AI Based Diabetes Prediction System

Project summary:

This project is based on Artificial intelligence that uses machine learning models to train using the dataset and produce results on whether the person has diabetes or not.

Contents:

* Introduction
* Problem statement
* Problem definition
* Diabetes Mellitus
* Objective
* Design Thinking Process
* Innovations
* Loading and preprocessing
* Model training and Evaluation
* Conclusion

Introduction:

A diabetes prediction system is an advanced application that uses data analysis and machine learning to estimate the risk of an individual developing diabetes. By analyzing various factors such as medical history, lifestyle, and biomarkers, it forecasts the likelihood of diabetes onset within a certain period. The system aims to enable early identification, personalized intervention, and proactive management to potentially prevent or better manage diabetes, offering significant potential in healthcare for improving outcomes and reducing the burden of the condition.

Problem statement:

Develop an AI-powered diabetes prediction system that leverages machine learning algorithms to analyze medical data and predict the likelihood of an individual developing diabetes, providing early risk assessment and personalized preventive measures.

Problem definition:

The problem is to build an AI-powered diabetes prediction system that uses machine learning algorithms to analyze medical data and predict the likelihood of an individual developing diabetes. The system aims to provide early risk assessment and personalized preventive measures, allowing individuals to take proactive actions to manage their health.

Diabetes Mellitus:

Diabetes Mellitus (DM), commonly known as diabetes, is a group of metabolic disorders characterized by high blood sugar levels over a prolonged period. Diabetes is due to either the pancreas not producing enough insulin, or the cells of the body not responding properly to the insulin produced. There are three main types of diabetes mellitus:

* Type 1 diabetes results from the pancreas's failure to produce enough insulin due to loss of beta cells. This form was previously referred to as "insulin-dependent diabetes mellitus" (IDDM) or "juvenile diabetes". The cause is unknown.
* Type 2 diabetes begins with insulin resistance, a condition in which cells fail to respond to insulin properly. As the disease progresses, a lack of insulin may also develop. This form was previously referred to as "non insulin-dependent diabetes mellitus" (NIDDM) or "adult-onset diabetes". The most common cause is a combination of excessive body weight and insufficient exercise.
* Gestational diabetes is the third main form, and occurs when pregnant women without a previous history of diabetes develop high blood sugar levels.

OBJECTIVE:

To build a machine learning model to accurately predict whether the patients in the dataset have diabetes or not.

Data Collection:

The system collects a comprehensive set of input data, Including the patient’s medical history, demographic information, Lifestyle factors (such as diet and physical activity), and genetic Predisposition if available.

Data Preprocessing:

Advanced data preprocessing techniques are employed to Clean, normalize, and transform the input data into a suitable format For analysis. Missing data is handled appropriately to ensure the Accuracy of predictions.

Machine Learning Models:

State-of-the-art machine learning algorithms, including logistic Regression, decision trees, random forests, support vector machines, And neural networks, are utilized to build predictive models. These models are trained on a diverse dataset containing information from Diabetic and non-diabetic individuals.

Model Training and Validation:

Models are trained using historical data, and their Performance is rigorously validated through cross-validation Techniques to ensure their accuracy, sensitivity, and specificity.

Model deployment:

Deploying the AI model is various fields to get the use of this Model fully.

INNOVATIONS IN DIABETES PREDICTION SYSTEM:

Personalized Risk Assessment:

The system calculates a personalized risk score for each user, Indicating their likelihood of developing diabetes in the future. This Score is based on the individual’s specific data and health profile.

Continuous Monitoring:

For individuals with elevated risk scores or diagnosed with prediabetes, the system offers continuous monitoring and feedback, relevant health metrics over time.

Privacy and Security:

Strict privacy and security measures are in place to protect sensitive medical and personal information, ensuring healthcare regulations.

Accessibility:

The system is designed to be accessible via web and mobile platforms, making it convenient for users to access their health information and recommendations on the go.

Integration of Wearable Devices and IoT:

Incorporating wearable devices such as continuous glucose monitors, fitness trackers, and smartwatches can provide real-time data for a more comprehensive understanding of an individual’s health. These devices can monitor blood glucose levels, physical activity, sleep patterns, and other relevant parameters, feeding this data into the prediction system.

Artificial Intelligence and Deep Learning:

Advancements in deep learning and artificial intelligence Techniques, including convolutional neural networks (CNNs) and Recurrent neural networks (RNNs), can help analyze complex, high-dimensional data to identify subtle patterns and trends. These models can be used to improve the accuracy of predictions.

