**TEAM 16: ARTIFICIAL INTELLIGENCE**

**PROJECT TITLE:** **DIABETES PREDICTION SYSTEM**

**PHASE V: Project Submission**

**DESCRIPTION:**

A diabetes prediction system is a technology or model that uses data analysis and artificial intelligence (AI) techniques to predict the risk of developing diabetes in individuals or to help manage the condition in those who are already diagnosed. Such systems are valuable for early intervention and personalized care. Here's an overview of how a diabetes prediction system typically works in below

**Problem Statement**:

Developing an AI-driven diabetes prediction system that utilizes machine learning and patient data analysis to accurately forecast the risk of diabetes onset in individuals. The system's objective is to enhance proactive healthcare by identifying patterns, risk factors, and early indicators, thereby enabling personalized interventions and timely preventative measures to reduce the incidence and impact of diabetes on individuals' health.

**Platform used for implementing:**

We have to use **Jupyter Notebook** for implementing the Machine Learning Algorithm. It has default libraries installed in it. It basically run on a Python program.

**Machine Learning Algorithms Used**:

In this diabetes prediction we have used as Regression and classification algorithm. In that we have to used a matplotlib library to describe the dataset in graphic model

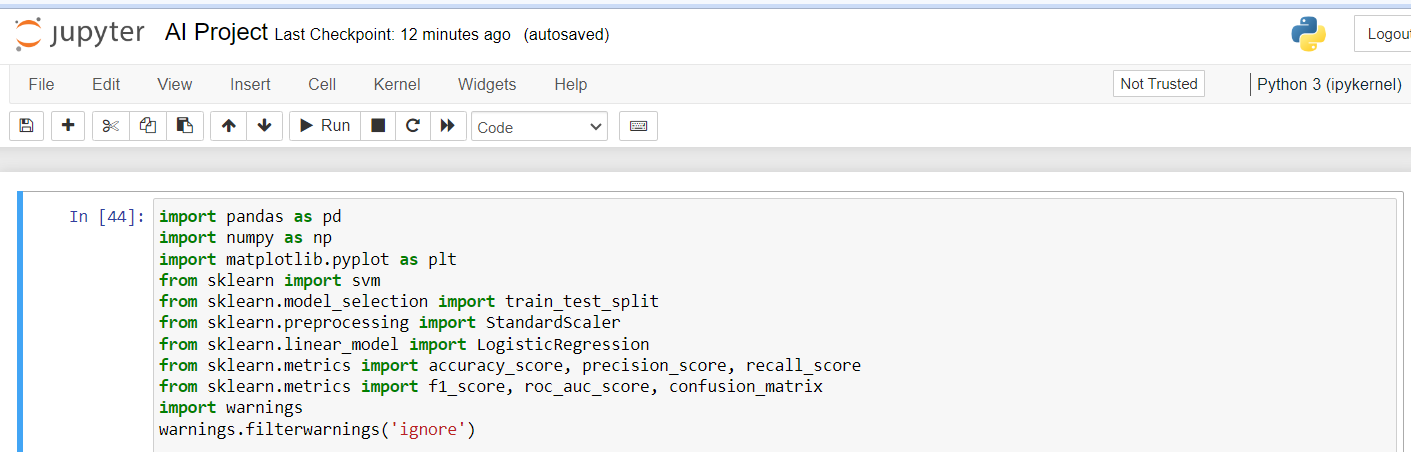
**Program:**

1. **Import libraries:**

**In Python, you use the import keyword to make code in one module available in another. Imports in Python are important for structuring your code effectively.**

Some important libraries known as in python is **Pandas, NumPy**

* **import pandas as pd**
* **from sklearn.model\_selection import train\_test\_split**
* **from sklearn.preprocessing import StandardScaler**
* **from sklearn.linear\_model import LogisticRegression**
* **from sklearn.metrics import accuracy\_score, precision\_score, recall\_score**
* **from sklearn.metrics import f1\_score, roc\_auc\_score, confusion\_matrix**

****

1. **Dataset:**

* Dataset is a collection of various types of data stored in a digital format
* Data is the key component of any Machine Learning project. Datasets primarily consist of images, texts, audio, videos, numerical data points, etc., for solving various Artificial Intelligence challenges such as. Image or video classification.
* Dataset contains the attributes, attributes in which it input from the user and gives output
* Basically the dataset in file format especially in CSV format.

