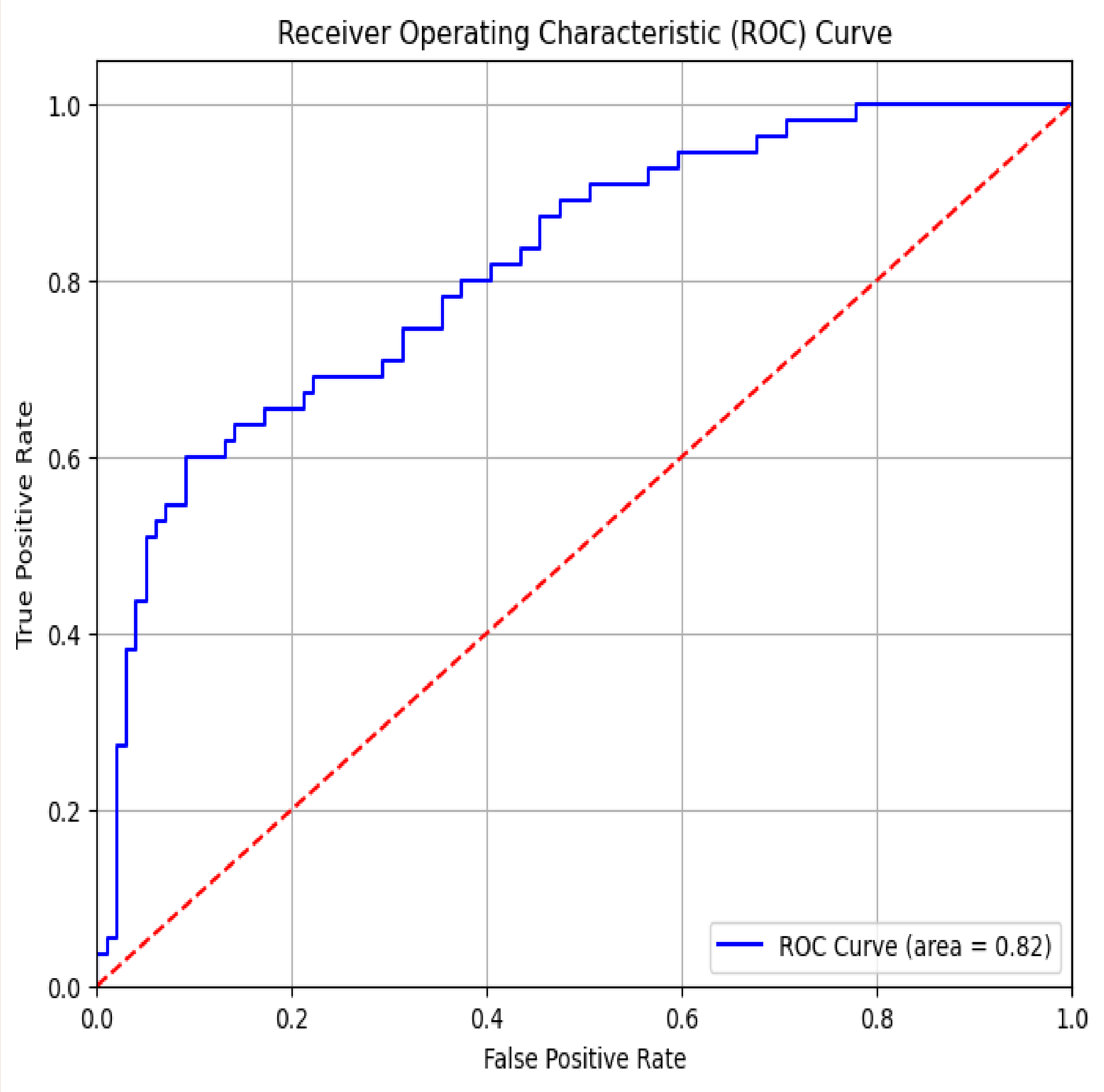




<--Count Plot (Diabetic or non-diabetic)

