**PHASE 3 PROJECT SUBMISSION**

**PROJECT NAME : AI BASED DIABETES PREDICTION SYSTEM**

**NAME : KATHIRAVAN S**

**REGISTER NUMBER : 312821121012**

**AI – BASED DIABETES PREDICTION SYSTEM**

**Abstract:**

Diabetes is a prevalent chronic disease affecting millions worldwide, characterized by the body's inability to regulate blood sugar levels effectively. Timely and accurate diagnosis of diabetes can significantly improve patient outcomes and enable effective management strategies. In this work, we present an AI-based diabetes prediction system aimed at assisting healthcare professionals in early detection and management of diabetes.

The system leverages machine learning algorithms and advanced data analysis techniques to predict the likelihood of an individual having diabetes based on a set of relevant features. These features include critical health indicators such as the number of pregnancies, glucose levels, blood pressure, skin thickness, insulin levels, BMI (Body Mass Index), diabetes pedigree function, and age. By integrating these key features, the system employs sophisticated predictive models to provide reliable and efficient diabetes risk assessment.

The development of the AI-based diabetes prediction system involves a comprehensive data preprocessing phase, encompassing data cleaning, feature selection, and normalization. Subsequently, the selected features are used to train a robust machine learning model, such as logistic regression, decision trees, or neural networks, to accurately classify individuals as either diabetic or non-diabetic. The system is evaluated using standard performance metrics, including accuracy, precision, recall, and F1 score, to ensure its reliability and effectiveness.

Our AI-based diabetes prediction system holds the potential to enhance early intervention strategies, facilitate personalized patient care, and contribute to the overall improvement of diabetes management. With its ability to swiftly assess the risk of diabetes onset, the system can assist healthcare providers in making informed decisions, thereby reducing the risk of complications and promoting better health outcomes for individuals at risk of or already living with diabetes.

**\***

**Context**

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective is to predict based on diagnostic measurements whether a patient has diabetes.

**Content**

Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

* Pregnancies: Number of times pregnant
* Glucose: Plasma glucose concentration a 2 hours in an oral glucose tolerance test
* BloodPressure: Diastolic blood pressure (mm Hg)
* SkinThickness: Triceps skin fold thickness (mm)
* Insulin: 2-Hour serum insulin (mu U/ml)
* BMI: Body mass index (weight in kg/(height in m)^2)
* DiabetesPedigreeFunction: Diabetes pedigree function
* Age: Age (years)
* Outcome: Class variable (0 or 1)

**Sources:**

1. Original owners: National Institute of Diabetes and Digestive and

Kidney Diseases

1. Donor of database: Vincent Sigillito (vgs@aplcen.apl.jhu.edu)

Research Center, RMI Group Leader

Applied Physics Laboratory

The Johns Hopkins University

Johns Hopkins Road

Laurel, MD 20707

(301) 953-6231

(c) Date received: 9 May 1990

**Past Usage:**

1. Smith,~J.~W., Everhart,~J.~E., Dickson,~W.~C., Knowler,~W.~C., \&

Johannes,~R.~S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In {\it Proceedings of the Symposium on Computer Applications and Medical Care} (pp. 261--265). IEEE Computer Society Press.

The diagnostic, binary-valued variable investigated is whether the patient shows signs of diabetes according to World Health Organization criteria (i.e., if the 2 hour post-load plasma glucose was at least 200 mg/dl at any survey examination or if found during routine medical care). The population lives near Phoenix, Arizona, USA.

Results: Their ADAP algorithm makes a real-valued prediction between

1. and 1. This was transformed into a binary decision using a cutoff of 0.448. Using 576 training instances, the sensitivity and specificity of their algorithm was 76% on the remaining 192 instances.

**Relevant Information:**

Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. ADAP is an adaptive learning routine that generates and executes digital analogs of perceptron-like devices. It is a unique algorithm; see the paper for details.

