**Diabetes Disease Prediction System - A Machine Learning Approach**

## **Overview:**

Diabetes is a chronic medical condition characterized by high levels of glucose in the blood due to the body's inability to produce or effectively use insulin. The condition can lead to serious health complications such as heart disease, kidney failure, nerve damage, and vision problems if not properly managed. Early diagnosis, lifestyle modifications, and proper treatment are key to managing diabetes and preventing complications. As AI, tools continue to evolve and find its application in medical, image diagnosis, it should be considered as a supportive tool, and assist the clinicians in their assessments.

In this project, we are using two Machine Learning models named as Logistic Regression and Support Vector Machines in detecting whether the particular person has diabetes or not based on the heath conditions provided in the dataset.

## **Objective:**

The main motto of this project is to detect whether the person is having diabetes or not with the help of two machine learning models named as Logistic Regression and Support Vector Machines. From these two models, we can predict which model will provide better accuracy in detecting whether the person is having diabetes or not.

## **Problem Statement:**

Diabetes is a chronic medical condition characterized by high levels of glucose in the blood due to the body's inability to produce or effectively use insulin. Detecting the diabetes at an early stage can serve as potential indicators of saving our lives, which is crucial in our fight against this formidable advisory. Therefore, to detect whether the person is having diabetes or not, we aim to use two machine-learning models named as Logistic Regression and Support Vector Machines for detecting the diabetes in the person.

## **Dataset:**

There are total 768 records of the diabetic patients in the dataset. From this records, there are total eight input features('Pregnancies' ,'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin' , 'BMI','DiabetesPedigreeFunction','Age') and one Output feature ('Outcome')

1. Pregnancies ---------> Number of times pregnant
2. Glucose ---------> Plasma glucose concentration a 2 hours in an oral glucose tolerance test
3. Blood Pressure ---------> Diastolic blood pressure (mm Hg)
4. Skin Thickness ---------> Triceps skin fold thickness (mm)
5. Insulin ---------> 2-Hour serum insulin (mu U/ml)
6. BMI ---------> Body mass index (weight in kg/(height in m)^2)
7. Diabetes Pedigree Function ---------> Diabetes pedigree function
8. Age ---------> Age (years)
9. Outcome ---------> Class variable (0 or 1)

## **Approach**

1. First, the diabetes dataset is obtained.
2. Then the dataset is cleaned and preprocessed.
3. The dataset is divided into independent (input features) and dependent (output features).
4. As all the input features in the dataset are in different range, so Standardization is done to keep all the input features in one range.
5. The dataset is divided into training and testing data.
6. Two Machine Learning Models (Logistic Regression) and (Support Vector Machines) are used to test the accuracy level of the model.
7. Using the above two models, the data is first trained, then using the test data, the model is predicted and we get the predicted output for each algorithms.
8. Then we calculate the accuracy value for each training and testing data for both the models.
9. The model that has better accuracy will be considered for determining the diabetic condition of the patient.
10. A prediction system is also built that helps in determining whether the person is having diabetes or not using a test data.
11. Then we have developed the web application for depicting the diabetic condition of the patient.

## **The hyperparameters used in this project is of below:**

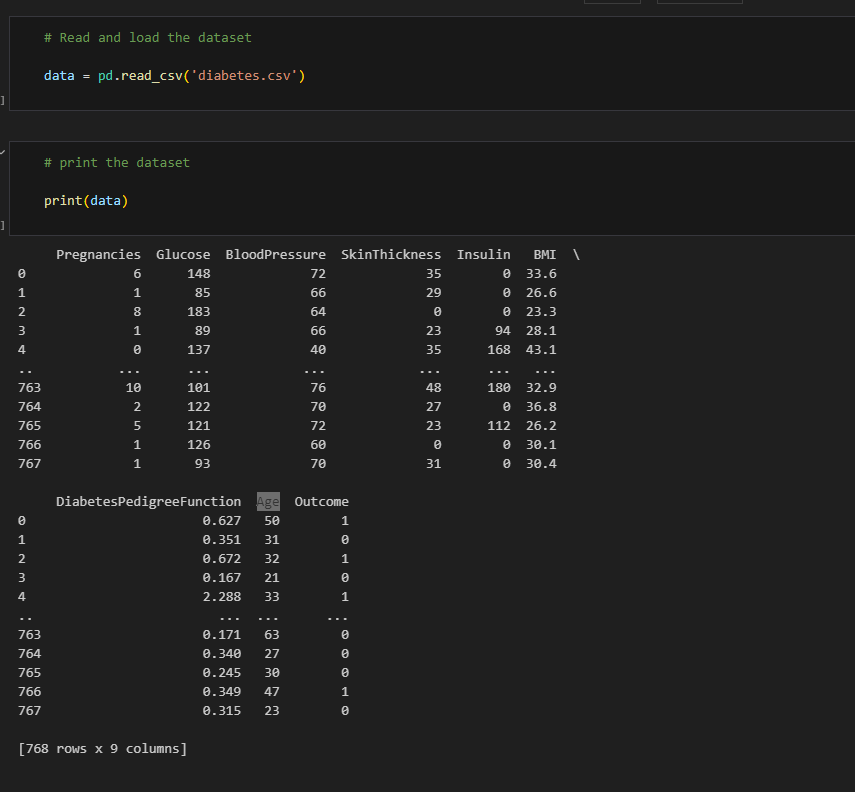
. Input = Training data: 80 % data. Testing data: 20 % data

