**DIABETES**

**“Diabetes in female body”**

-Shruti Kharche

**Overview**

- The population for this study was the Pima Indian population near Phoenix, Arizona. The population has been under continuous study since 1965 by the [National Institute of Diabetes and Digestive and Kidney Diseases](http://www.niddk.nih.gov/Pages/default.aspx) because of its high incidence rate of diabetes.

-For the purposes of this dataset, diabetes was diagnosed according to [World Health Organization](http://www.who.int/diabetes/publications/en/) Criteria, which stated that if the 2 hour post-load glucose was at least 200 mg/dl at any survey exam or if the Indian Health Service Hospital serving the community found a glucose concentration of at least 200 mg/dl during the course of routine medical care.

-Given the medical data we can gather about people, we should be able to make better predictions on how likely a person is to suffer the onset of diabetes, and therefore act appropriately to help. We can start analyzing data and experimenting with algorithms that will help us study the onset of diabetes in Pima Indians.

Content of the Table

It contains data on female patients at least 21 years and above.

We have 768 instances and the following 8 attributes:

* Number of times pregnant (preg)
* Plasma glucose concentration a 2 hours in an oral glucose tolerance test (plas)
* Diastolic blood pressure in mm Hg (pres)
* Triceps skin fold thickness in mm (skin)
* 2-Hour serum insulin in mu U/ml (insu)
* Body mass index measured as weight in kg/(height in m)^2 (mass)
* Diabetes pedigree function (pedi)
* Age in years (age)
* Outcome(1 or 0)

A particularly interesting attribute used in the study was the Diabetes Pedigree Function, pedi. It provided some data on diabetes mellitus history in relatives and the genetic relationship of those relatives to the patient. This measure of genetic influence gave us an idea of the hereditary risk one might have with the onset of diabetes mellitus. Based on observations in the proceeding section, it is unclear how well this function predicts the onset of diabetes.

DATA MODELLING

 We’ll be using Machine Learning to predict whether a person has diabetes or not, based on information about the patient such as blood pressure, body mass index (BMI), age, etc. The tutorial walks through the various stages of the data science workflow. In particular, the tutorial has the following sections

* Overview
* Data Description
* Data Exploration
* Data Preparation
* Training and Evaluating the Machine Learning Model
* Interpreting the ML Model
* Making Predictions with the Model

**Overview**

- The population for this study was the Pima Indian population near Phoenix, Arizona. The population has been under continuous study since 1965 by the [National Institute of Diabetes and Digestive and Kidney Diseases](http://www.niddk.nih.gov/Pages/default.aspx) because of its high incidence rate of diabetes.

-For the purposes of this dataset, diabetes was diagnosed according to [World Health Organization](http://www.who.int/diabetes/publications/en/) Criteria, which stated that if the 2 hour post-load glucose was at least 200 mg/dl at any survey exam or if the Indian Health Service Hospital serving the community found a glucose concentration of at least 200 mg/dl during the course of routine medical care.

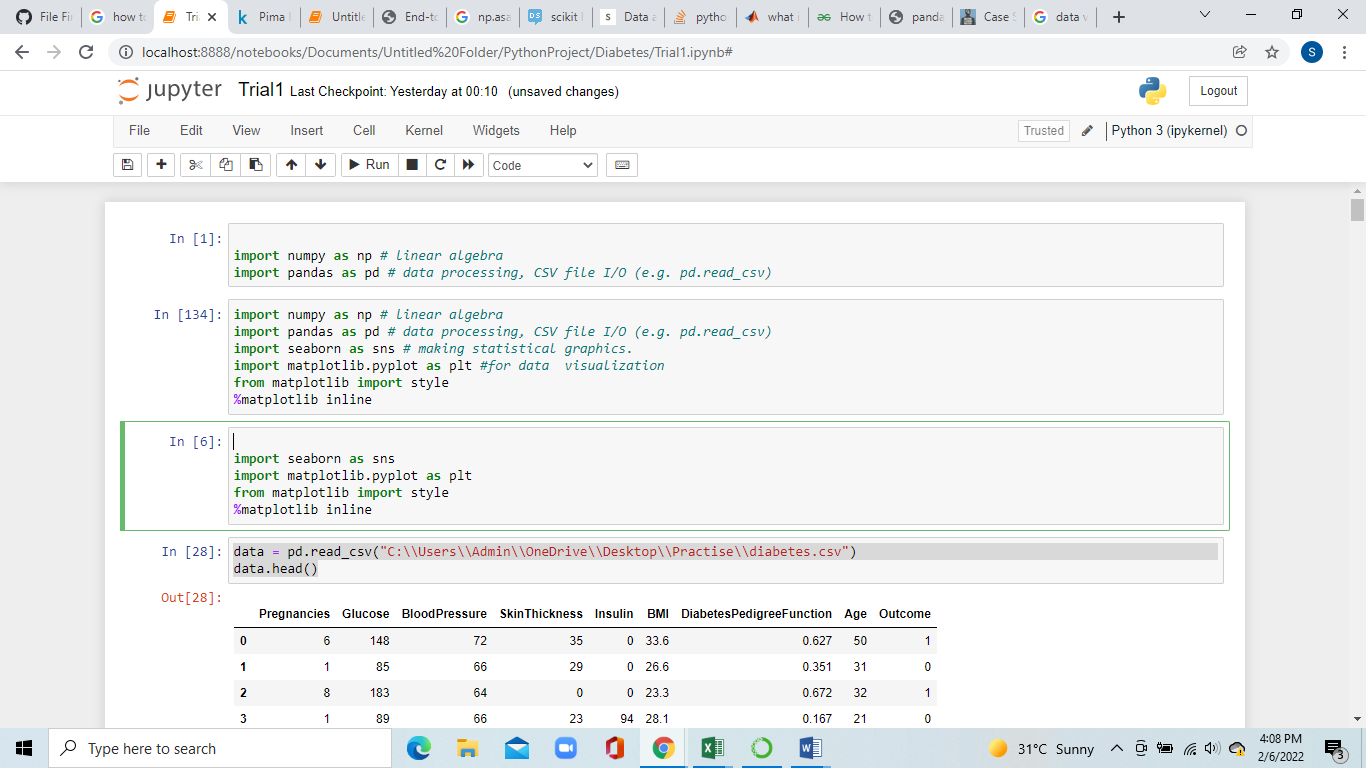
-Given the medical data we can gather about people, we should be able to make better predictions on how likely a person is to suffer the onset of diabetes, and therefore act appropriately to help. We can start analyzing data and experimenting with algorithms that will help us study the onset of diabetes in Pima Indians.