Genomic Data Integration:

Genetic information plays a significant role in diabetes risk. Innovations in genomic sequencing and analysis can allow the Integration of genetic data into prediction models. This can help Identify individuals with a higher genetic predisposition to diabetes.

Explainable AI:

To build trust and transparency, incorporating explainable AI Techniques can help users understand why a particular prediction was made. This is especially important in healthcare, where decisions can have a profound impact on a person’s life.

Mobile Health Apps:

Mobile apps can serve as a platform for individuals to input their health data, receive personalized recommendations, and track their progress. These apps can also integrate features like reminders, gamification, and social support to encourage healthy behavior.

Personalized Nutrition and Lifestyle Recommendations:

Innovations in nutrition science and personalized medicine can enable the system to provide highly tailored dietary and lifestyle recommendations based on an individual's unique health profile.

LOADING AND PREPROCESSING THE DATASET

Importing necessary libraries:

The development of a diabetes prediction system in Python requires the use of a variety of libraries and tools to handle data, build machine learning models, and create a user-friendly interface. The tools and libraries are:

NumPy:

NumPy is the fundamental library for scientific computing in Python. It provides support for large, multi-dimensional arrays and matrices, along with mathematical functions to operate on these arrays. For diabetes prediction, NumPy is essential for data manipulation and mathematical operations.

Pandas:

Pandas is foundational library that provides data structures and data analysis tools. It is used to handle and manipulate structured data, such as datasets containing patient information, medical records, and other relevant information. I can use Pandas to clean and preprocess the data for my prediction model.

Scikit-Learn:

Scikit-Learn is one of the most popular machine learning libraries in Python. It offers a wide range of tools for tasks like classification, regression, clustering, model selection, and dimensionality reduction. Scikit-Learn is particularly valuable for building and evaluating machine learning models to predict diabetes based on patient data.

Code

import numpy as np

import pandas as pd

from sklearn.preprocessing import StandardScaler

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

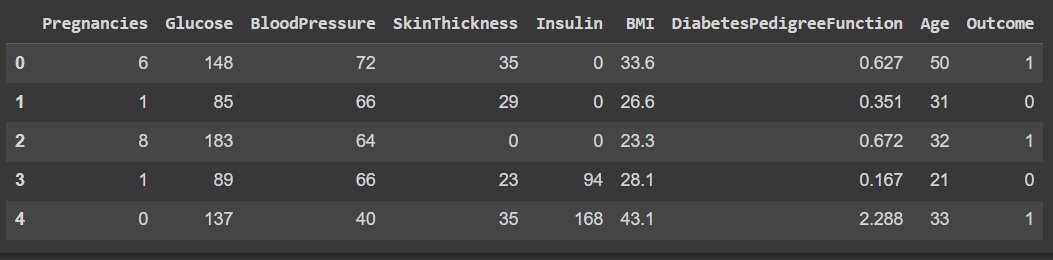
from sklearn import svm

from sklearn.metrics import accuracy\_score

Loading the dataset to the dataframe using the pandas library:

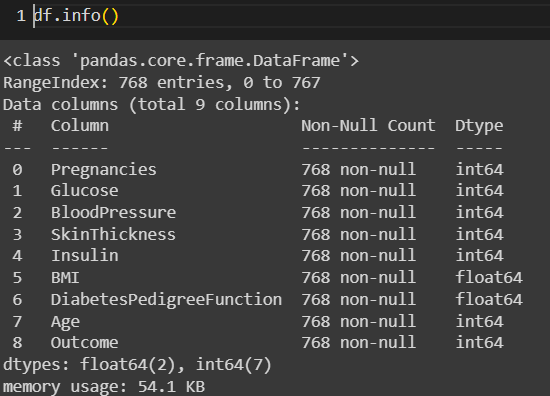
df=pd.read\_csv("/content/diabetes.csv") #loading the dataset

df.head()



DATA EXPLORATION:

Basic information about the dataset:

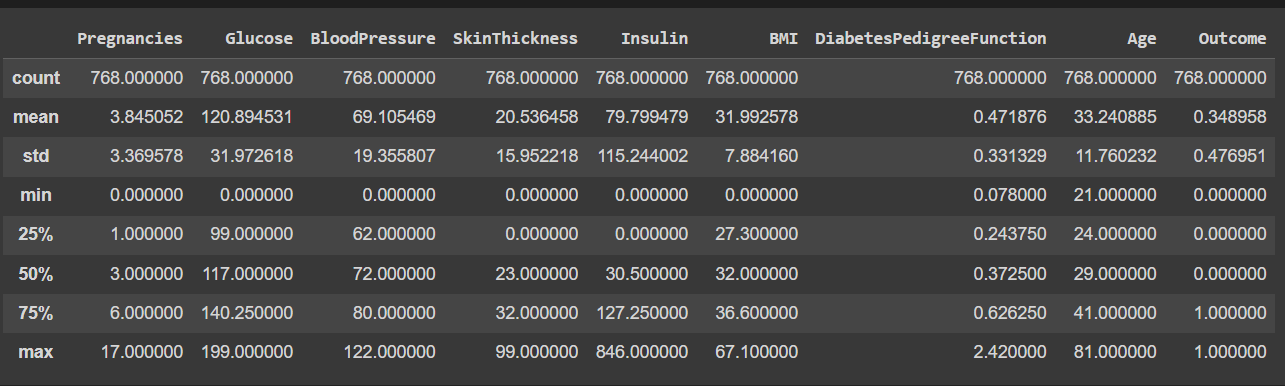
df.info()

Getting dimensions of the dataset:

df.shape



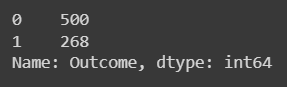
Statistical measures of data:

df.describe()

Analysing outcome:

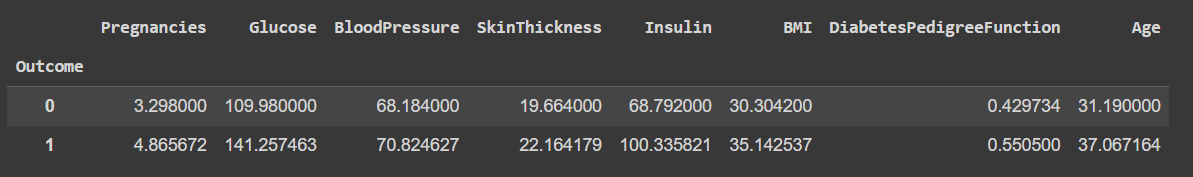
df['Outcome'].value\_counts()

# 0--> non-diabetic people

# 1--> diabetic people

Checking the description of the ‘Outcome’:

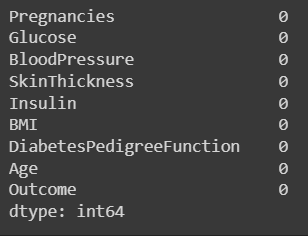
df.groupby('Outcome').mean()



Checking missing values:

missing\_values = df.isnull().sum()

print(missing\_values)



Filling missing values:

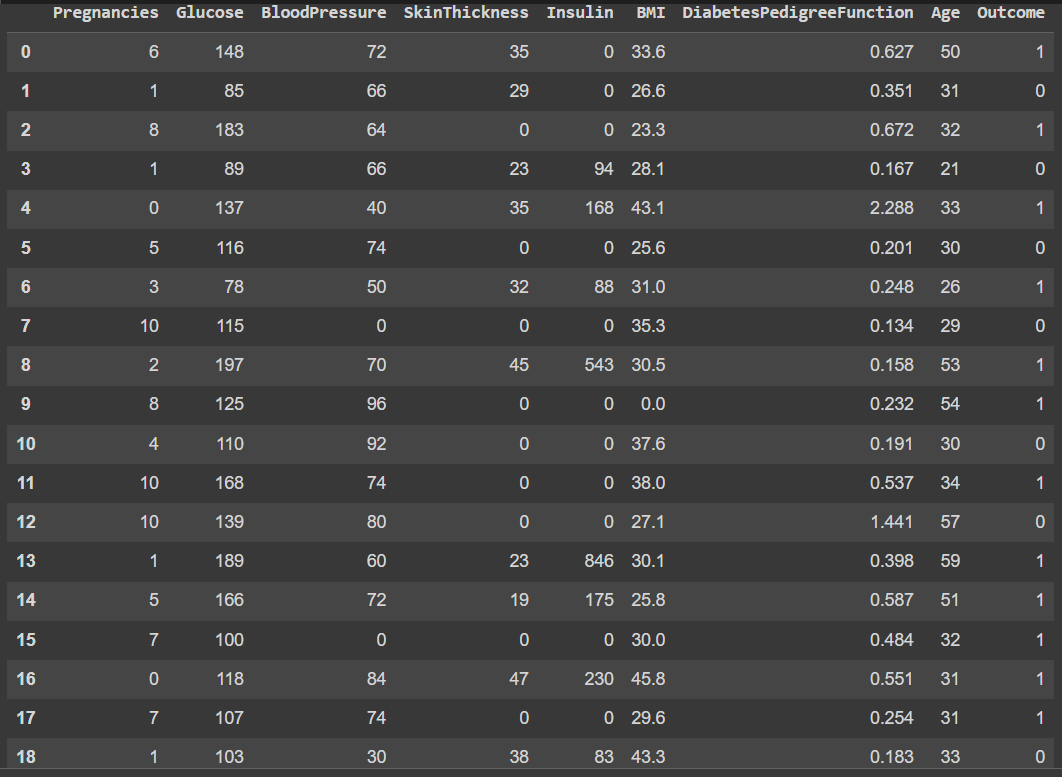
mean\_fill = df.fillna(df.mean())