ROC Curve (Reciever Operating Characteristics)→

-Visualization: ROC curves provide a visual representation of a model's performance, helping to understand how well the SVC model distinguishes between positive and negative classes at different classification thresholds.
-Comparison: ROC curves can be used to compare the performance of multiple SVC models or other classification algorithms, identifying the model with the best overall performance.
-Bias Detection: ROC curves can help detect potential biases in the model. If the curve is skewed towards one class, it might indicate that the model is performing better for one class than the other.



Accuracy for diabetes(SVC) :- 78%

Evaluating Model Performance

```
print("Accuracy score of the training data is: ", training_data_accuracy)
```

```
Accuracy score of the training data is:  0.7866449511400652
```

Discussion :-

--Model Performance: Both logistic regression and SVC demonstrated strong performance in predicting heart disease and diabetes.

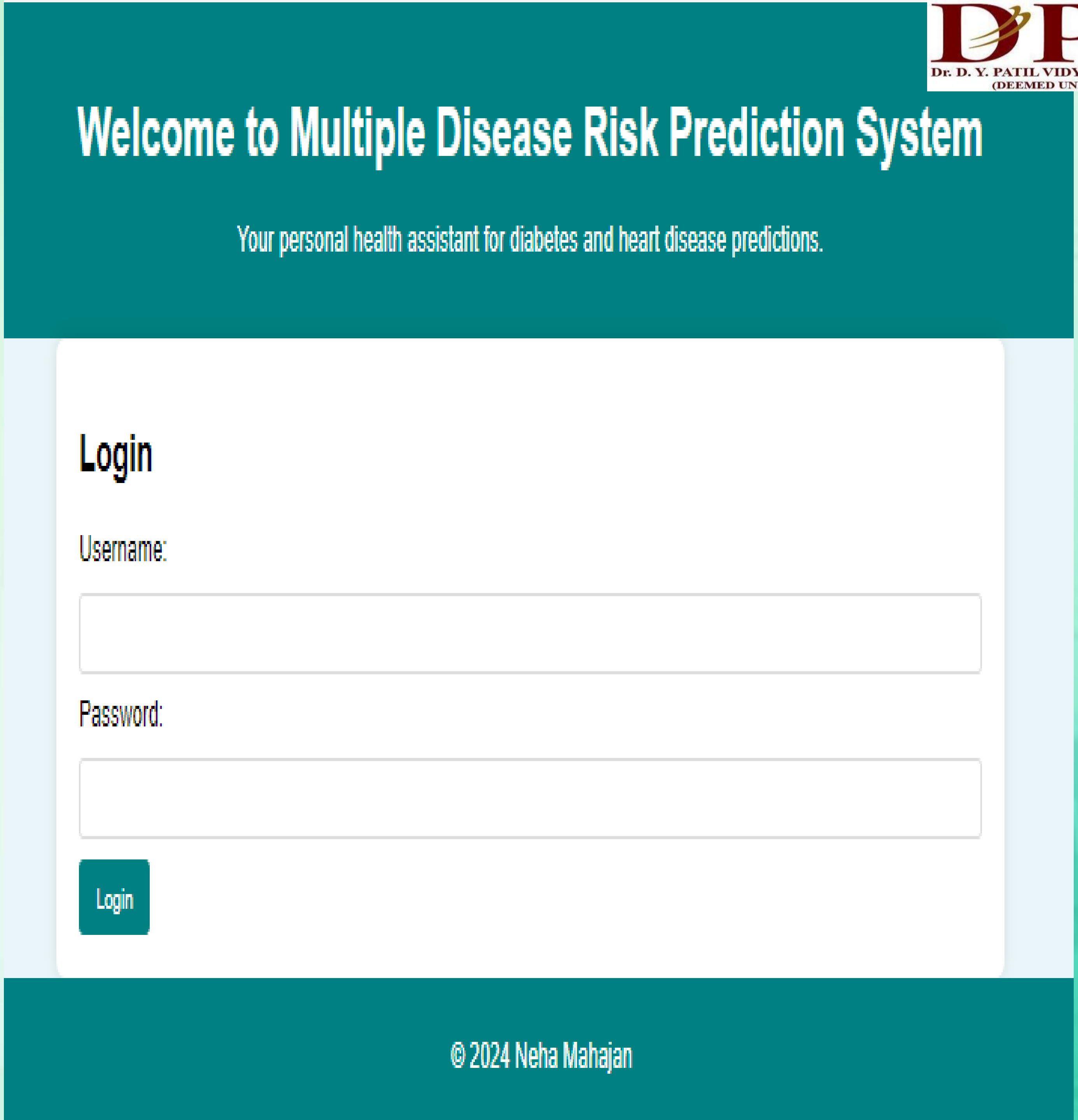
--Model Selection: Logistic regression slightly outperformed SVC in terms of accuracy for heart disease prediction, while SVC was competitive for diabetes prediction.