# Data set used was from Kaggel – Pima Indian Diabetes Database .

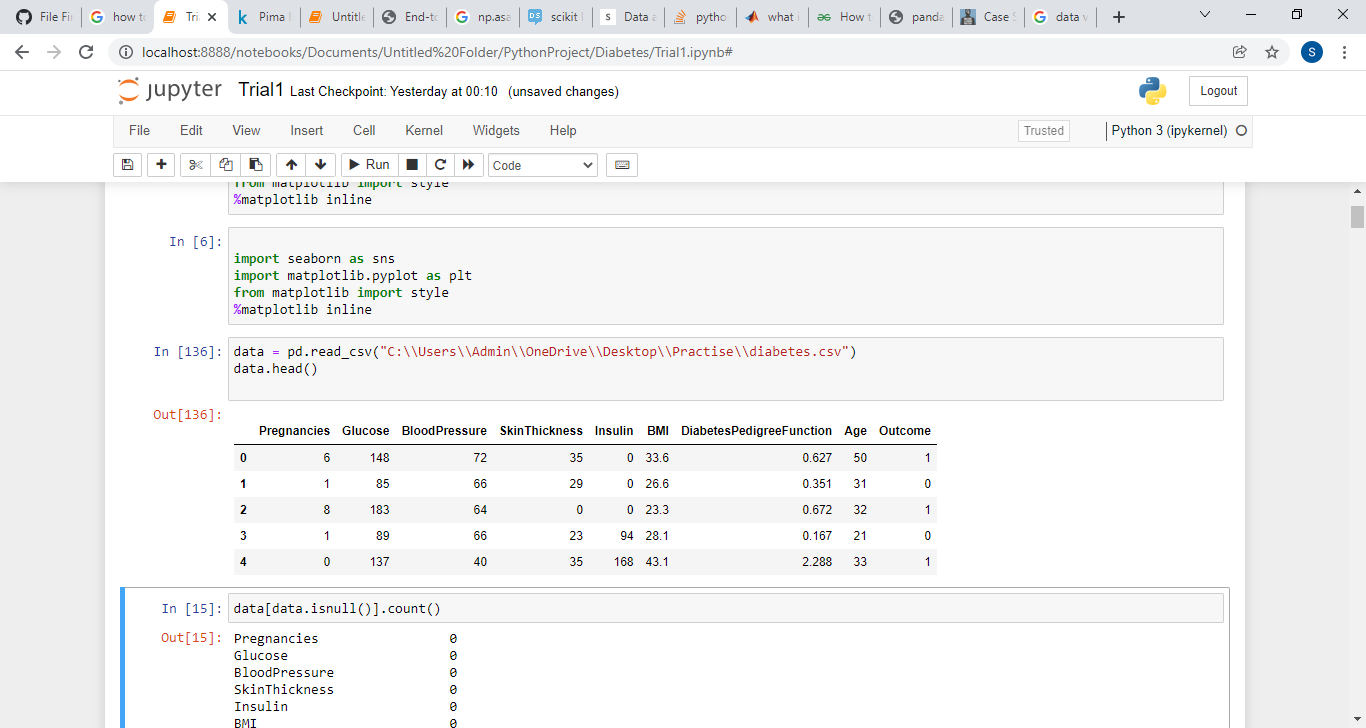
We’ll be using Python and some of its popular data science related packages. First of all, we will import pandas to read our data from a CSV file and manipulate it for further use. We will also use numpy to convert out data into a format suitable to feed our classification model. We’ll use seaborn and matplotlib for visualizations. We will then import Logistic Regression algorithm from sklearn. This algorithm will help us build our classification model. Lastly, we will use joblib available in sklearn to save our model for future use.