We have to load the dataset into program by in a **csv format**

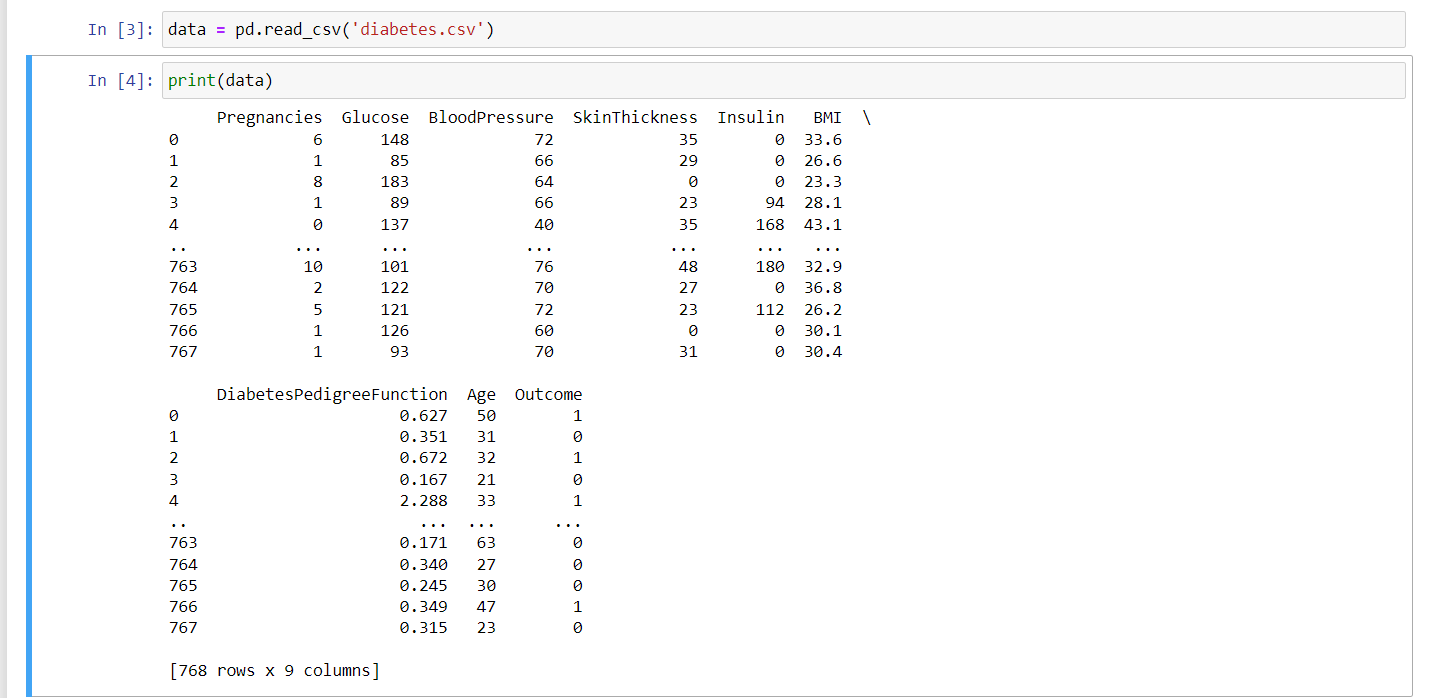
**Dataset Link:**[**https://www.kaggle.com/datasets/mathchi/diabetes-data-set**](https://www.kaggle.com/datasets/mathchi/diabetes-data-set)

* **Loading the dataset into a program by**

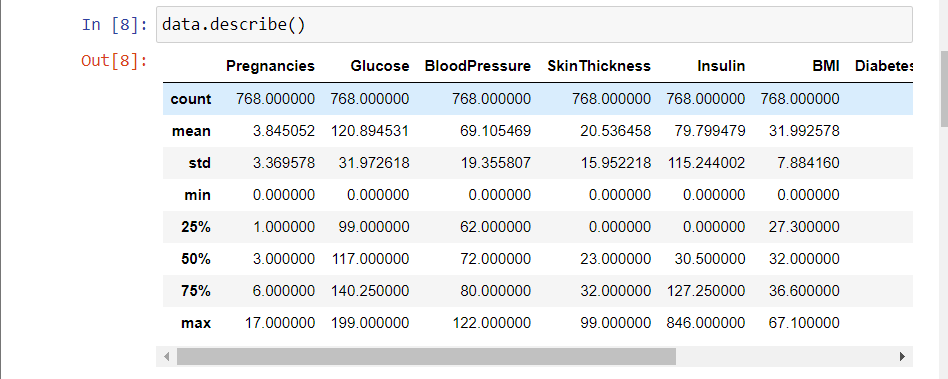
**data = pd.read\_csv("diabetes.csv")**

In our Diabetes prediction system dataset contains attributes such as

* **Pregnancies**
* **Glucose**
* **Blood pressure**
* **Skin Thickness**
* **Insulin**
* **BMI**
* **Diabetes Pedigree Function**
* **Age**
* **Outcome**