--Regularization: Regularization techniques had a minimal impact on model performance, suggesting that the datasets were well-balanced and the models were not overly complex.

--Future Directions: Future research could explore more advanced feature engineering techniques, ensemble methods, and explainable AI approaches to further enhance model performance and interpretability.

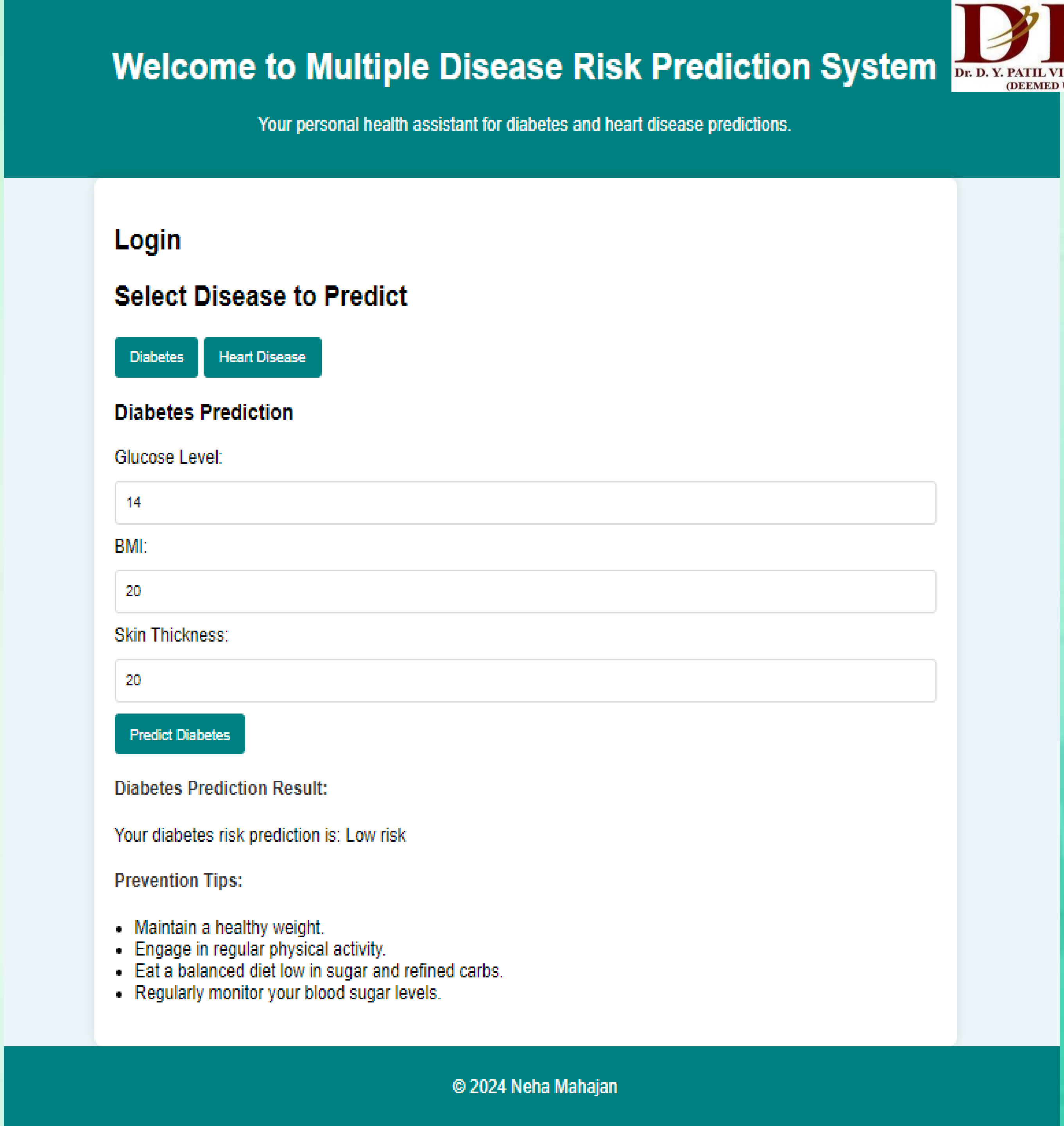
Login page :

- The homepage features a dean and intuitive design that enhances user experience,
- The application serves as a personal health assistant, specifically designed for diabetes and heart disease predictions, providing valuable insights to users regarding their health risks.



Choice -1
Predict Diabetes

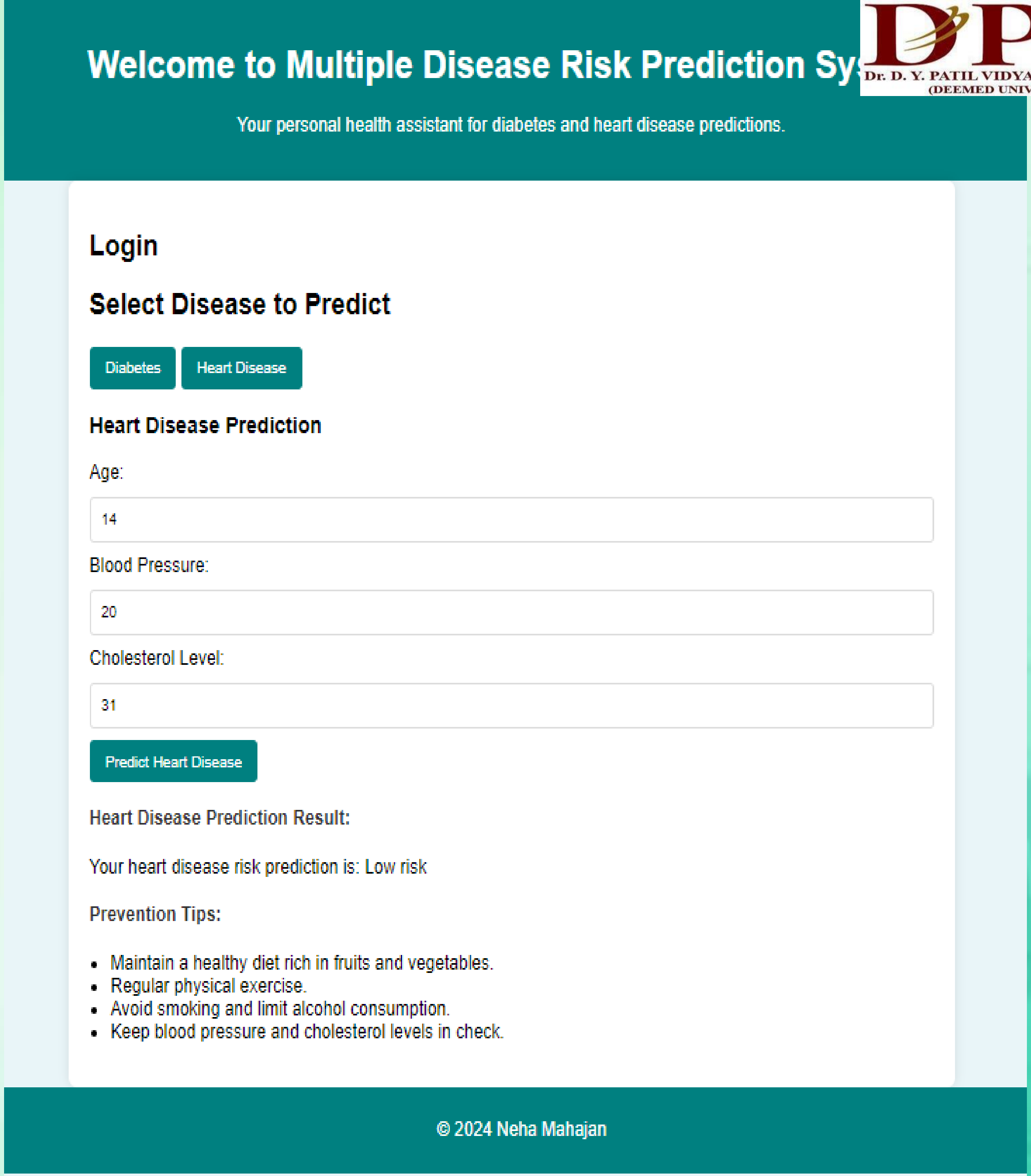
- A higher BMI indicates increased body fat, which raises the risk of insulin resistance and type 2 diabetes.
- Elevated blood glucose levels reflect impaired glucose metabolism, a key indicator of prediabetes or diabetes.
- Increased skin thickness can signify higher subcutaneous fat, correlating with obesity and a higher likelihood of developing diabetes.
- Combining BMI, glucose levels, and skin thickness provides a comprehensive assessment of an individual's risk for diabetes, aiding in early intervention and management strategies.



Dr. D. Y. PATIL VIDYAPEETH, PUNE
(DEEMED UNIVRSITY)

Choice -2
Predict Heart Disease

- Increased age is a significant risk factor for heart disease due to declining cardiovascular health over time.
- High blood pressure can damage arteries and elevate the risk of heart disease and stroke.
- Elevated cholesterol levels lead to plaque buildup in arteries, narrowing them and reducing blood flow.
- A family history of heart disease suggests a genetic predisposition, increasing vulnerability to similar conditions.



DPU
Dr. D. Y. PATIL VIDYAPEETH, PUNE
(DEEMED UNIVERSITY)

Conclusion and Future scope :-

Conclusion: -

--Model Performance: Both logistic regression and SVC performed well, with logistic regression achieving a slightly higher accuracy of 80% compared to SVC's 78%.
--Model Selection: While both models are effective, logistic regression might be a slightly better choice based on the accuracy results in this specific case.
--Further Analysis: It's important to consider other factors beyond accuracy, such as interpretability, computational cost, and domain knowledge, when selecting a model for a particular application.

Future Scope: -

- Explore deep learning models for disease prediction
- Leverage transfer learning for disease prediction
- Develop multi-class classification models for predicting multiple diseases
- Handle imbalanced data to improve model performance
- Improve model interpretability through explainability techniques
- Deploy models in real-world settings for practical evaluation
- Use data augmentation to increase dataset size and improve model performance
- Optimize model performance through hyperparameter tuning

THANK YOU

Dr. D. Y. PATIL VIDYAPEETH, PUNE
(DEEMED UNIVRSITY)