Data Description

* Liabraries used

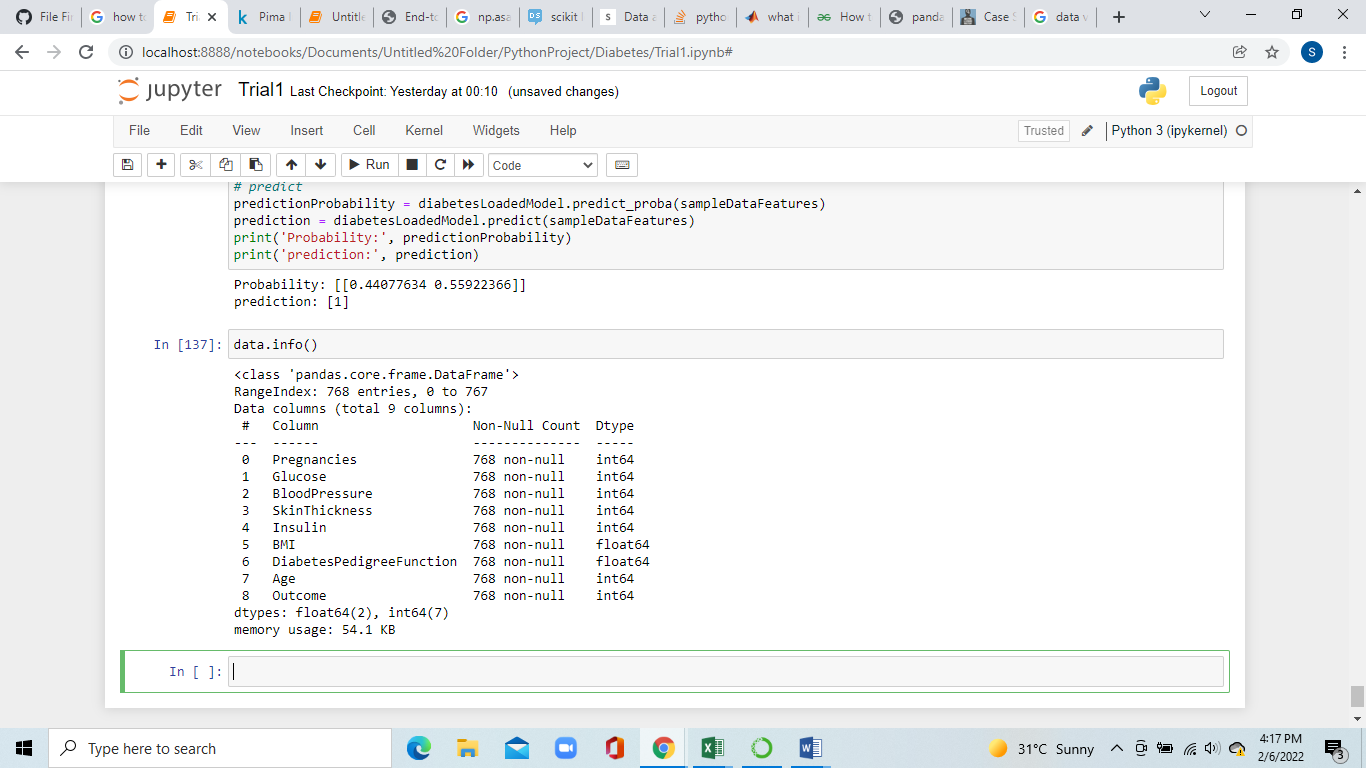


* We have our data saved in a CSV file called diabetes.csv . We first read our dataset into a pandas dataframe called Data and then use the head() function to show the first five records from our dataset.



The following features have been provided to help us predict whether a person is diabetic or not:

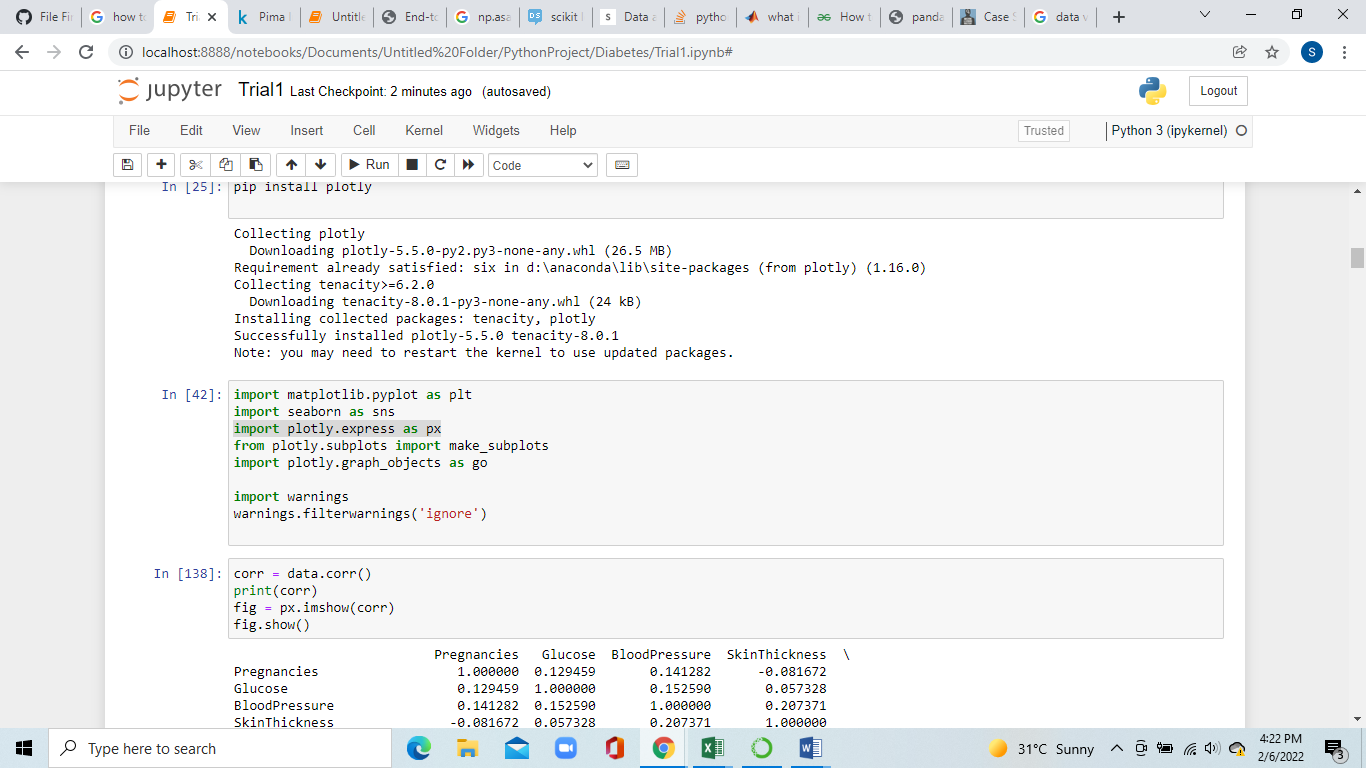
* **Pregnancies:**Number of times pregnant
* **Glucose:** Plasma glucose concentration over 2 hours in an oral glucose tolerance test
* **BloodPressure:**Diastolic blood pressure (mm Hg)
* **SkinThickness:** Triceps skin fold thickness (mm)
* **Insulin:** 2-Hour serum insulin (mu U/ml)
* **BMI:** Body mass index (weight in kg/(height in m)2)
* **DiabetesPedigreeFunction:** Diabetes pedigree function (a function which scores likelihood of diabetes based on family history)
* **Age:** Age (years)
* **Outcome:** Class variable (0 if non-diabetic, 1 if diabetic)
* Cheking if the data is clean (has no null values)