**EX:**

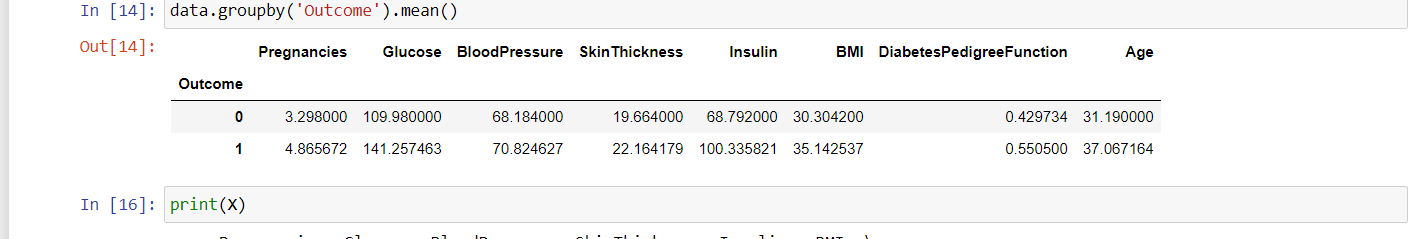
**We have to describe the data by describe() method:**

****

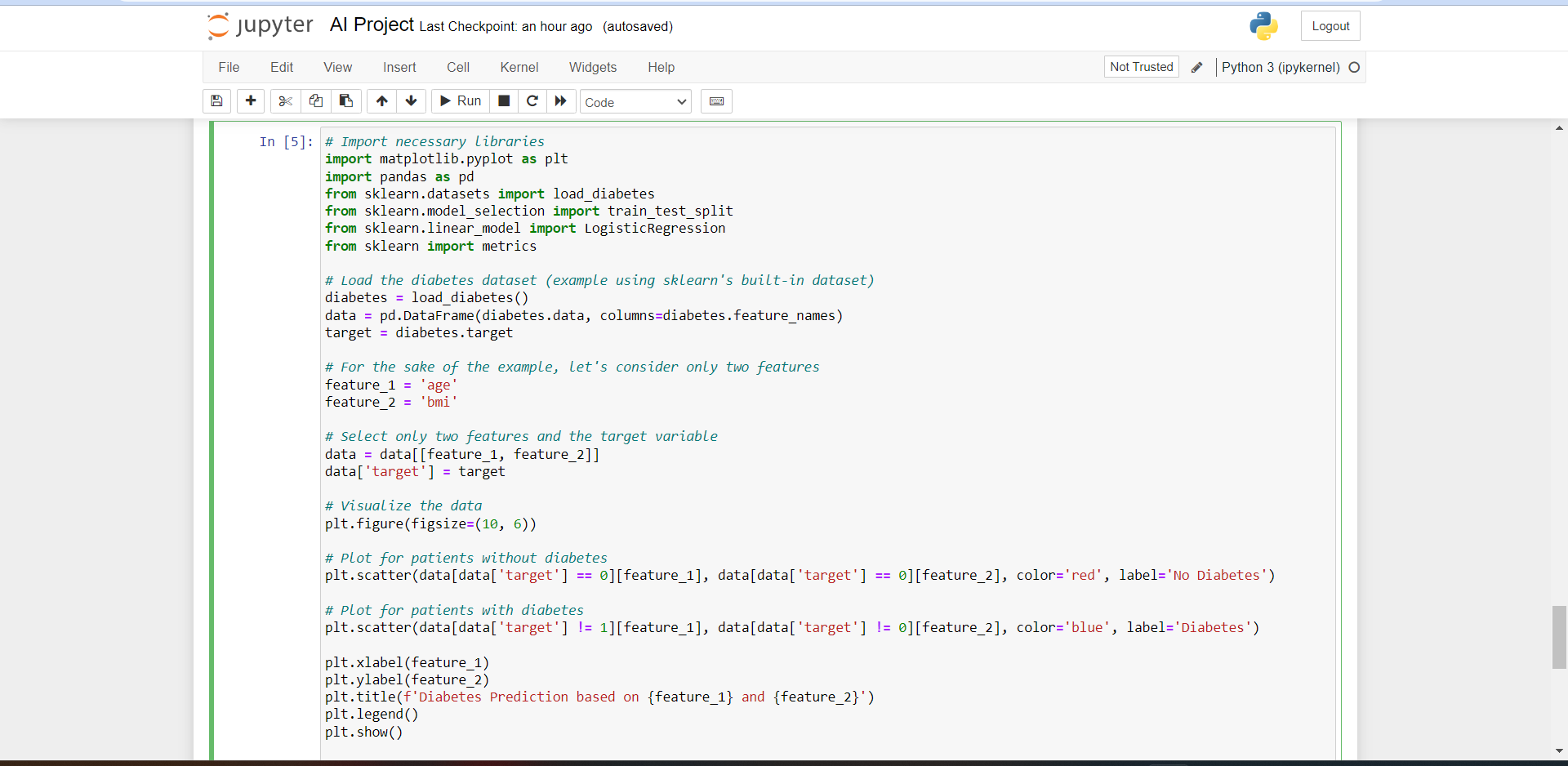
**Source of Dataset:**

We have to choose a dataset from Kaggle. Kaggle is an online community that have collection of data is used for building the model.

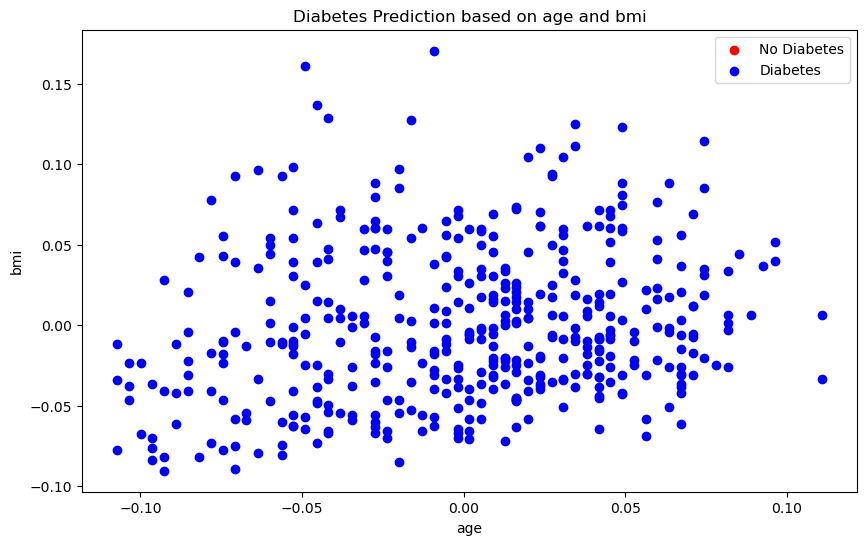
By training the model we have find the mean of value in dataset



**Describe Dataset in Graphic Model:**

****

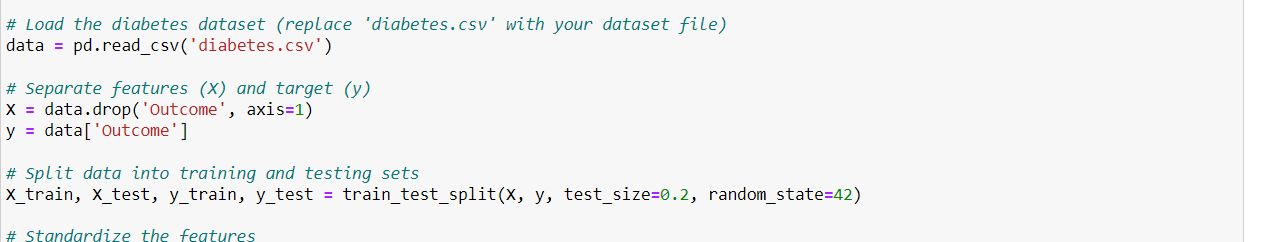
**O/p:**

****

1. **Data preprocessing:**

**It is used to handle missing values, feature selection, and feature engineering**

**By**

****

1. **Split the data into training and testing sets:**

****

1. **Evaluating Model Performance:**

**By evaluating the performance of the model by the following parameters**

* **Accuracy:**

The proportion of correctly predicted instances among the total instances

* **Precision:**

The ratio of true positive predictions to the total positive predictions, focusing on the accuracy of positive predictions.