**For Each Attribute: (all numeric-valued)**

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

**Class Distribution: (class value 1 is interpreted as "tested positive for** diabetes")

**Program Code:**

*# Import necessary libraries* import numpy as np import pandas as pd import matplotlib.pyplot as plt import seaborn as sns

from sklearn.model\_selection import train\_test\_split from sklearn.preprocessing import StandardScaler

from sklearn.svm import SVC

from sklearn.metrics import accuracy\_score, classification\_report, confusion\_ matrix

*# Load the dataset (replace 'diabetes.csv' with your dataset file)* df = pd.read\_csv('/kaggle/input/diabetes-data-set/diabetes.csv')

*# Step 1: Data Cleaning*

*# Check for Missing Values* missing\_values = df.isnull().sum() print("Missing Values:") print(missing\_values)

*# Handle missing values (if any)*

*# For example, fill missing values with the mean of the column* mean\_fill = df.mean()

df.fillna(mean\_fill, inplace=True)

*# Check for Duplicate Rows* duplicate\_rows = df[df.duplicated()] print("**\n**Duplicate Rows:") print(duplicate\_rows)

*# Handle duplicate rows (if any)* *# For example, drop duplicate rows* df.drop\_duplicates(inplace=True)

*# Step 2: Data Analysis*

*# Summary Statistics* summary\_stats = df.describe() print("**\n**Summary Statistics:") print(summary\_stats)

*# Class Distribution (for binary classification problems)* class\_distribution = df['Outcome'].value\_counts() print("**\n**Class Distribution:") print(class\_distribution)

*# Step 3: Support Vector Machine (SVM) Modeling*

*# Separate features and target variable* X = df.drop('Outcome', axis=1) y = df['Outcome']

*# Split the dataset into a training and testing set*

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, rand om\_state=42)

*# Standardize features* scaler = StandardScaler()

X\_train = scaler.fit\_transform(X\_train)

X\_test = scaler.transform(X\_test)

*# Initialize and train the SVM model* model = SVC(kernel='linear', random\_state=42) model.fit(X\_train, y\_train)

*# Make predictions*

y\_pred = model.predict(X\_test)

*# Evaluate the model*

accuracy = accuracy\_score(y\_test, y\_pred) print(f'Accuracy: **{**accuracy**:**.2f**}**')

*# Classification report and confusion matrix* print(classification\_report(y\_test, y\_pred)) cm = confusion\_matrix(y\_test, y\_pred) sns.heatmap(cm, annot=True, fmt='d')

plt.show()

Missing Values:

Pregnancies 0

Glucose 0

BloodPressure 0

SkinThickness 0

Insulin 0

BMI 0

DiabetesPedigreeFunction 0

Age 0 Outcome 0 dtype: int64

Duplicate Rows:

Empty DataFrame

Columns: [Pregnancies, Glucose, BloodPressure, SkinThickness, Insulin, BMI

, DiabetesPedigreeFunction, Age, Outcome]

Index: []

Summary Statistics:

Pregnancies Glucose BloodPressure SkinThickness Insulin

\

count 768.000000 768.000000 768.000000 768.000000 768.000000 mean 3.845052 120.894531 69.105469 20.536458 79.799479 std 3.369578 31.972618 19.355807 15.952218 115.244002 min 0.000000 0.000000 0.000000 0.000000 0.000000 25% 1.000000 99.000000 62.000000 0.000000 0.000000

50% 3.000000 117.000000 72.000000 23.000000 30.500000 75% 6.000000 140.250000 80.000000 32.000000 127.250000 max 17.000000 199.000000 122.000000 99.000000 846.000000

BMI DiabetesPedigreeFunction Age Outcome count 768.000000 768.000000 768.000000 768.000000 mean 31.992578 0.471876 33.240885 0.348958 std 7.884160 0.331329 11.760232 0.476951 min 0.000000 0.078000 21.000000 0.000000 25% 27.300000 0.243750 24.000000 0.000000

50% 32.000000 0.372500 29.000000 0.000000 75% 36.600000 0.626250 41.000000 1.000000 max 67.100000 2.420000 81.000000 1.000000

Class Distribution:

Outcome

1. 500
2. 268

Name: count, dtype: int64 Accuracy: 0.76

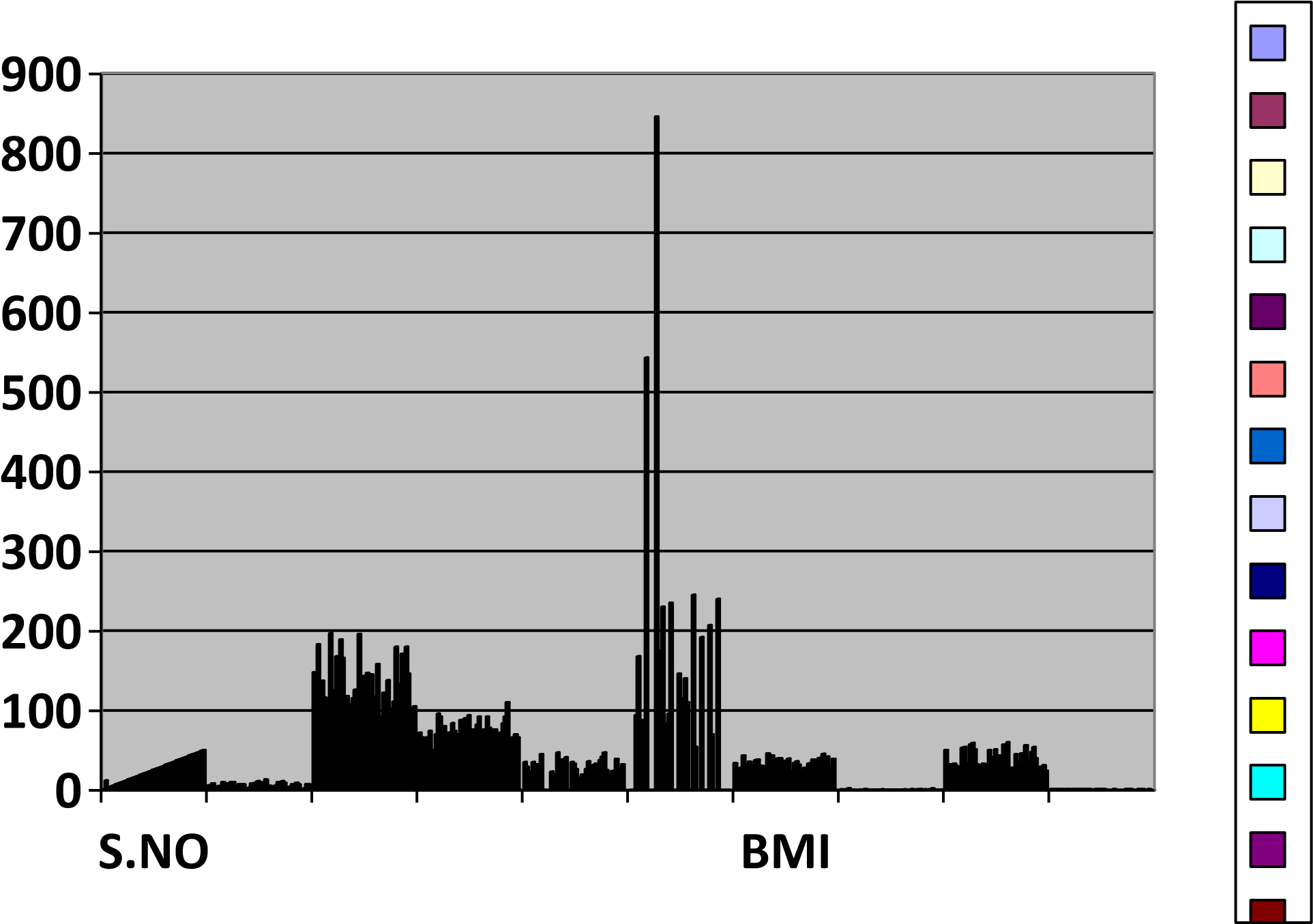
precision recall f1-score support

* 1. 0.81 0.82 0.81 99
  2. 0.67 0.65 0.66 55

accuracy 0.76 154 macro avg 0.74 0.74 0.74 154 weighted avg 0.76 0.76 0.76 154

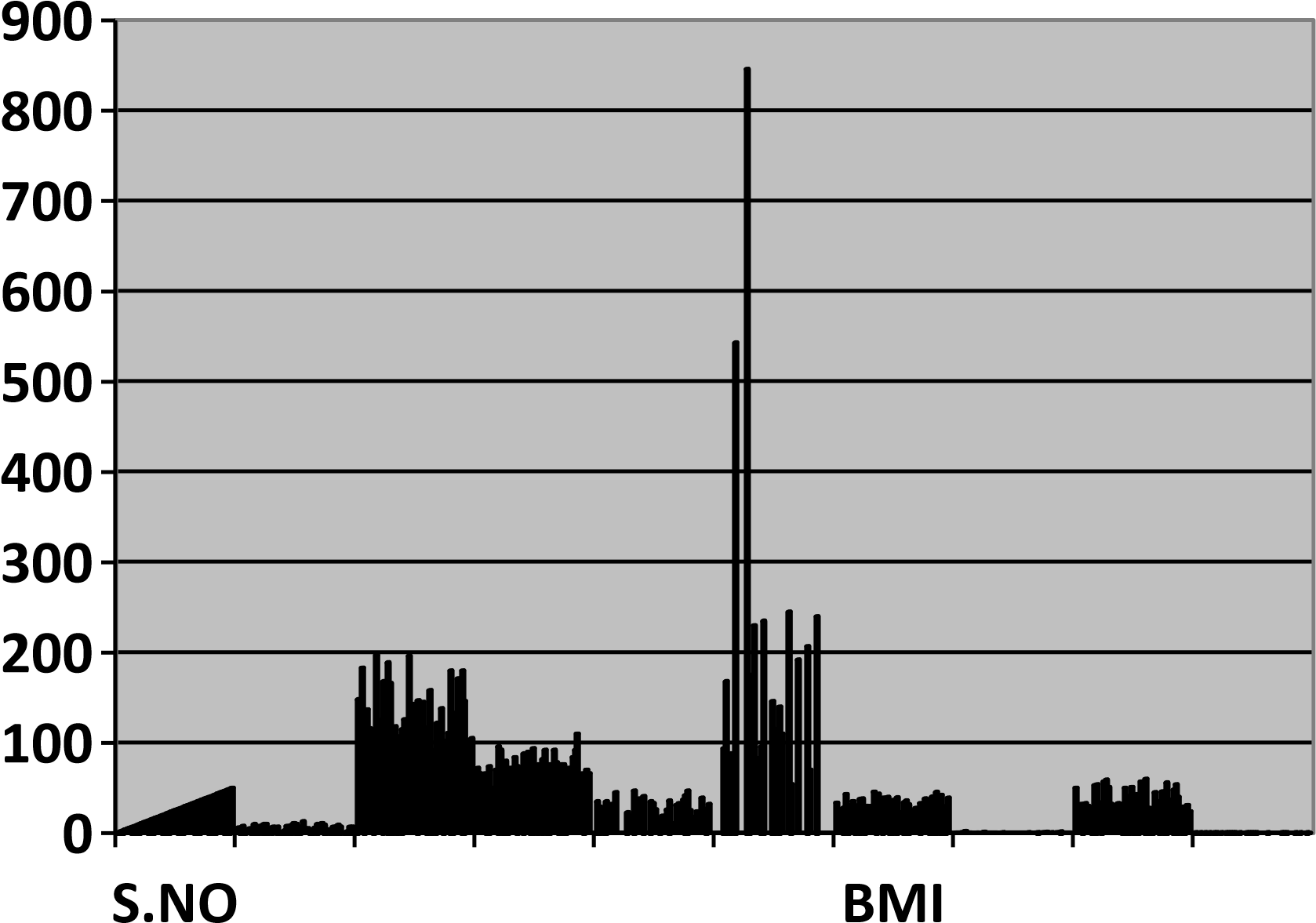
**INPUTS**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **S.N O** | **Pregnancies** | **Glucose** | **BloodPressure** | **SkinThickness** | **Insulin** | **BMI** | **DiabetesPedigreeFunction** | **Age** |  | **Outco me** |
| 1 | 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 |  | 50 | 1 |
| 12 | 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 |  | 31 | 0 |
| 3 | 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 |  | 32 | 1 |
| 4 | 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 |  | 21 | 0 |
| 5 | 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 |  | 33 | 1 |
| 6 | 5 | 116 | 74 | 0 | 0 | 25.6 | 0.201 |  | 30 | 0 |
| 7 | 3 | 78 | 50 | 32 | 88 | 31 | 0.248 |  | 26 | 1 |
| 8 | 10 | 115 | 0 | 0 | 0 | 35.3 | 0.134 |  | 29 | 0 |
| 9 | 2 | 197 | 70 | 45 | 543 | 30.5 | 0.158 |  | 53 | 1 |
| 10 | 8 | 125 | 96 | 0 | 0 | 0 | 0.232 |  | 54 | 1 |
| 11 | 4 | 110 | 92 | 0 | 0 | 37.6 | 0.191 |  | 30 | 0 |
| 12 | 10 | 168 | 74 | 0 | 0 | 38 | 0.537 |  | 34 | 1 |
| 13 | 10 | 139 | 80 | 0 | 0 | 27.1 | 1.441 |  | 57 | 0 |
| 14 | 1 | 189 | 60 | 23 | 846 | 30.1 | 0.398 |  | 59 | 1 |
| 15 | 5 | 166 | 72 | 19 | 175 | 25.8 | 0.587 |  | 51 | 1 |
| 16 | 7 | 100 | 0 | 0 | 0 | 30 | 0.484 |  | 32 | 1 |
| 17 | 0 | 118 | 84 | 47 | 230 | 45.8 | 0.551 |  | 31 | 1 |
| 18 | 7 | 107 | 74 | 0 | 0 | 29.6 | 0.254 |  | 31 | 1 |
| 19 | 1 | 103 | 30 | 38 | 83 | 43.3 | 0.183 |  | 33 | 0 |
| 20 | 1 | 115 | 70 | 30 | 96 | 34.6 | 0.529 |  | 32 | 1 |
| 21 | 3 | 126 | 88 | 41 | 235 | 39.3 | 0.704 |  | 27 | 0 |
| 22 | 8 | 99 | 84 | 0 | 0 | 35.4 | 0.388 |  | 50 | 0 |
| 23 | 7 | 196 | 90 | 0 | 0 | 39.8 | 0.451 |  | 41 | 1 |
| 24 | 9 | 119 | 80 | 35 | 0 | 29 | 0.263 |  | 29 | 1 |
| 25 | 11 | 143 | 94 | 33 | 146 | 36.6 | 0.254 |  | 51 | 1 |
| 26 | 10 | 125 | 70 | 26 | 115 | 31.1 | 0.205 |  | 41 | 1 |
| 27 | 7 | 147 | 76 | 0 | 0 | 39.4 | 0.257 |  | 43 | 1 |
| 28 | 1 | 97 | 66 | 15 | 140 | 23.2 | 0.487 |  | 22 | 0 |
| 29 | 13 | 145 | 82 | 19 | 110 | 22.2 | 0.245 |  | 57 | 0 |
| 30 | 5 | 117 | 92 | 0 | 0 | 34.1 | 0.337 |  | 38 | 0 |
| 31 | 5 | 109 | 75 | 26 | 0 | 36 | 0.546 |  | 60 | 0 |
| 32 | 3 | 158 | 76 | 36 | 245 | 31.6 | 0.851 |  | 28 | 1 |
| 33 | 3 | 88 | 58 | 11 | 54 | 24.8 | 0.267 |  | 22 | 0 |
| 34 | 6 | 92 | 92 | 0 | 0 | 19.9 | 0.188 |  | 28 | 0 |
| 35 | 10 | 122 | 78 | 31 | 0 | 27.6 | 0.512 |  | 45 | 0 |
| 36 | 4 | 103 | 60 | 33 | 192 | 24 | 0.966 |  | 33 | 0 |
| 37 | 11 | 138 | 76 | 0 | 0 | 33.2 | 0.42 |  | 35 | 0 |
| 38 | 9 | 102 | 76 | 37 | 0 | 32.9 | 0.665 |  | 46 | 1 |
| 39 | 2 | 90 | 68 | 42 | 0 | 38.2 | 0.503 |  | 27 | 1 |
| 40 | 4 | 111 | 72 | 47 | 207 | 37.1 | 1.39 |  | 56 | 1 |
| 41 | 3 | 180 | 64 | 25 | 70 | 34 | 0.271 |  | 26 | 0 |
| 42 | 7 | 133 | 84 | 0 | 0 | 40.2 | 0.696 |  | 37 | 0 |
| 43 | 7 | 106 | 92 | 18 | 0 | 22.7 | 0.235 |  | 48 | 0 |
| 44 | 9 | 171 | 110 | 24 | 240 | 45.4 | 0.721 |  | 54 | 1 |
| 45 | 7 | 159 | 64 | 0 | 0 | 27.4 | 0.294 | 40 | | 0 |
| 46 | 0 | 180 | 66 | 39 | 0 | 42 | 1.893 | 25 | | 1 |
| 47 | 1 | 146 | 56 | 0 | 0 | 29.7 | 0.564 | 29 | | 0 |
| 48 | 2 | 71 | 70 | 27 | 0 | 28 | 0.586 | 22 | | 0 |
| 49 | 7 | 103 | 66 | 32 | 0 | 39.1 | 0.344 | 31 | | 1 |
| 50 | 7 | 105 | 0 | 0 | 0 | 0 | 0.305 | 24 | | 0 |



**OUTPUTS**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **S.NO** | **Pregnancies** | **Glucose** | **BloodPressure** | **SkinThickness** | **Insulin** | **BMI** | **DiabetesPedigreeFunction** | **Age** | **Outcome** |
| 1 | 6 | 148 | 72 | 35 | 0 | 33.6 | 0.627 | 50 | 1 |
| 2 | 1 | 85 | 66 | 29 | 0 | 26.6 | 0.351 | 31 | 0 |
| 3 | 8 | 183 | 64 | 0 | 0 | 23.3 | 0.672 | 32 | 1 |
| 4 | 1 | 89 | 66 | 23 | 94 | 28.1 | 0.167 | 21 | 0 |
| 5 | 0 | 137 | 40 | 35 | 168 | 43.1 | 2.288 | 33 | 1 |
| 6 | 5 | 116 | 74 | 0 | 0 | 25.6 | 0.201 | 30 | 0 |
| 7 | 3 | 78 | 50 | 32 | 88 | 31 | 0.248 | 26 | 1 |
| 8 | 10 | 115 | 0 | 0 | 0 | 35.3 | 0.134 | 29 | 0 |
| 9 | 2 | 197 | 70 | 45 | 543 | 30.5 | 0.158 | 53 | 1 |
| 10 | 8 | 125 | 96 | 0 | 0 | 0 | 0.232 | 54 | 1 |
| 11 | 4 | 110 | 92 | 0 | 0 | 37.6 | 0.191 | 30 | 0 |
| 12 | 10 | 168 | 74 | 0 | 0 | 38 | 0.537 | 34 | 1 |
| 13 | 10 | 139 | 80 | 0 | 0 | 27.1 | 1.441 | 57 | 0 |
| 14 | 1 | 189 | 60 | 23 | 846 | 30.1 | 0.398 | 59 | 1 |
| 15 | 5 | 166 | 72 | 19 | 175 | 25.8 | 0.587 | 51 | 1 |
| 16 | 7 | 100 | 0 | 0 | 0 | 30 | 0.484 | 32 | 1 |
| 17 | 0 | 118 | 84 | 47 | 230 | 45.8 | 0.551 | 31 | 1 |
| 18 | 7 | 107 | 74 | 0 | 0 | 29.6 | 0.254 | 31 | 1 |
| 19 | 1 | 103 | 30 | 38 | 83 | 43.3 | 0.183 | 33 | 0 |
| 20 | 1 | 115 | 70 | 30 | 96 | 34.6 | 0.529 | 32 | 1 |
| 21 | 3 | 126 | 88 | 41 | 235 | 39.3 | 0.704 | 27 | 0 |
| 22 | 8 | 99 | 84 | 0 | 0 | 35.4 | 0.388 | 50 | 0 |
| 23 | 7 | 196 | 90 | 0 | 0 | 39.8 | 0.451 | 41 | 1 |
| 24 | 9 | 119 | 80 | 35 | 0 | 29 | 0.263 | 29 | 1 |
| 25 | 11 | 143 | 94 | 33 | 146 | 36.6 | 0.254 | 51 | 1 |
| 26 | 10 | 125 | 70 | 26 | 115 | 31.1 | 0.205 | 41 | 1 |
| 27 | 7 | 147 | 76 | 0 | 0 | 39.4 | 0.257 | 43 | 1 |
| 28 | 1 | 97 | 66 | 15 | 140 | 23.2 | 0.487 | 22 | 0 |
| 29 | 13 | 145 | 82 | 19 | 110 | 22.2 | 0.245 | 57 | 0 |
| 30 | 5 | 117 | 92 | 0 | 0 | 34.1 | 0.337 | 38 | 0 |
| 31 | 5 | 109 | 75 | 26 | 0 | 36 | 0.546 | 60 | 0 |
| 32 | 3 | 158 | 76 | 36 | 245 | 31.6 | 0.851 | 28 | 1 |
| 33 | 3 | 88 | 58 | 11 | 54 | 24.8 | 0.267 | 22 | 0 |
| 34 | 6 | 92 | 92 | 0 | 0 | 19.9 | 0.188 | 28 | 0 |