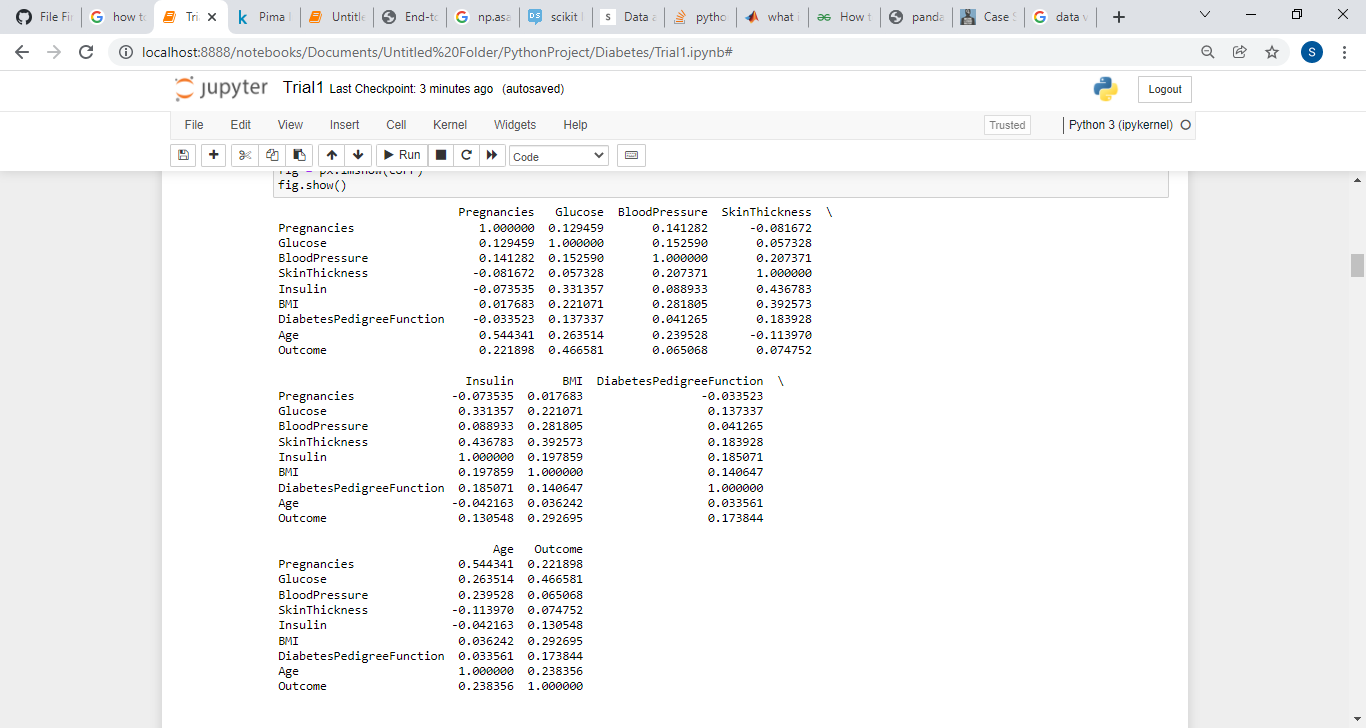


Conclusion- Data is clean and has no null values.

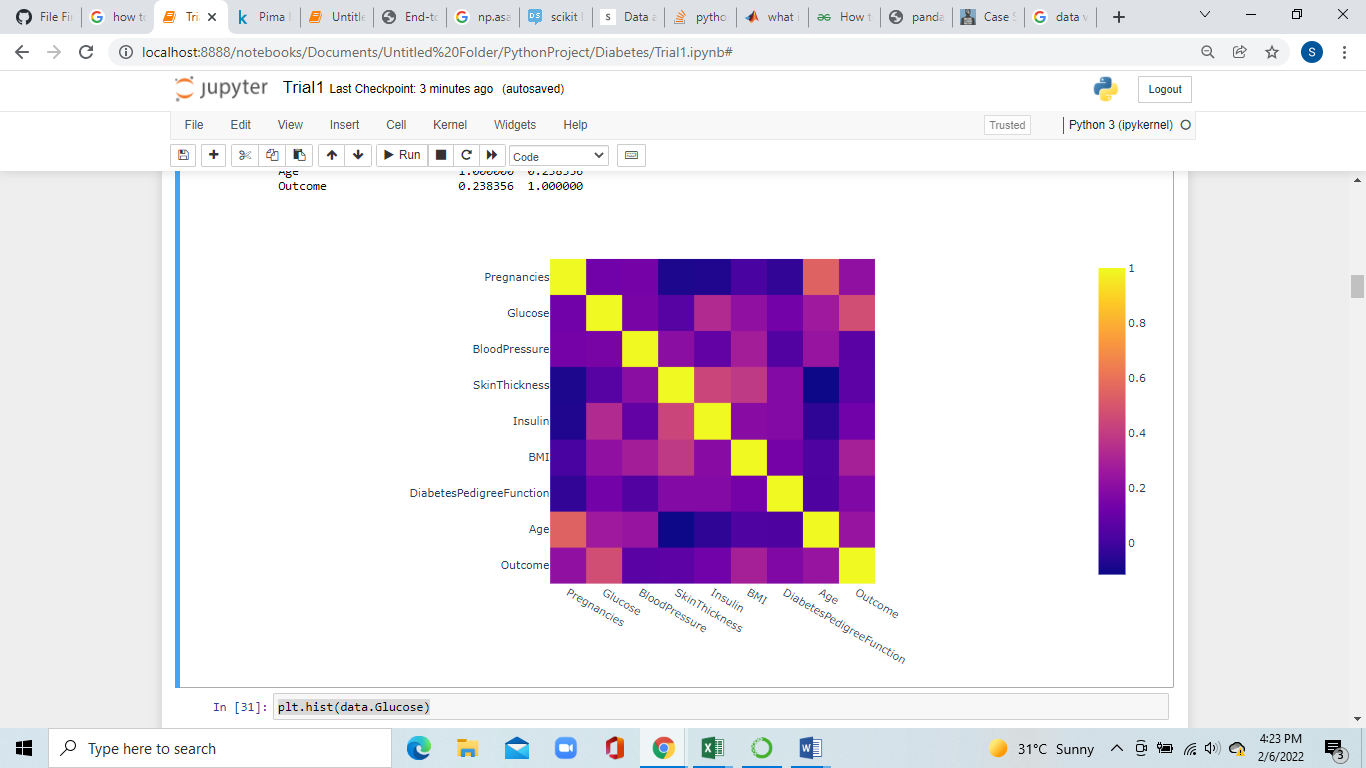
Data Exploration

* Data Correlation was made





Heat Map

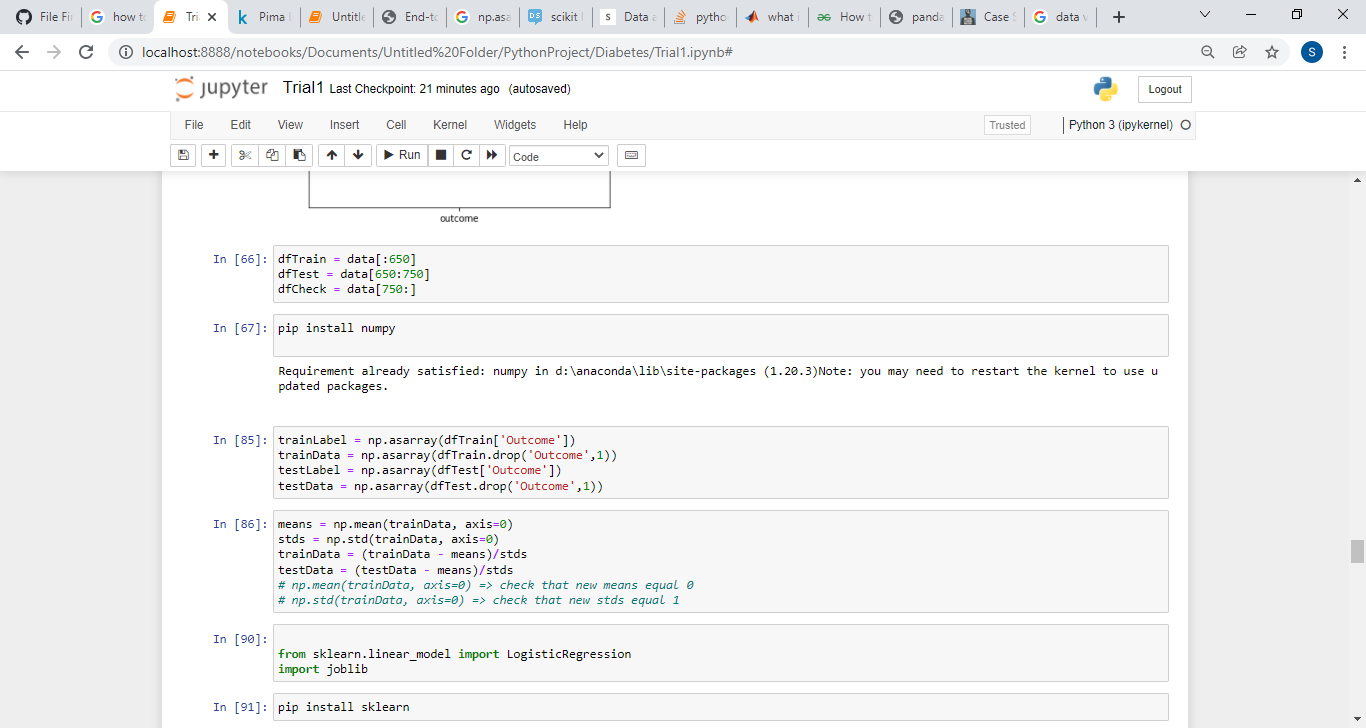


Conclusion

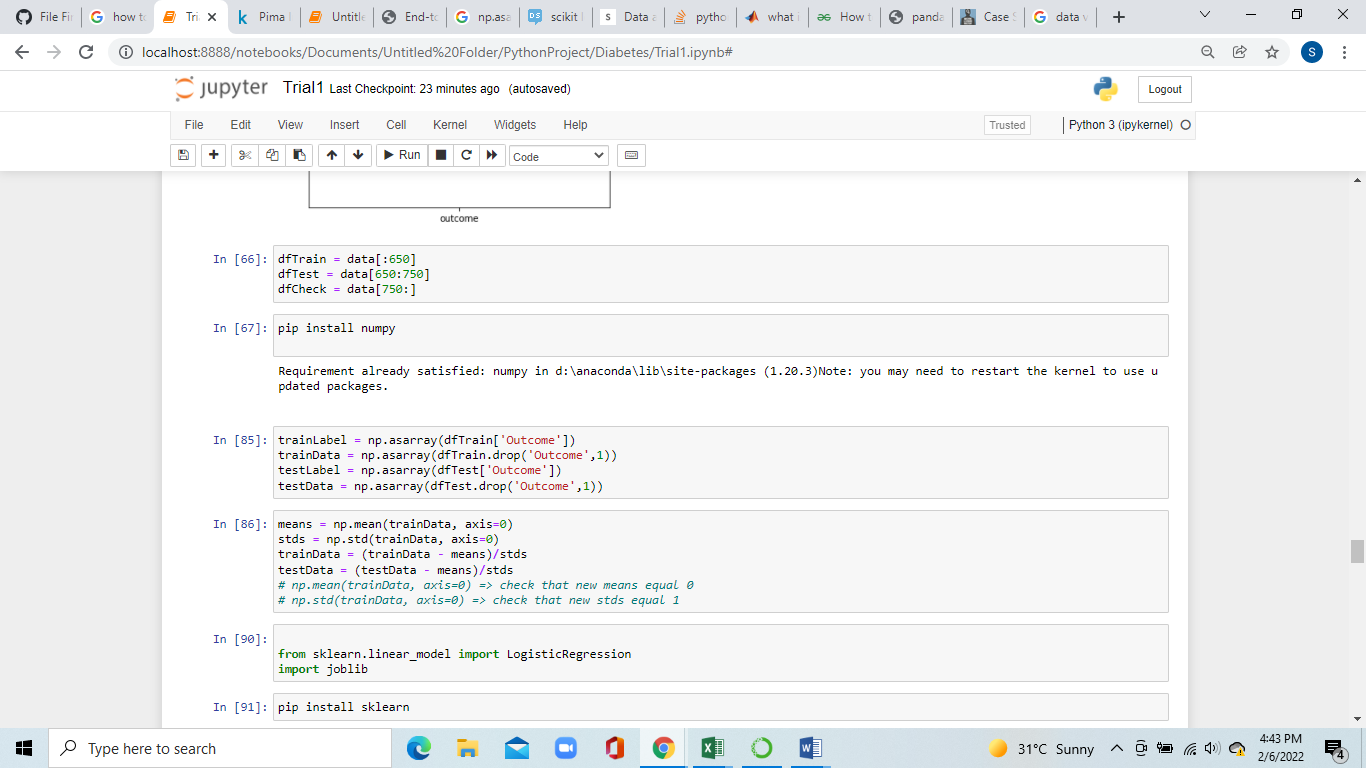
* Values closer to 0 have a non linear relationship , eg: insulin-pregnancies
* Values closer to 1 are directly correlated (directly propoertional) , eg: Pregnanicies-Age