* **F1\_score:**

The harmonic mean of precision and recall, providing a balance between precision and recall**.**

* **Recall:**

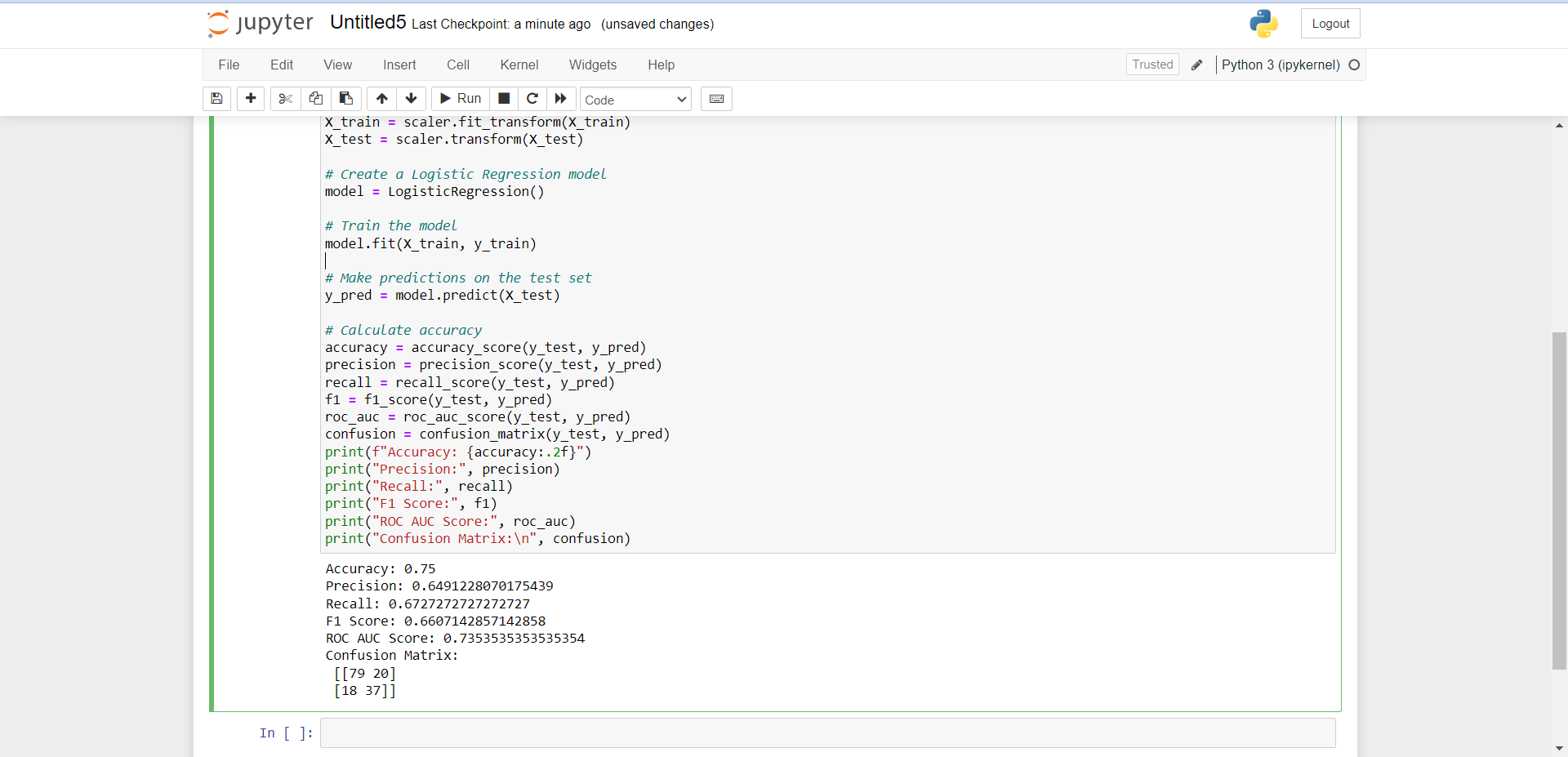
The ratio of true positive predictions to the actual positives, focusing on how many actual positives were captured by the model.

* **ROC AUC: The ROC AUC (Receiver Operating Characteristic Area Under the Curve):**