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

Checking duplicate values and deleting them:

duplicate\_values = df.duplicated().sum()

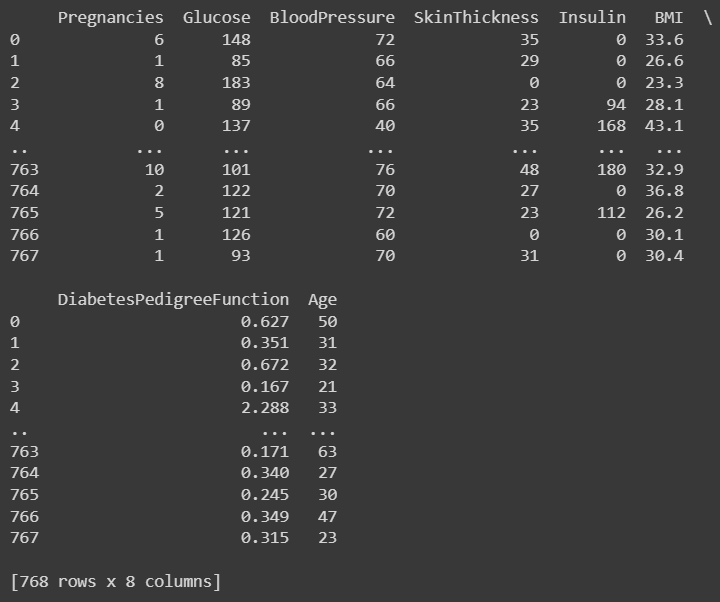
df.drop\_duplicates(inplace=True)

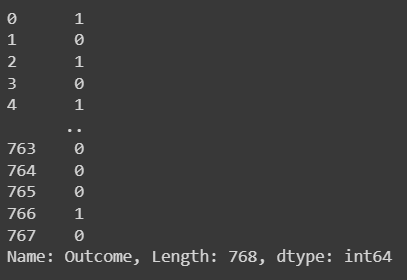
df.head(20)

Separating the data and labels:

X = df.drop(columns='Outcome', axis=1)

Y = df['Outcome']

print(X)

print(Y)

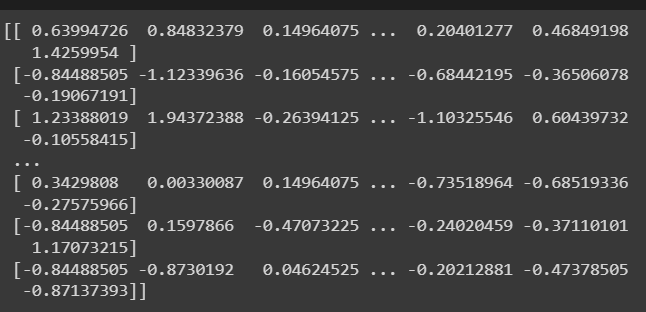
Data Standardization:

Data standardization, also known as data scaling or data normalization, is a crucial preprocessing step in machine learning. It involves transforming the features (variables) in our dataset into a standard format or scale to ensure that the machine learning algorithms can work effectively and make meaningful predictions.

scale = StandardScaler()

scale.fit(X)

sd\_data = scale.transform(X)

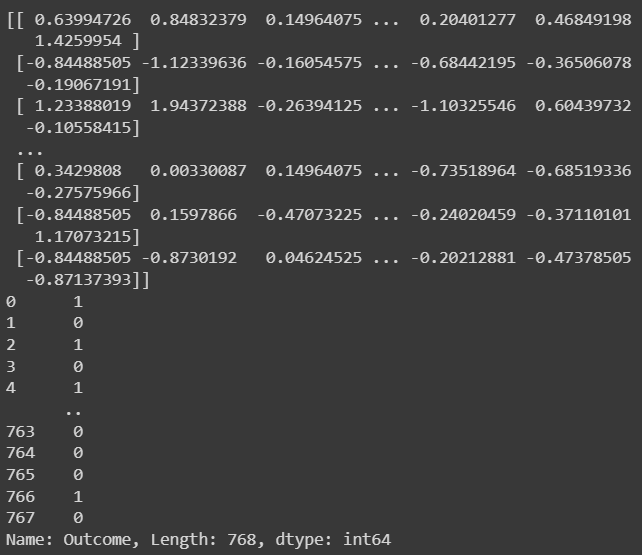
print(sd\_data)

X = sd\_data

Y= df['Outcome']

print(X)

print(Y)



BUILDING THE MODEL

Splitting data for training:

The splitting of the dataset using the train\_test\_split will split the dataset into 4 parts

The train – X which is the input for the model to train

The train – Y which is the input for the model to train

Y,X-test which are used to test the model and increasing the efficiency.