* Values equal to 1 are perfectly correlating , eg: Age-Age

Data Preparation

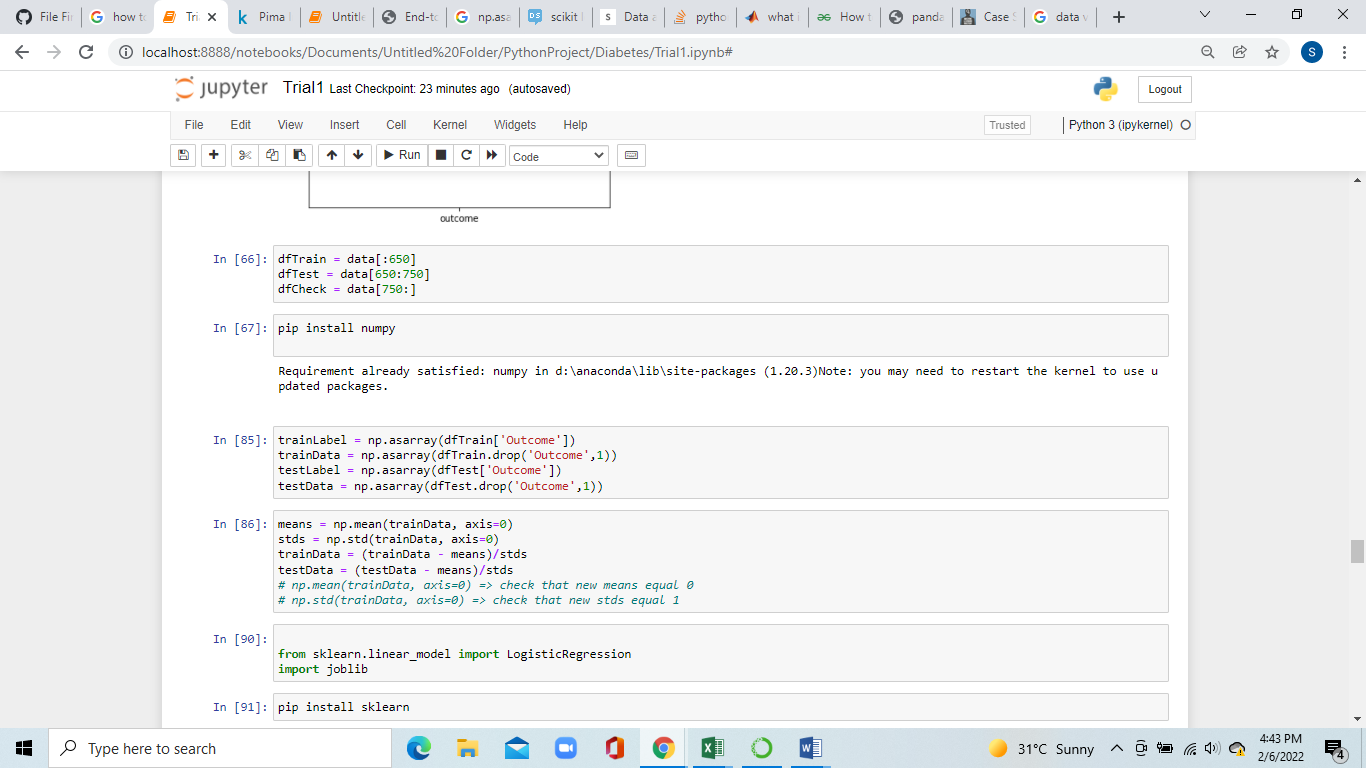
Splitting the dataset into Train and Test .The data set consists of record of 767 patients in total. To train our model we will be using 650 records. We will be using 100 records for testing, and the last 17 records to cross check our model.