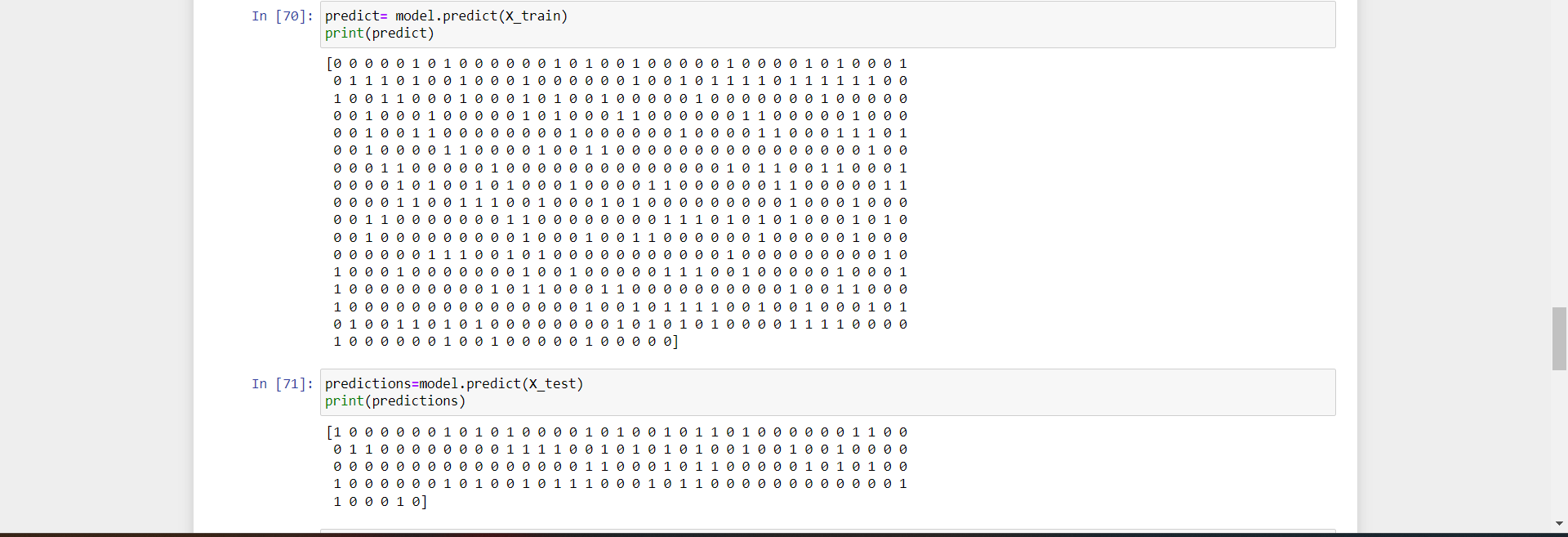
Receiver Operating Characteristic curve and Area Under the Curve help assess the trade-off between true positive rate and false positive rate.

