**AI BASED DIABETES PREDICTION SYSTEM**

Phase 3: Submission

**Project:** Diabetes prediction system



**Introduction:**

* An AI-based diabetes prediction system is a sophisticated healthcare application that employs artificial intelligence (AI) and machine learning (ML) techniques to analyze and interpret diverse sets of data to predict the likelihood of an individual developing diabetes.
* Step into the future of healthcare with our AI-based Diabetes Prediction System. This innovative solution leverages the prowess of artificial intelligence to analyze comprehensive datasets, ranging from medical history to lifestyle factors.
* Through advanced machine learning algorithms, it accurately predicts the likelihood of diabetes onset, empowering individuals and healthcare professionals with proactive insights for early intervention and personalized preventive measures.

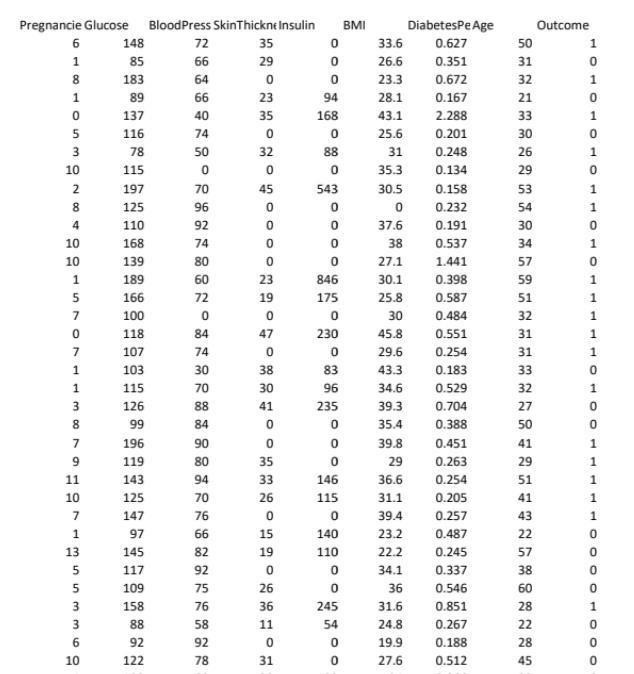
# Phase 3: Development Part 1

In this part you will begin building your project by loading and preprocessing the dataset.Start building the AI based diabetes prediction model by loading and preprocessing the dataset.Load the diabetes prediction dataset and preprocess the data.

## Data Source

A good data source for diabetes prediction using machine learning should be Accurate whether the person has diabetes or not.

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



**Import Libraries:**

Import the necessary libraries for your project. You'll likely need libraries such as Pandas and NumPy and seaborn.

Source Code:

*import pandas as pd*

*import numpy as np*

*importmatplotlib.pyploaslt*

*import seaborn as sns*

**Load the dataset:**

Load your diabetes prediction dataset into a Pandas Data Frame. You can typically load data from a CSV file using the pd.read\_csv() function .

Source Code:

df=pd.read\_csv('../input/pimab indians diabetes database/diabetes.csv')

df.head()

**Explore the data:**

Take a look at the data to understand its structure and content. You can use functions like data.head(), data.info(), and data.describe() to get an overview.

Source Code:

# Display the first few rows of the dataset print(data.head())

# Get information about the dataset print(data.info())

# Summary statistics of the data print(data.describe())

Data Preprocessing:

* Handling Missing Data:

Check for missing values in the dataset and decide how to handle them (e.g., by filling missing values with the mean, median, or using other techniques).

Source Code:

# Handle missing values (if any)

data.fillna(method='ffill', inplace=True)

# Example: Forward fill missing values

* Categorical Data:

If your dataset contains categorical variables, you may need to encode them using techniques like one-hot encoding.

Source Code:

# Example one-hot encoding

data = pd.get\_dummies(data, columns=['categorical\_column'])

* Scaling:

Normalize or standardize numerical features to ensure that they are on a similar scale.

Source Code:

from sklearn.preprocessing import StandardScaler scaler = StandardScaler()

data[['numerical\_column1', 'numerical\_column2']] =

scaler.fit\_transform(data[['numerical\_column1', 'numerical\_column2']])

**Conclusion and Future Work:**

We have loaded and preprocessed the dataset. Next step we can proceed to build the model of diabetes prediction system in the next phase of the project.

**Project Conclusion:**

Using Random forest algorithm in our prediction system, we can evaluate the performance using the accuracy score, comparing the performance between train and test data and produce accurate prediction values of whether a patient has diabetes or not.