. Kernel = linear

. Random state = 1

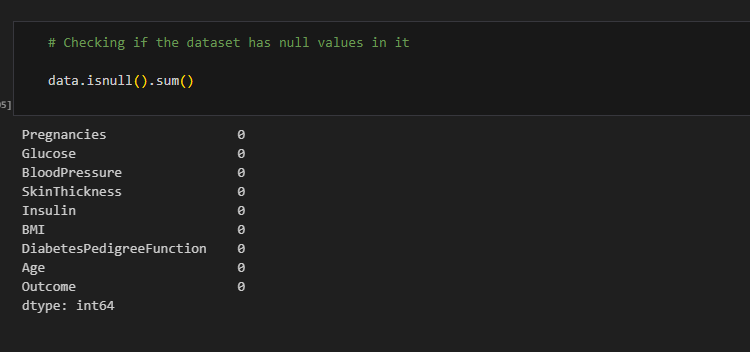
**Implementation:**

* 1. **The dataset is fetched.**



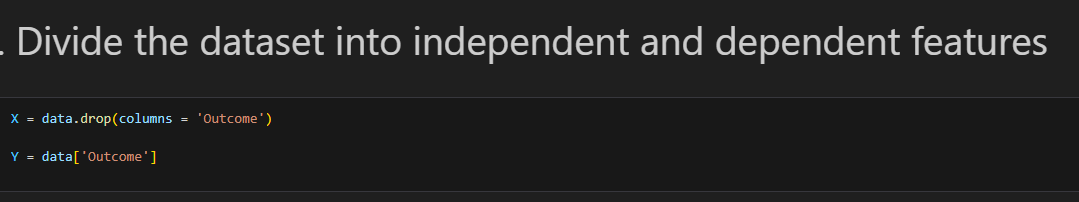
**OBSERVATION:**

* The diabetes dataset is loaded in the system.
  1. **The dataset is preprocessed.**



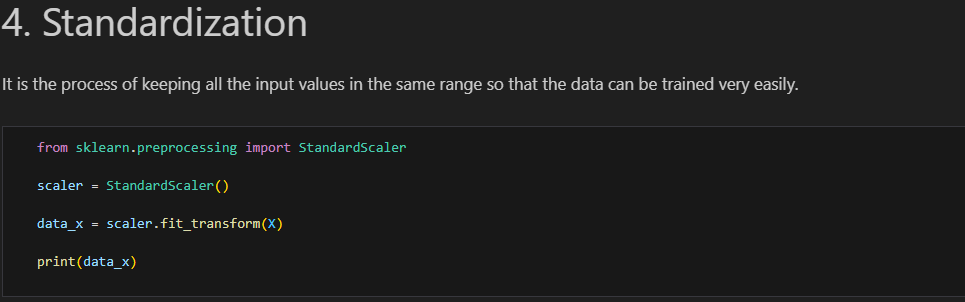
**OBSERVATION:**

* The dataset is preprocessed and checked if there are any NULL entities in it.
* From the above screenshot, it is confirmed that there are no NULL entries in the dataset.
  1. **The dataset is divided into independent and dependent features.**