* **Confusion Matrix:**

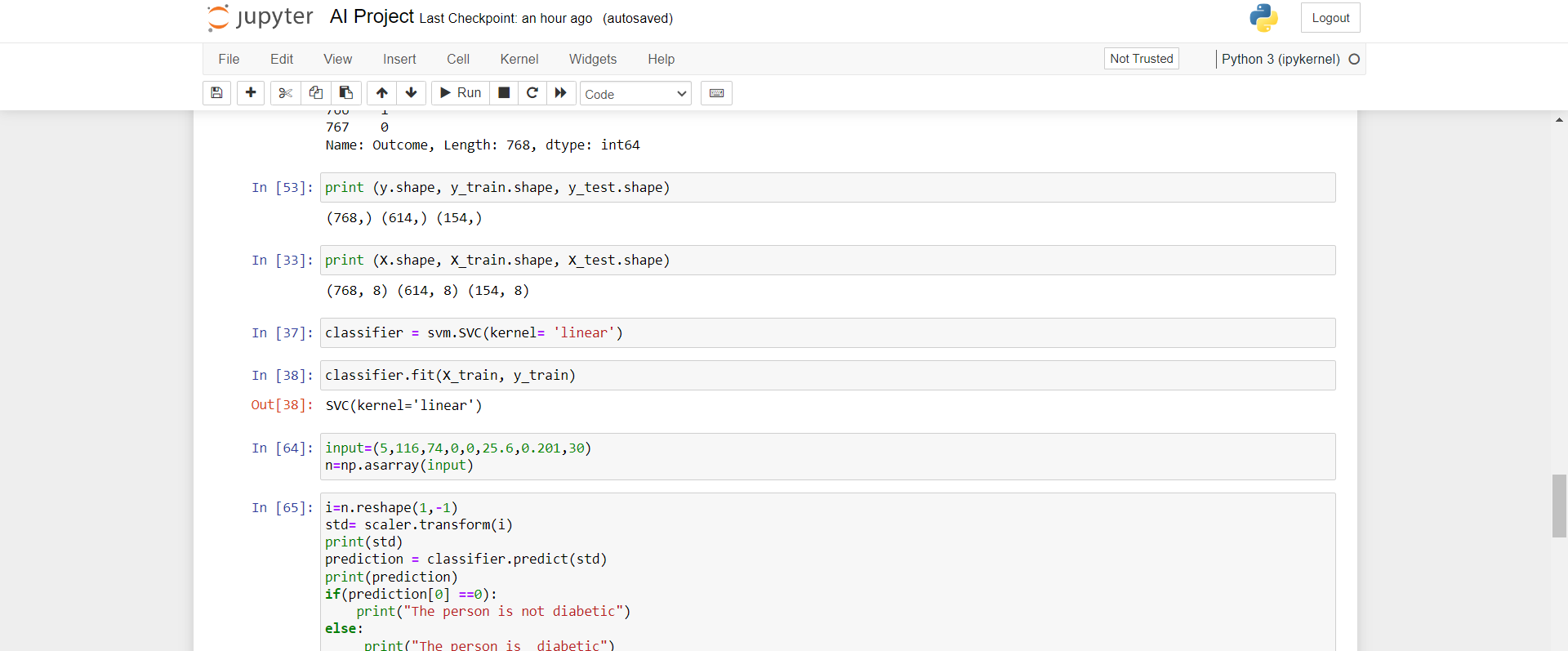
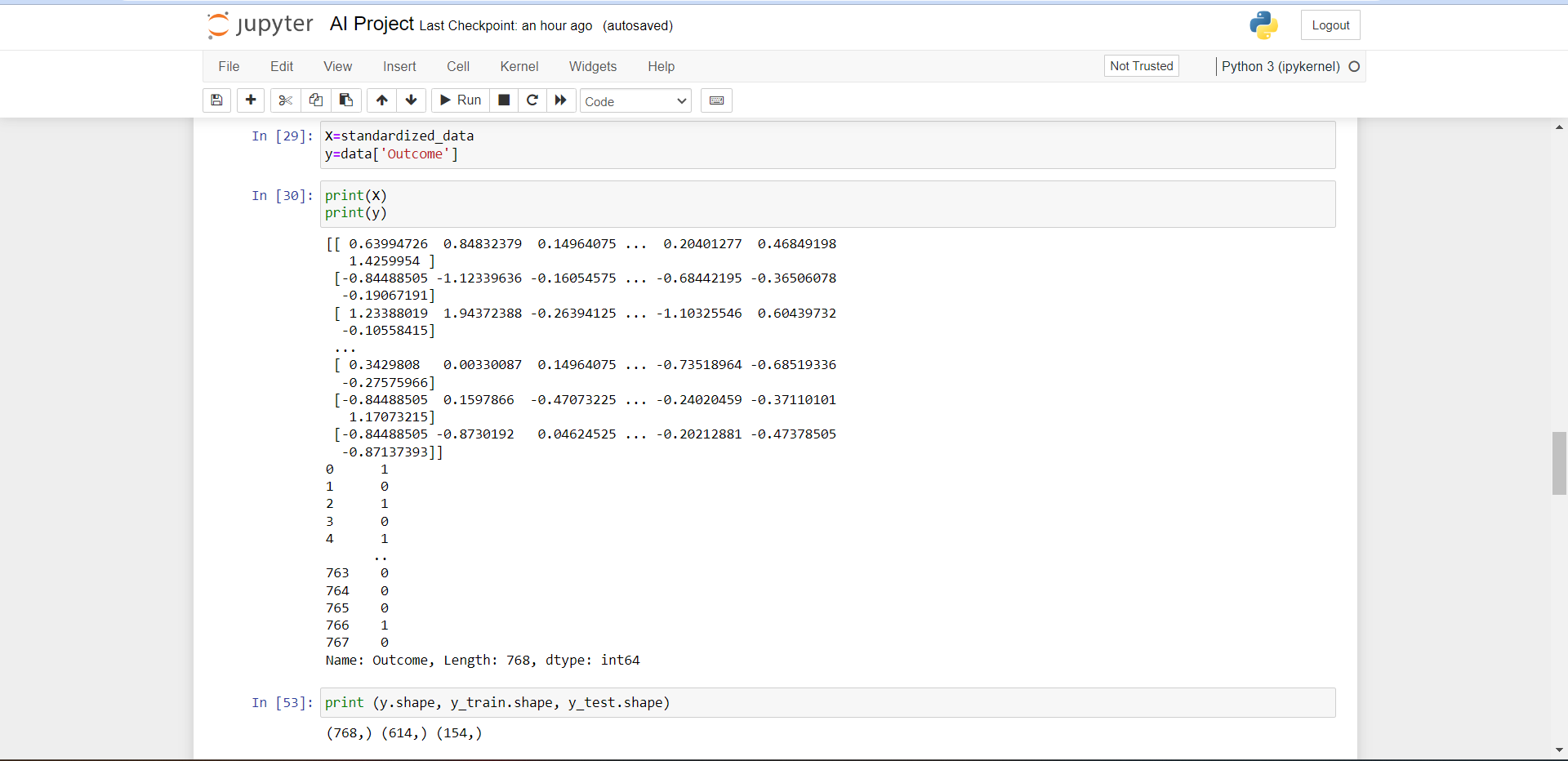
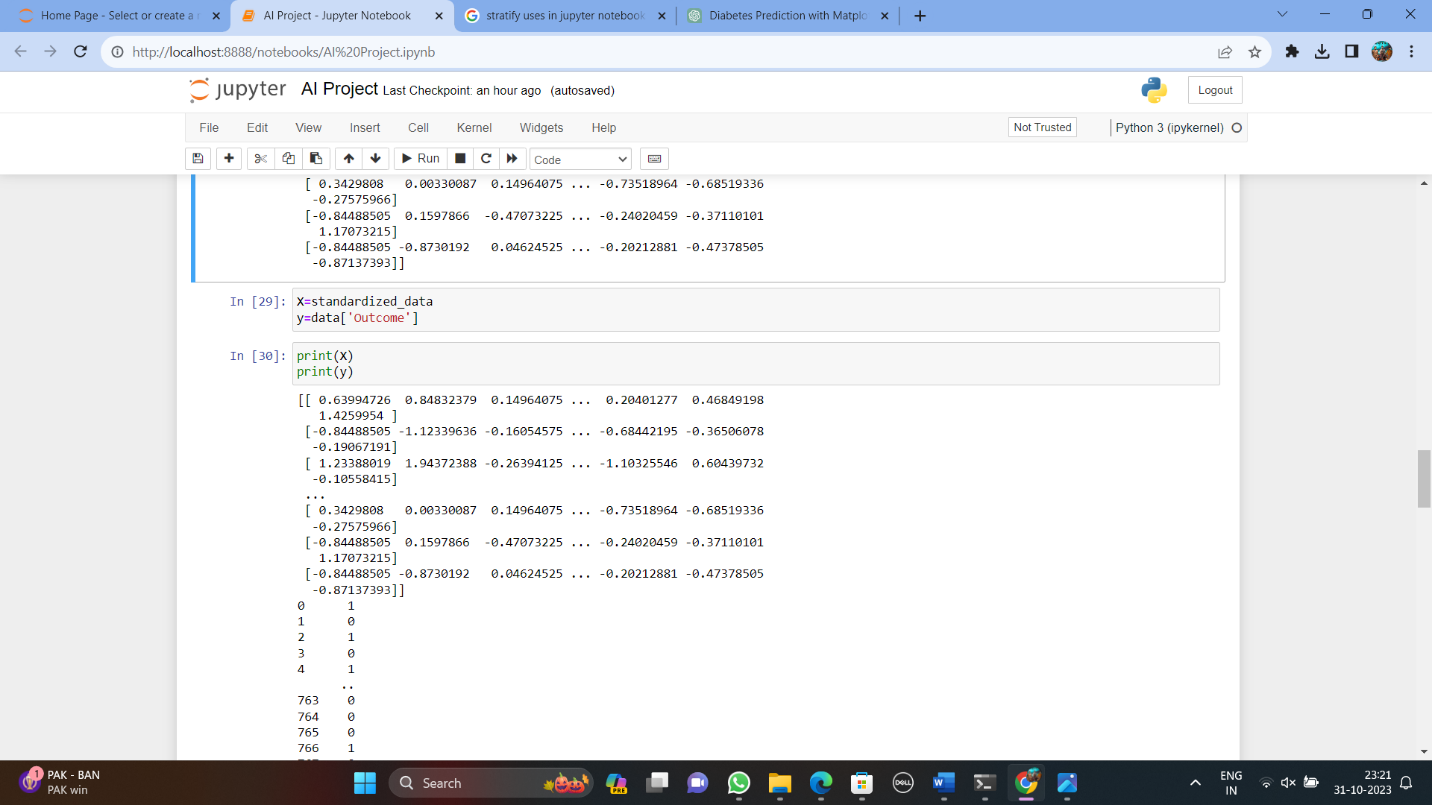
Provides a detailed breakdown of correct and incorrect predictions, including true positives, true negatives, false positives, and false negatives.

****

Model predict on X train and test set:



Other codes:



Output:

