**AI-Based Diabetes Prediction System**

**211521243058: m.j guru venkatesh   
 Phase-1 Document Submission**

**Problem Definition:**

The challenge is to develop an AI-powered Diabetes Prediction System that utilizes machine learning algorithms to analyze medical data and forecast the likelihood of an individual developing diabetes. This system aims to offer early risk assessment and personalized preventive measures, empowering individuals to proactively manage their health and reduce the risks associated with diabetes.

**Design Thinking Approach:**

**1. Data Collection:** The foundation of this system lies in the collection of a comprehensive medical dataset. It encompasses vital health attributes like glucose levels, blood pressure, BMI, and information about whether the individual has diabetes or not. The richness and diversity of this dataset are crucial for robust predictions.

Dataset Link:[**https://www.kaggle.com/datasets/mathchi/diabetes-data-set**](https://www.kaggle.com/datasets/mathchi/diabetes-data-set)





**2. Data Preprocessing:** To ensure data quality and model effectiveness, the collected medical data undergoes a meticulous process of cleaning, normalization, and preparation. This step addresses issues such as missing values and outliers, ensuring that the data is ready for machine learning model training.

**3. Feature Selection:** The identification of relevant features plays a pivotal role in the system's success. By selecting the most informative attributes, we reduce model complexity and enhance prediction accuracy. Techniques like mutual information guide this process.

**4. Model Selection:** Experimentation with a variety of machine learning algorithms, including Logistic Regression, Random Forest, and Gradient Boosting, allows us to determine the most suitable approach for diabetes risk prediction. Model selection is a critical step in achieving high accuracy.

**PYTHON PROGRAMING:**

import numpy as np # linear algebra  
import pandas as pd # data processing, CSV file I/O  
import matplotlib.pyplot as plt  
import seaborn as sns  
dataset=pd.read\_csv("/kaggle/input/diabetes-dataset/diabetes.csv")

*# Replacing zero values with NaN*  
dataset\_new = dataset  
dataset\_new[["Glucose","BloodPressure","SkinThickness", "Insulin", "BMI"]] = dataset\_new[["Glucose", "BloodPressure","SkinThickness","Insulin","BMI"]].replace(0, np.NaN)

# Replacing NaN with mean values  
dataset\_new["Glucose"].fillna(dataset\_new["Glucose"].mean(), inplace = True)  
dataset\_new["BloodPressure"].fillna(dataset\_new["BloodPressure"].mean(), inplace = True)  
dataset\_new["SkinThickness"].fillna(dataset\_new["SkinThickness"].mean(), inplace = True)  
dataset\_new["Insulin"].fillna(dataset\_new["Insulin"].mean(), inplace = True)  
dataset\_new["BMI"].fillna(dataset\_new["BMI"].mean(), inplace = True)  
dataset\_new.isnull().sum()

#Logistic regression  
y = dataset\_new['Outcome']  
X = dataset\_new.drop('Outcome', axis=1)

# Splitting X and Y  
from sklearn.model\_selection import train\_test\_split  
X\_train, X\_test, Y\_train,Y\_test=train\_test\_split(X, y, test\_size = 0.20, random\_state = 42, stratify = dataset\_new['Outcome'] )

from sklearn.linear\_model import LogisticRegression  
model = LogisticRegression()  
model.fit(X\_train, Y\_train)  
y\_predict = model.predict(X\_test)  
y\_predict

# Confusion matrix  
from sklearn.metrics import confusion\_matrix  
cm = confusion\_matrix(Y\_test, y\_predict)

# Heatmap of Confusion matrix

sns.heatmap(pd.DataFrame(cm), annot=True)

from sklearn.metrics import accuracy\_score  
accuracy =accuracy\_score(Y\_test, y\_predict)  
accuracy

#Example: Let's check whether the person have diabetes or not using some random values  
y\_predict=model.predict([[1,148,72,35,79.799,33.6,0.627,50]])  
print(y\_predict)  
if y\_predict==1:  
 print("Diabetic")  
else:  
 print("Non Diabetic")

**OUTPUT:**

Pregnancies 0

Glucose 0

BloodPressure 0

SkinThickness 0

Insulin 0

BMI 0

DiabetesPedigreeFunction 0

Age 0

Outcome 0

dtype: int64

array([1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1,

0, 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0,

0, 0, 0, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 1, 0, 0,

1, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,

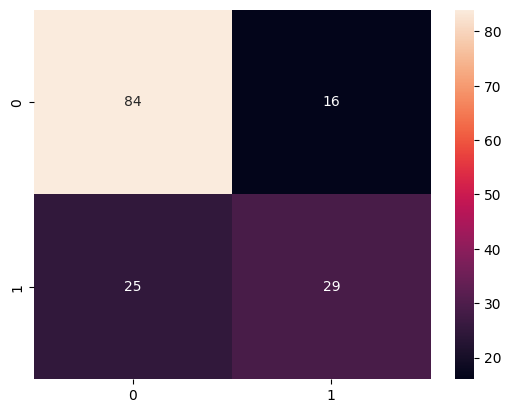
0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0, 0, 1, 0, 1, 0, 1, 1,

0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 1, 1, 1, 0, 0, 0, 1,

0, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0])

array([[84, 16],

[25, 29]])



<Axes:>

0.7337662337662337

Diabetic

/opt/conda/lib/python3.10/site-packages/sklearn/base.py:439: UserWarning: X does not have valid feature names, but LogisticRegression was fitted with fe ature names

warnings.warn(

**5. Evaluation:** Rigorous model evaluation is essential. Metrics such as accuracy, precision, recall, F1-score, and ROC-AUC are employed to assess the model's performance. Validation against separate datasets ensures the reliability of predictions.

**6. Iterative Improvement:** Continuous enhancement of the system is fundamental for long-term effectiveness. Fine-tuning model parameters, exploring feature engineering techniques, and adapting to emerging research findings are ongoing practices that contribute to improved prediction accuracy and relevance.

In summary, the AI-Based Diabetes Prediction System is a proactive solution that harnesses the power of machine learning to provide early risk assessment and personalized preventive measures. This comprehensive approach, from data collection to iterative improvement, aims to empower individuals in managing their health and combating the challenges associated with diabetes.