AI-Based diabetes prediction system

Outline of problem statement, design thinking process, and the phases of development.

problem statement:

Design an AI-based Diabetes Prediction System that can accurately predict the likelihood of an individual developing diabetes based on their health and demographic information. The system should aim to assist healthcare professionals and individuals in early diabetes detection, thus enabling timely interventions and personalized care.

Design thinking process:

Problem Definition:

Clearly define the problem and its scope. Specify the objectives, target audience (e.g., healthcare providers, individuals), and the desired accuracy of the prediction.

Data Collection:

- Gather a comprehensive dataset that includes features such as age, gender, family history, body mass index (BMI), blood pressure, glucose levels, and other relevant health parameters.
- Ensure that the dataset is diverse and representative of the population in question.

Data Preprocessing:

- Clean the dataset by handling missing values and outliers.
- Normalize or standardize features to ensure they are on a common scale.
- Encode categorical variables if necessary (e.g., one-hot encoding for gender).

Feature Selection:

- **Use feature selection techniques to identify the most relevant features for diabetes prediction. This can help improve the model's efficiency and interpretability.**

Model Selection:

- **Choose appropriate machine learning or deep learning algorithms for the task. Common choices include logistic regression, decision trees, random forests, support vector machines, and neural networks.**
- **Experiment with multiple models to compare their performance.**

Model Training:

- **Split the dataset into training and testing sets for model evaluation.**
- **Train the selected models using the training data.**
- **Tune hyperparameters to optimize model performance.**

Model Evaluation:

- **Evaluate the models using appropriate metrics such as accuracy, precision, recall, F1-score, and area under the ROC curve.**
- **Perform cross-validation to ensure the model's generalization ability.**

Interpretability:

- **Ensure the model provides interpretable results by explaining its predictions. Techniques like SHAP (SHapley Additive exPlanations) values and LIME (Local Interpretable Model-agnostic Explanations) can help with this.**

Deployment:

- **Develop a user-friendly interface (web app, mobile app, or API) for healthcare professionals and individuals to interact with the system.**
- **Ensure the deployment environment is secure, scalable, and compliant with privacy regulations (e.g., HIPAA).**

Description of the dataset used, data preprocessing steps, and feature extraction Techniques:

diabetes.csv

```
Pregnancies,Glucose,BloodPressure,SkinThickness,Insulin,BMI,DiabetesPedigreeFunc
tion,Age,Outcome
6,148,72,35,0,33.6,0.627,50,1
1,85,66,29,0,26.6,0.351,31,0
8,183,64,0,0,23.3,0.672,32,1
1,89,66,23,94,28.1,0.167,21,0
0,137,40,35,168,43.1,2.288,33,1
5,116,74,0,0,25.6,0.201,30,0
3,78,50,32,88,31,0.248,26,1
10,115,0,0,0,35.3,0.134,29,0
2,197,70,45,543,30.5,0.158,53,1
8,125,96,0,0,0,0.232,54,1
4,110,92,0,0,37.6,0.191,30,0
10,168,74,0,0,38,0.537,34,1
10,139,80,0,0,27.1,1.441,57,0
1,189,60,23,846,30.1,0.398,59,1
5,166,72,19,175,25.8,0.587,51,1
7,100,0,0,0,30,0.484,32,1
0,118,84,47,230,45.8,0.551,31,1
7,107,74,0,0,29.6,0.254,31,1
1,103,30,38,83,43.3,0.183,33,0
1,115,70,30,96,34.6,0.529,32,1
3,126,88,41,235,39.3,0.704,27,0
8,99,84,0,0,35.4,0.388,50,0
7,196,90,0,0,39.8,0.451,41,1
9,119,80,35,0,29,0.263,29,1
11,143,94,33,146,36.6,0.254,51,1
10,125,70,26,115,31.1,0.205,41,1
7,147,76,0,0,39.4,0.257,43,1
1,97,66,15,140,23.2,0.487,22,0
13,145,82,19,110,22.2,0.245,57,0
5,117,92,0,0,34.1,0.337,38,0
5,109,75,26,0,36,0.546,60,0
3,158,76,36,245,31.6,0.851,28,1
3,88,58,11,54,24.8,0.267,22,0
6,92,92,0,0,19.9,0.188,28,0
10,122,78,31,0,27.6,0.512,45,0
4,103,60,33,192,24,0.966,33,0
11,138,76,0,0,33.2,0.42,35,0
9,102,76,37,0,32.9,0.665,46,1
2,90,68,42,0,38.2,0.503,27,1
4,111,72,47,207,37.1,1.39,56,1
3,180,64,25,70,34,0.271,26,0
7,133,84,0,0,40.2,0.696,37,0
7,106,92,18,0,22.7,0.235,48,0
9,171,110,24,240,45.4,0.721,54,1
7,159,64,0,0,27.4,0.294,40,0
0,180,66,39,0,42,1.893,25,1
1,146,56,0,0,29.7,0.564,29,0
2,71,70,27,0,28,0.586,22,0
7,103,66,32,0,39.1,0.344,31,1
7,105,0,0,0,0,0.305,24,0
1,103,80,11,82,19.4,0.491,22,0
1,101,50,15,36,24.2,0.526,26,0
5,88,66,21,23,24.4,0.342,30,0
8,176,90,34,300,33.7,0.467,58,1
7,150,66,42,342,34.7,0.718,42,0
1,73,50,10,0,23,0.248,21,0
7,187,68,39,304,37.7,0.254,41,1
0,100,88,60,110,46.8,0.962,31,0
0,146,82,0,0,40.5,1.781,44,0
0,105,64,41,142,41.5,0.173,22,0
2,84,0,0,0,0,0.304,21,0
8,133,72,0,0,32.9,0.27,39,1
```

Dataset Description:

The document used for that model training is the dataset from kaggle "diabetes.csv" which includes age, gender, family history, body mass index (BMI), blood pressure, glucose levels .