Next, we separate the label and features (for both training and test dataset). In addition to that, we will also convert them into NumPy arrays as our machine learning algorithm process data in NumPy array format.



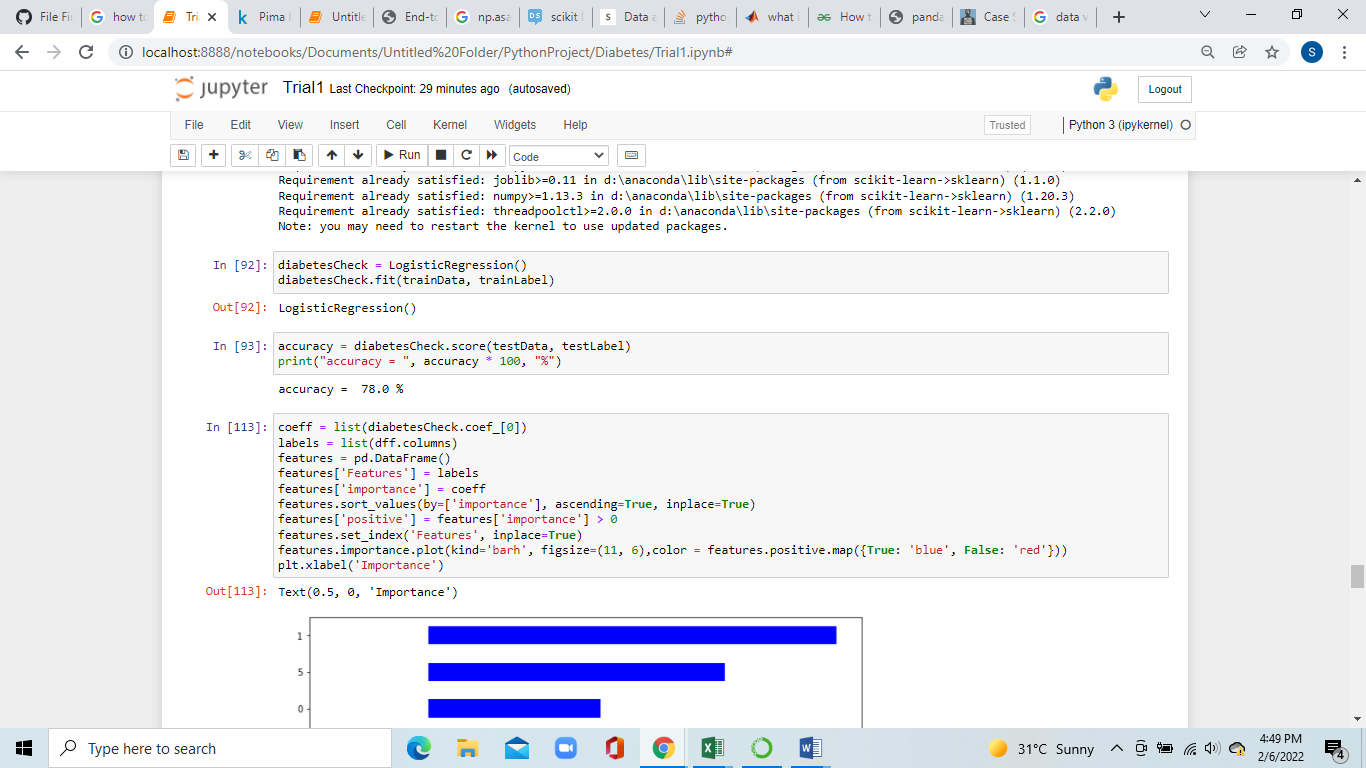
As the final step before using machine learning, we will normalize our inputs. Machine Learning models often benefit substantially from input normalization. It also makes it easier for us to understand the importance of each feature later, when we’ll be looking at the model weights. We’ll normalize the data such that each variable has 0 mean and standard deviation of 1.