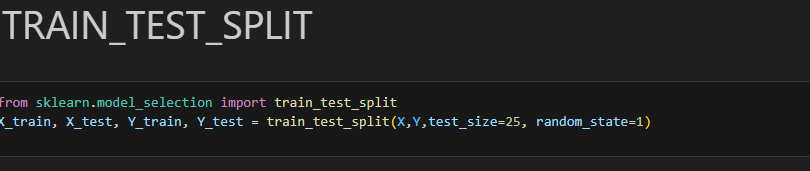
**OBSERVATION:**

* The dataset is divided into input and output features.
  1. **Standization:**



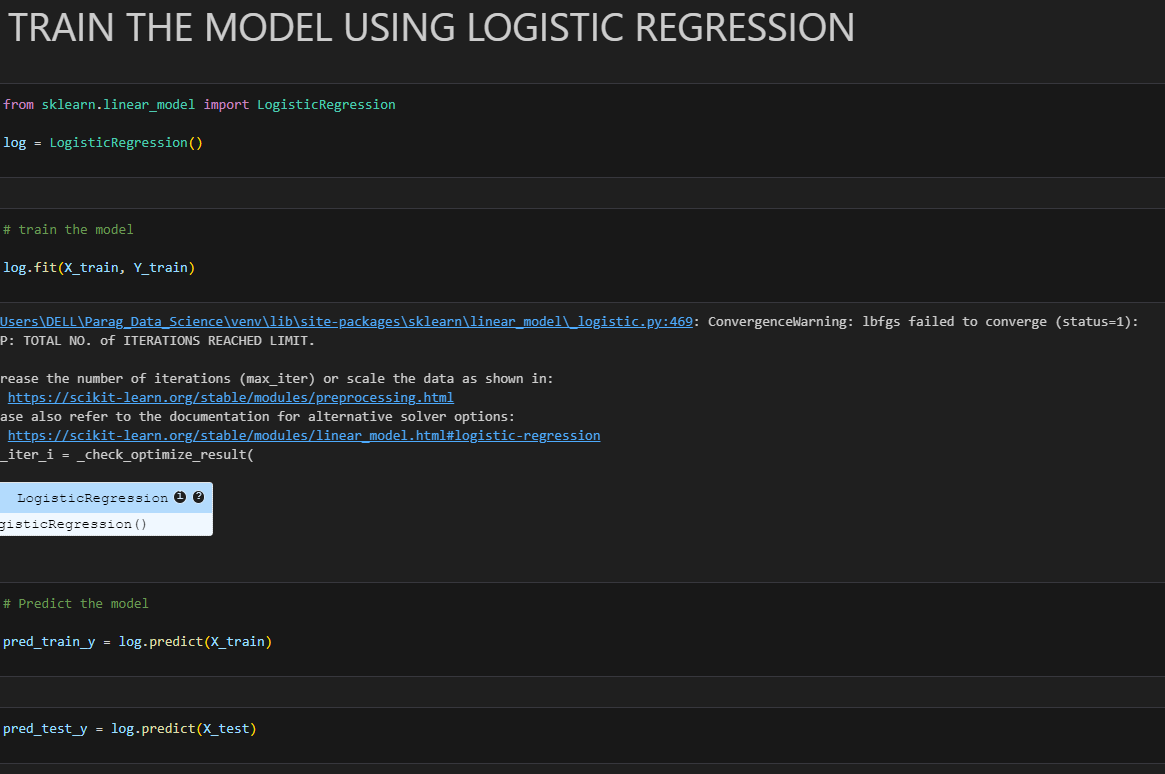
**OBSERVATION:**

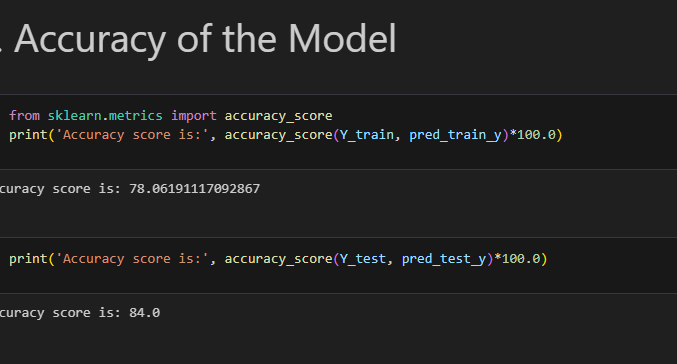
* All the input features are standardized in order to keep them in one range for easy training of the model.
  1. **Train Test Split:**

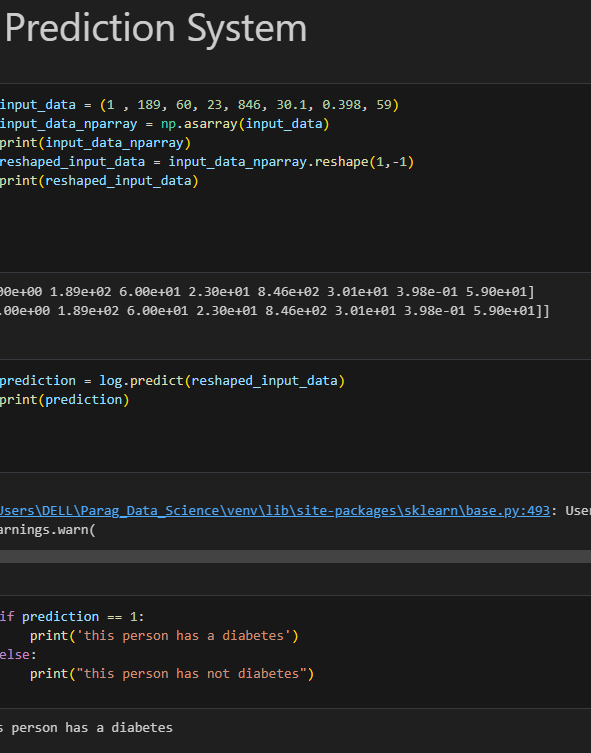


**OBSERVATION:**

* The dataset is divided into training and testing data.
  1. **FIRST MACHINE LEARNING MODEL – LOGISTIC REGRESSION:**

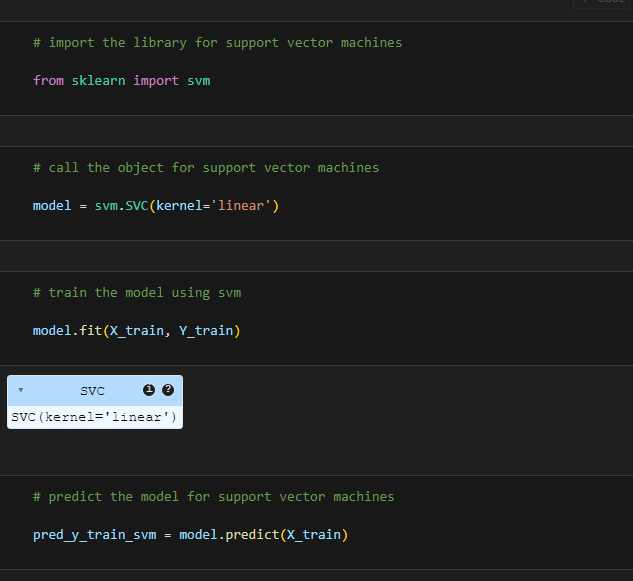


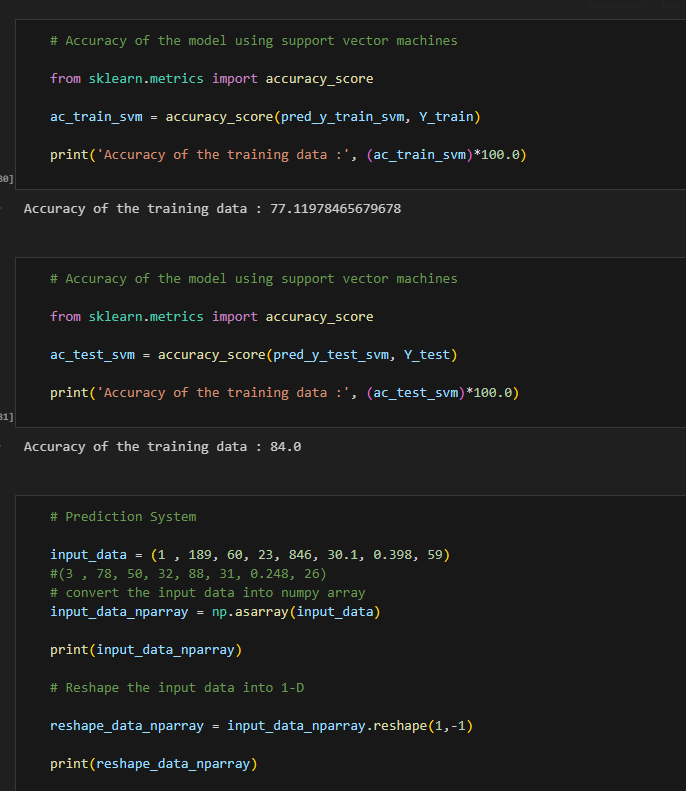




**OBSERVATION:**

* The model is trained using Logistic Regression.
* The model is predicted using the input values of training and testing data.
* The predicted value of training and testing data is obtained.
* The Accuracy score for training and testing data is obtained.
* A sample test data is taken from where based on the model; diabetic condition of the patient is predicted.
  1. **SECOND MACHINE LEARNING MODEL – SUPPORT VECTOR MACHINES:**







**OBSERVATION:**

* The model is trained using Support Vector Machines.
* The model is predicted using the input values of training and testing data.
* The predicted value of training and testing data is obtained.
* The Accuracy score for training and testing data is obtained.
* A sample test data is taken from where based on the model; diabetic condition of the patient is predicted.