Data Preprocessing:

- **Clean the dataset by handling missing values and outliers.**
- **Normalize or standardize features to ensure they are on a common scale.**
- **Encode categorical variables if necessary (e.g., one-hot encoding for gender).**

Feature Selection:

- **Use feature selection techniques to identify the most relevant features for diabetes prediction. This can help improve the model's efficiency and interpretability**

Explaination of the choice of machine learning algorithm, model training, and evaluation metrics:

Ai Model

- **Logistic Regression:**

  Logistic Regression is a widely used model for binary classification problems, such as predicting whether an individual will develop diabetes (yes or no). It's a simple yet effective algorithm that provides interpretable results. In logistic regression, you model the relationship between the features (e.g., age, BMI, blood pressure) and the probability of a person developing diabetes.

Machine Learning Algorithm:

- **Random forest:**

  Random Forest is an ensemble learning method that combines multiple decision trees to make predictions. It's a popular choice for binary classification tasks like diabetes prediction. Here's how you can use Random Forest for training and evaluation:

- **Hyper parameter tuning:**

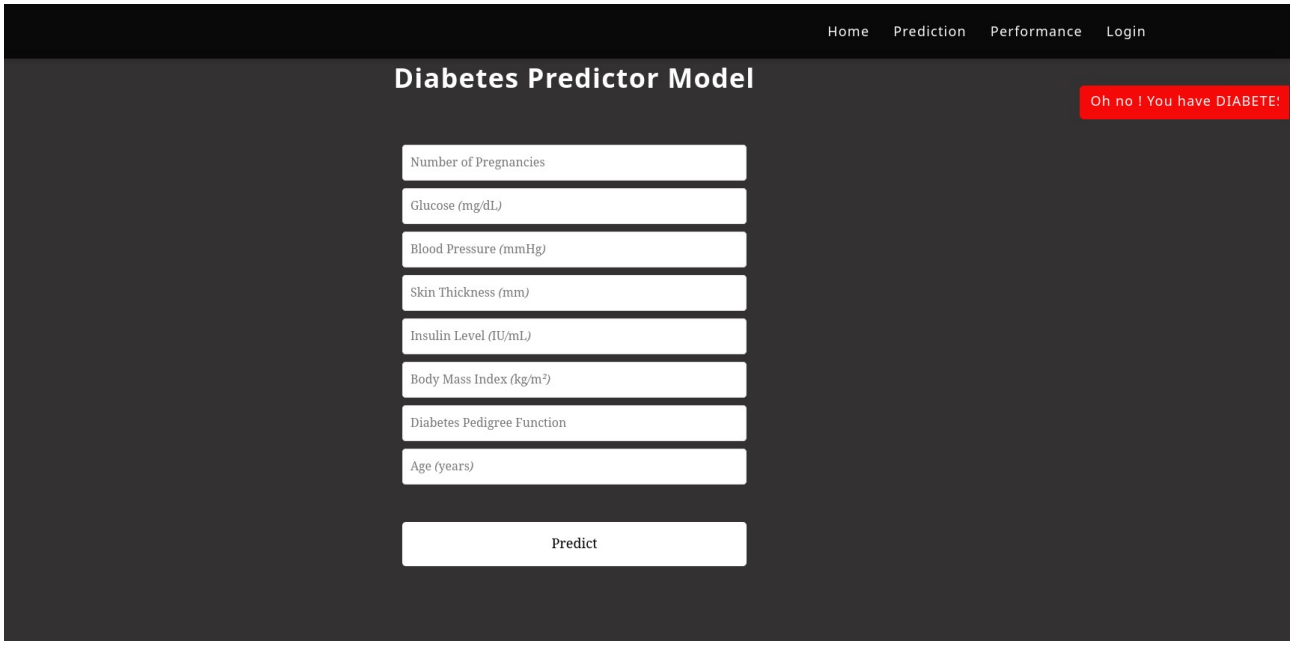
  Logistic Regression is a widely used model for binary classification problems, such as predicting whether an individual will develop diabetes (yes or no). It's a simple yet effective algorithm that provides interpretable results. In logistic regression, you model the relationship between the features (e.g., age, BMI, blood pressure) and the probability of a person developing diabetes.

Performance evaluation:

- **Accuracy for my model is just above 75% .**
- **It predicts the result using the given dataset with high efficiency.**
- **It measures the proprtion of the data value and classified as a diabetic or not diabetic.**
- **A harmonic mean of precision and recall, balancing both metrices.**

OUTPUT :

Diabetic result :



Not diabetic :