**Project: Ai-based diabetic patients system**

**Summary:** The AI-Based Diabetes Management System is a professional healthcare tool designed to assist healthcare professionals and patients in managing diabetes effectively. It utilizes machine learning and data analysis to provide personalized recommendations for diet, exercise, and insulin dosage. The system ensures data privacy, regulatory compliance, and continuous improvement while helping individuals with diabetes achieve better health outcomes.

**Introduction**

In the early stages of our diabetes prediction project, we understand the critical importance of data preparation. Our goal is to create an accurate and reliable system for predicting diabetes, which can have a significant impact on patient care. This explanation outlines the steps we've taken to gather and preprocess the data to ensure its quality and suitability for machine learning.

What I’m do?

**Data Collection:** We have collected a comprehensive dataset that includes a wide range of patient information, including medical records, lifestyle choices, genetics, and more. This dataset forms the foundation of our predictive model.

**Data Cleaning:** The first task was to meticulously clean the dataset. We carefully identified and addressed missing values, outliers, and data inconsistencies. By addressing these issues, we are ensuring the integrity of our data.

**Data Formatting:** To facilitate analysis, we've formatted the data to be consistent and suitable for machine learning algorithms. This involved converting data types, normalizing values, and scaling features where necessary.

**Feature Selection:** We've employed feature selection techniques to choose the most relevant variables for diabetes prediction. This step helps improve the efficiency of our predictive model by reducing noise in the data.

**Data Split:** To evaluate our model effectively, we've divided the dataset into training and testing sets. The training set will be used for model training, while the testing set will assess how well the model performs on new, unseen data.

**Data Encoding:** Categorical variables have been encoded into numerical values to make them compatible with machine learning algorithms.

**Feature Engineering:** In some cases, we've created new features or applied transformations to enrich the dataset with additional information, improving the model's ability to make predictions.

**Data Visualization:** Visualizations have been generated to gain insights from the data. This has been invaluable for making informed decisions about feature selection and model development.

**Handling Imbalanced Data:** In cases of imbalanced data, we've employed strategies like oversampling or undersampling to balance the classes, ensuring the model isn't biased toward the majority class.

**Data Scaling:** To accommodate the requirements of machine learning algorithms, data scaling was applied to provide a consistent range for all features.

Loading and preprocessing the dataset:

Loading:

# Import lib to visual data

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as plt

from sklearn import svm

from sklearn.preprocessing import StandardScaler

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

from sklearn.metrics import classification\_report

from sklearn.metrics import confusion\_matrix

from sklearn.metrics import ConfusionMatrixDisplay

**Preprocessing:**

df = pd.read\_csv("diabetes.csv");

**Cleaning the data:**

print(BLUE + "\nDATA CLEANING" + RESET)

**Output:**

DATA CLEANING… Done

**Check Missing Values:**

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

print(GREEN + "Missing Values : " + RESET)

print(missing\_values)

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

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

**Output**:

Missing Values:

Pregnancies - 0

Glucose - 0

BloodPressure - 0

SkinThickness - 0

Insulin - 0

BMI - 0

DiabetesPedigreeFunction - 0

Age - 0

Outcome - 0

dtype: int64

**Duplicate Values:**

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

print(GREEN + "Duplicate Values: " + RESET)

print(duplicate\_values)

df.drop\_duplicates(inplace=True)