20% of the dataset are used for testing and stratify will split the data according to their columns

X\_train,X\_test,Y\_train,Y\_test = train\_test\_split(X,Y, test\_size=0.2, stratify = Y, random\_state=2)

Dimensions of the train and test dataset:

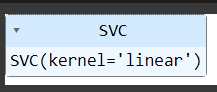
print(X.shape, X\_train.shape, Y\_test.shape)

Choosing the training model and loading it to the classifier:

A classifier is a type of algorithm used to categorize or label objects into different classes or categories.

classifier = svm.SVC(kernel= 'linear')

classifier.fit(X\_train,Y\_train)



Accuracy score on the given training data

The accuracy score is a commonly used metric to evaluate the performance of a classification model. It measures the proportion of correctly predicted instances (samples or observations) out of the total number of instances.

X\_train\_prediction = classifier.predict(X\_train)

training\_model\_accuracy = accuracy\_score(X\_train\_prediction, Y\_train)

print(training\_model\_accuracy)

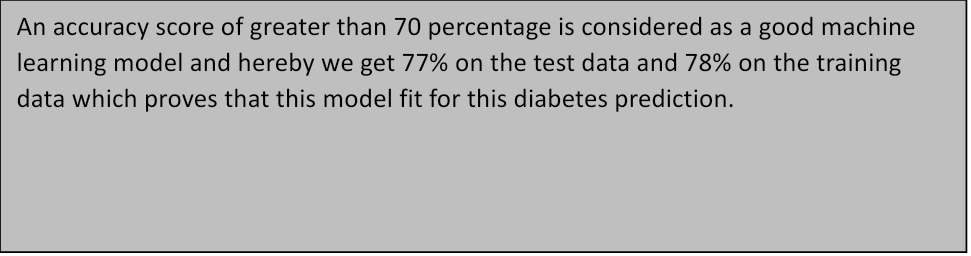


Accuracy score on the test data:

X\_test\_prediction = classifier.predict(X\_test)

test\_data\_accuracy= accuracy\_score(X\_test\_prediction, Y\_test

print(test\_data\_accuracy)



TESTING THE MODEL

Testing with custom input:

input\_data = (10, 125, 70, 26, 115, 31.1, 0.205, 41)

Transformed to numpy array

data\_array = np.asarray(input\_data)

Reshape array for one instance

reshaped\_data\_input = data\_array.reshape(1,-1)

Standardization

Since we use standardized input for training we need to use the standardized input for testing the model

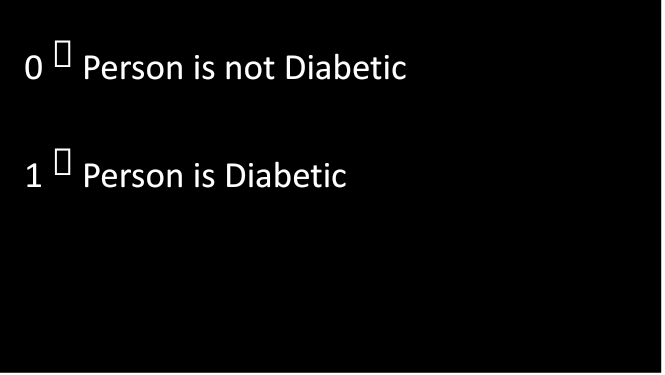
std\_input\_data= scale.transform(reshaped\_data\_input)

print(std\_input\_data)

prediction =  classifier.predict(std\_input\_data)

print(prediction)





EVALUATING THE MODEL

if(prediction[0] == 0):

  print("The person is not diabetic")

else:

  print("The person is diabetic")

The linear regression based machine learning model  is best fit for the  given dataset and gives us 78 percentage accuracy.  And also passed the  custom input.

Conclusion:

The diabetes prediction system represents a significant step towards proactive and personalized healthcare, empowering individuals to take preventive measures and make informed lifestyle choices. Ultimately, the aim is to create a future where the burden of diabetes can be alleviated through early identification and intervention, positively impacting the quality of life for individuals at risk.