| 35 | 10 | 122 | 78 | 31 | 0 | 27.6 | 0.512 | 45 | 0 |
| 36 | 4 | 103 | 60 | 33 | 192 | 24 | 0.966 | 33 | 0 |
| 37 | 11 | 138 | 76 | 0 | 0 | 33.2 | 0.42 | 35 | 0 |
| 38 | 9 | 102 | 76 | 37 | 0 | 32.9 | 0.665 | 46 | 1 |
| 39 | 2 | 90 | 68 | 42 | 0 | 38.2 | 0.503 | 27 | 1 |
| 40 | 4 | 111 | 72 | 47 | 207 | 37.1 | 1.39 | 56 | 1 |
| 41 | 3 | 180 | 64 | 25 | 70 | 34 | 0.271 | 26 | 0 |
| 42 | 7 | 133 | 84 | 0 | 0 | 40.2 | 0.696 | 37 | 0 |
| 43 | 7 | 106 | 92 | 18 | 0 | 22.7 | 0.235 | 48 | 0 |
| 44 | 9 | 171 | 110 | 24 | 240 | 45.4 | 0.721 | 54 | 1 |
| 45 | 7 | 159 | 64 | 0 | 0 | 27.4 | 0.294 | 40 | 0 |
| 46 | 0 | 180 | 66 | 39 | 0 | 42 | 1.893 | 25 | 1 |
| 47 | 1 | 146 | 56 | 0 | 0 | 29.7 | 0.564 | 29 | 0 |
| 48 | 2 | 71 | 70 | 27 | 0 | 28 | 0.586 | 22 | 0 |
| 49 | 7 | 103 | 66 | 32 | 0 | 39.1 | 0.344 | 31 | 1 |
| 50 | 7 | 105 | 0 | 0 | 0 | 0 | 0.305 | 24 | 0 |



**Conclusion:**

In conclusion, the AI-based diabetes prediction system represents a significant advancement in healthcare technology. By harnessing the power of artificial intelligence and machine learning, this system offers a non-invasive, cost-effective, and highly accurate means of identifying individuals at risk of diabetes. Early detection is crucial for effective management and prevention of complications associated with diabetes.

Furthermore, the system can be seamlessly integrated into existing healthcare infrastructures, providing valuable support to medical professionals. The ability to assess diabetes risk swiftly and accurately empowers healthcare providers to tailor interventions, monitor at-risk populations, and engage in proactive diabetes management.

While this system shows great promise, ongoing research and development are essential to refine predictive models, incorporate additional health parameters, and expand its applicability to diverse patient demographics. As we look to the future, the AIbased diabetes prediction system stands as a testament to the potential of artificial intelligence in revolutionizing healthcare and improving the lives of individuals affected by diabetes.