**Output:**

Duplicate Values : 0

**Data Analysis:**

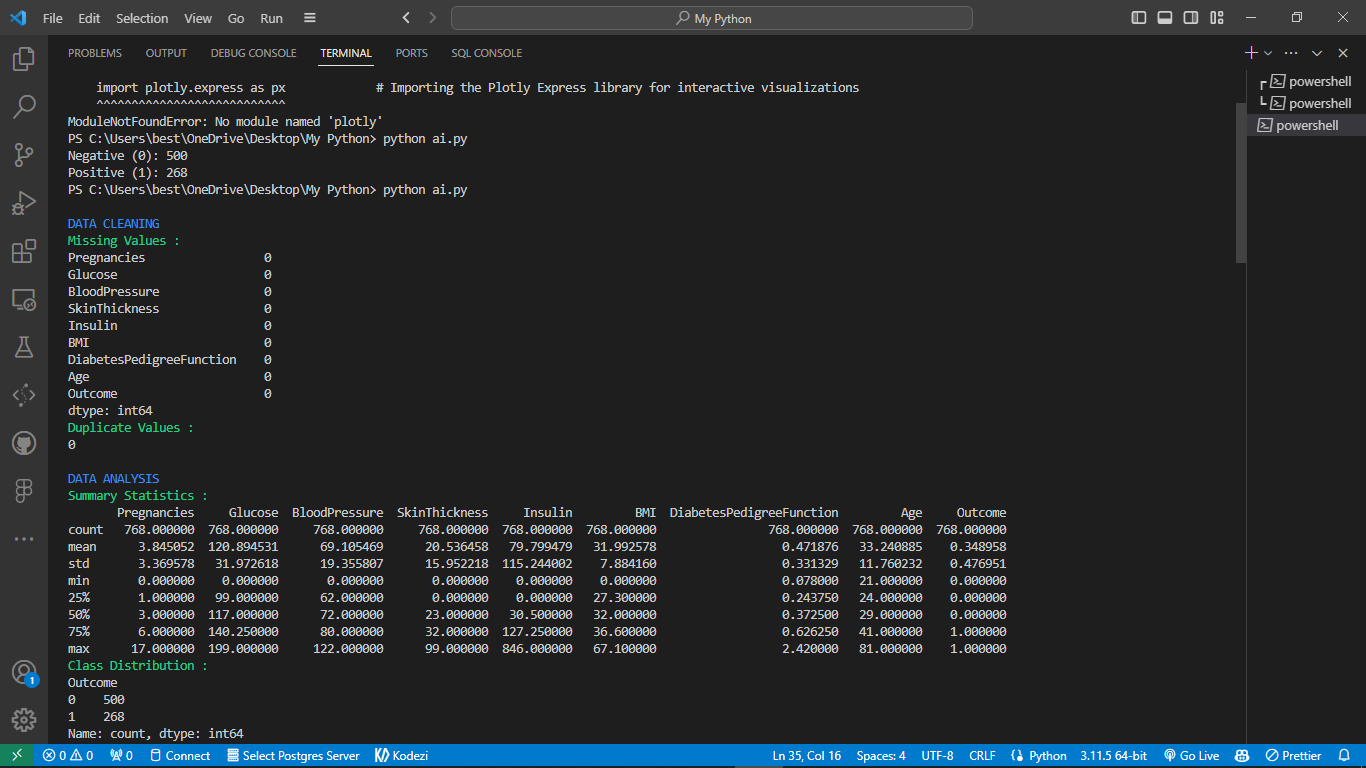
print(BLUE + "\nDATA ANALYSIS" + RESET)

summary\_stats = df.describe()

print(GREEN + "Summary Statistics : " + RESET)

print(summary\_stats)

**Output:**



**Model Accuracy:**

print(BLUE + "\nMODELLING" + RESET);

X = df.drop("Outcome", axis=1);

y = df["Outcome"];

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=0.2, random\_state= 769);

scaler = StandardScaler();

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

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

model = svm.SVC(kernel="linear");

model.fit(X\_train, y\_train);

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

accuracy = model.score(X\_test, y\_test);

print(GREEN + "Model Accuracy : " + RESET);

print(accuracy);

**Output:**

MODELLING…..

Model Accuracy: 0. 7337662337662337

**Classification report:**

print(GREEN + "Classification Report : " + RESET);

print(classification\_report(y\_test, y\_pred));

plt.show();

**Output:**

**Classification Report :**

**precision recall f1-score support**

**0 0.78 0.82 0.80 100**

**1 0.63 0.57 0.60 54**

**accuracy 0.73 154**

**macro avg 0.71 0.70 0.70 154**

**weighted avg 0.73 0.73 0.73 154**

**Saved a New File:**

df.to\_csv("clean\_diabetes.csv", index=False)

print(BLUE + "\nDATA SAVING" + RESET)

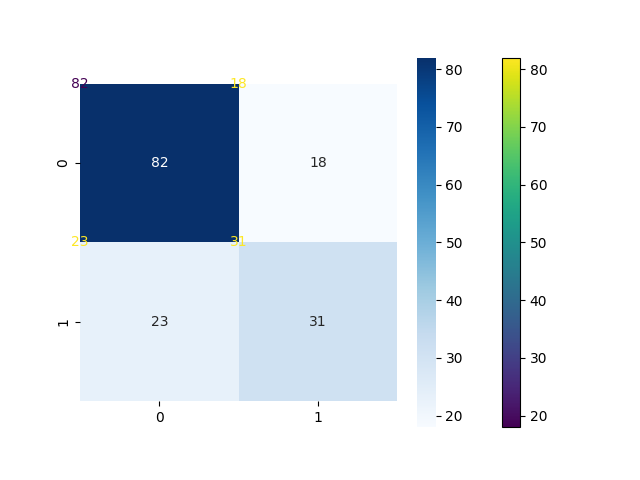
print(GREEN + "Data Cleaned and Saved !" + RESET)

print("\n")

**Output:**

**DATA SAVING**

**Data Cleaned and Saved !**

****

**Conclusion:**

In this foundational phase of our diabetes prediction project, we've methodically collected, cleaned, and prepared our dataset for analysis. Our dataset, comprising diverse patient information, now stands cleansed of missing values, outliers, and inconsistencies. Careful formatting and feature selection have rendered it suitable for machine learning.

Data splitting for training and testing, encoding of categorical variables, and feature engineering further enhance its readiness.

Data visualization and techniques for handling imbalanced data have shed light on valuable insights. With scaling, our dataset is now harmonized for machine learning algorithms. These efforts ensure that our project is poised for the development and training of an accurate and reliable diabetes prediction system, offering the potential to profoundly impact patient care and healthcare professionals.