## **Project Results**

The model used on MRI images test set provides the following result:

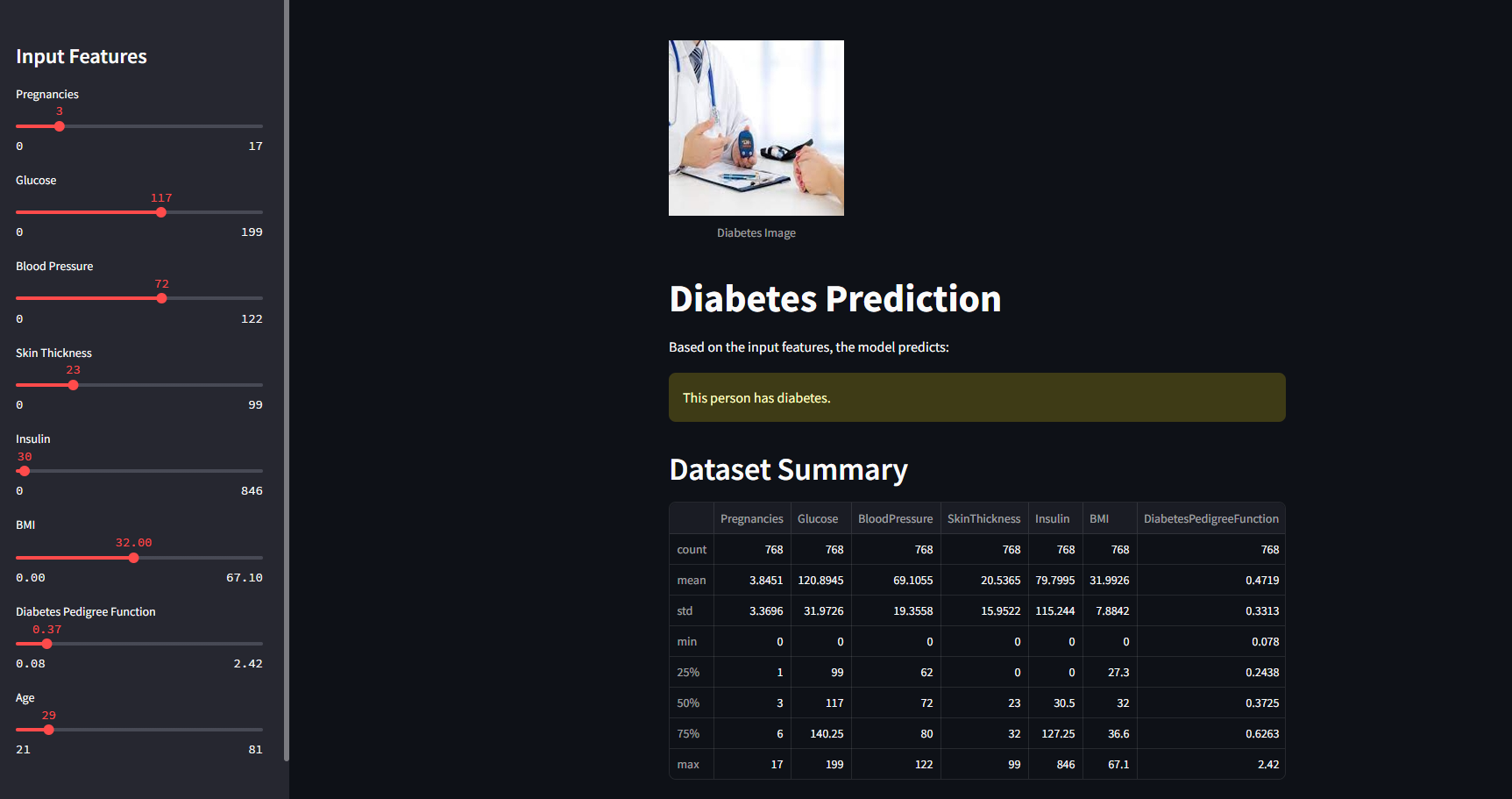
**Logistic Regression:**

* Accuracy for Training Dataset: (78% Accuracy).
* Accuracy for Testing Dataset : (84% Accuracy)

**Support Vector Machines:**

* Accuracy for Training Dataset: (77% Accuracy) .
* Accuracy for Testing Dataset : (84% Accuracy)
  1. **Conclusion:**
* Test data is taken for both the models and it is predicted that the person is having diabetes.

1. **WEB APPLICATION INTERFACE:**



**OBSERVATION:**

* A web application interface is built from where we can see that based on the values of the input features, a person is having diabetes.