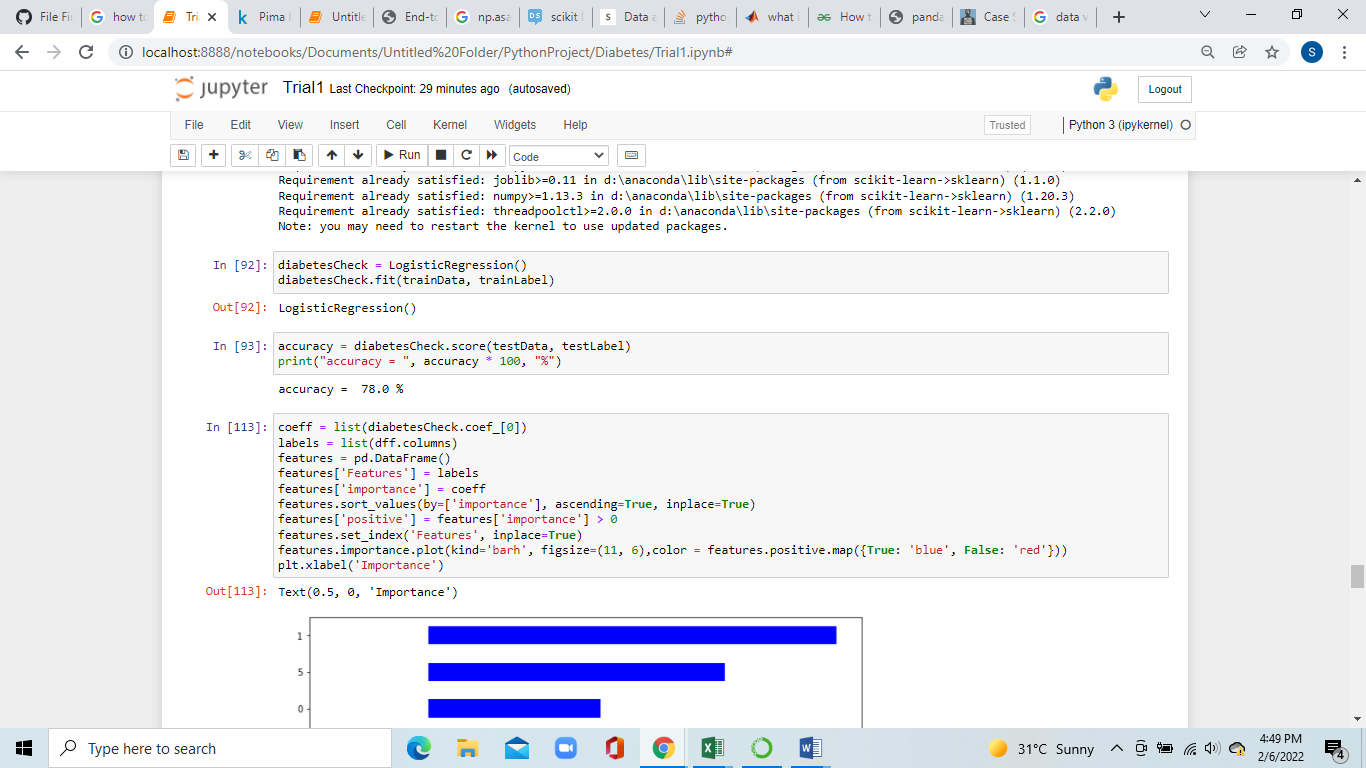
Conclusion –Data has been successfully divided.

# Training and Evaluating Machine Learning Model

We can now train our classification model. We’ll be using a machine simple learning model called logistic regression. Since the model is readily available in sklearn, the training process is quite easy and we can do it in few lines of code. First, we create an instance called diabetesCheck and then use the fit function to train the model.



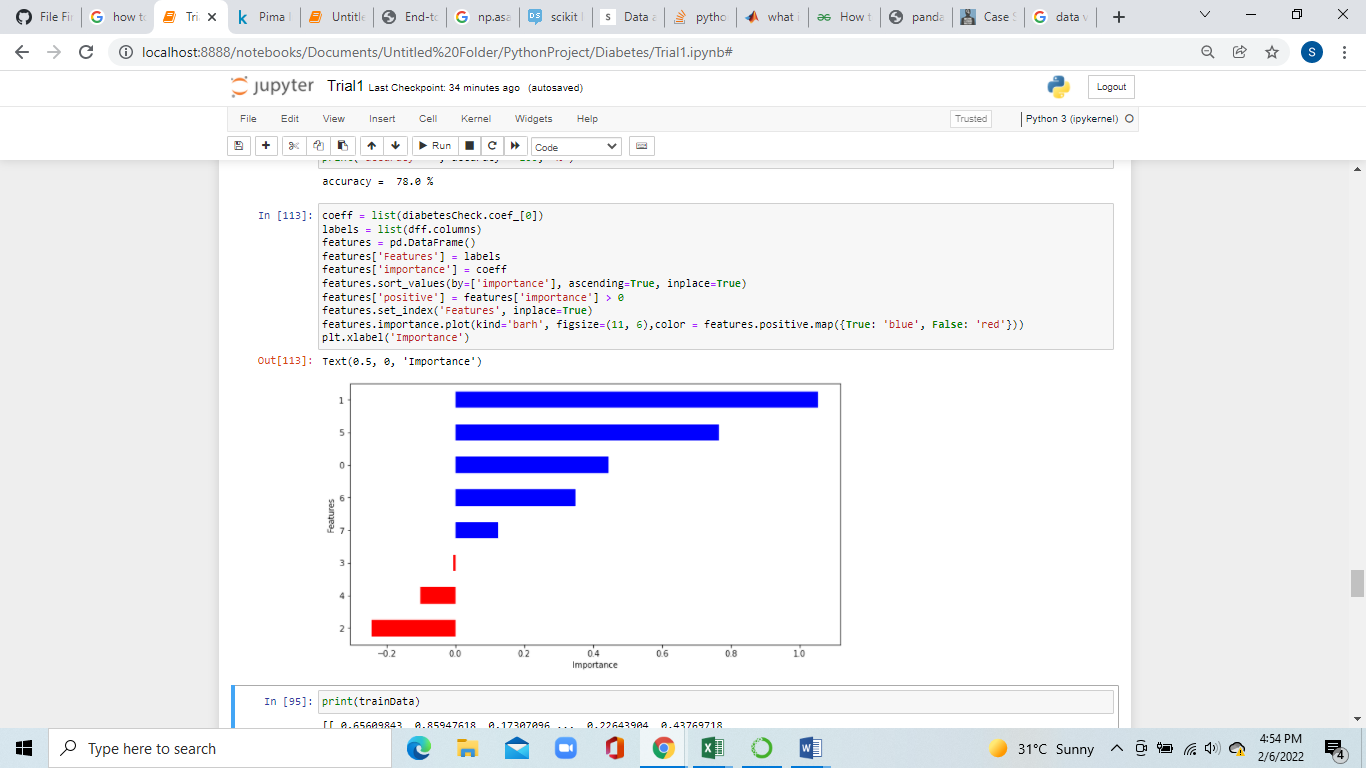
Next, we will use our test data to find out accuracy of the model.

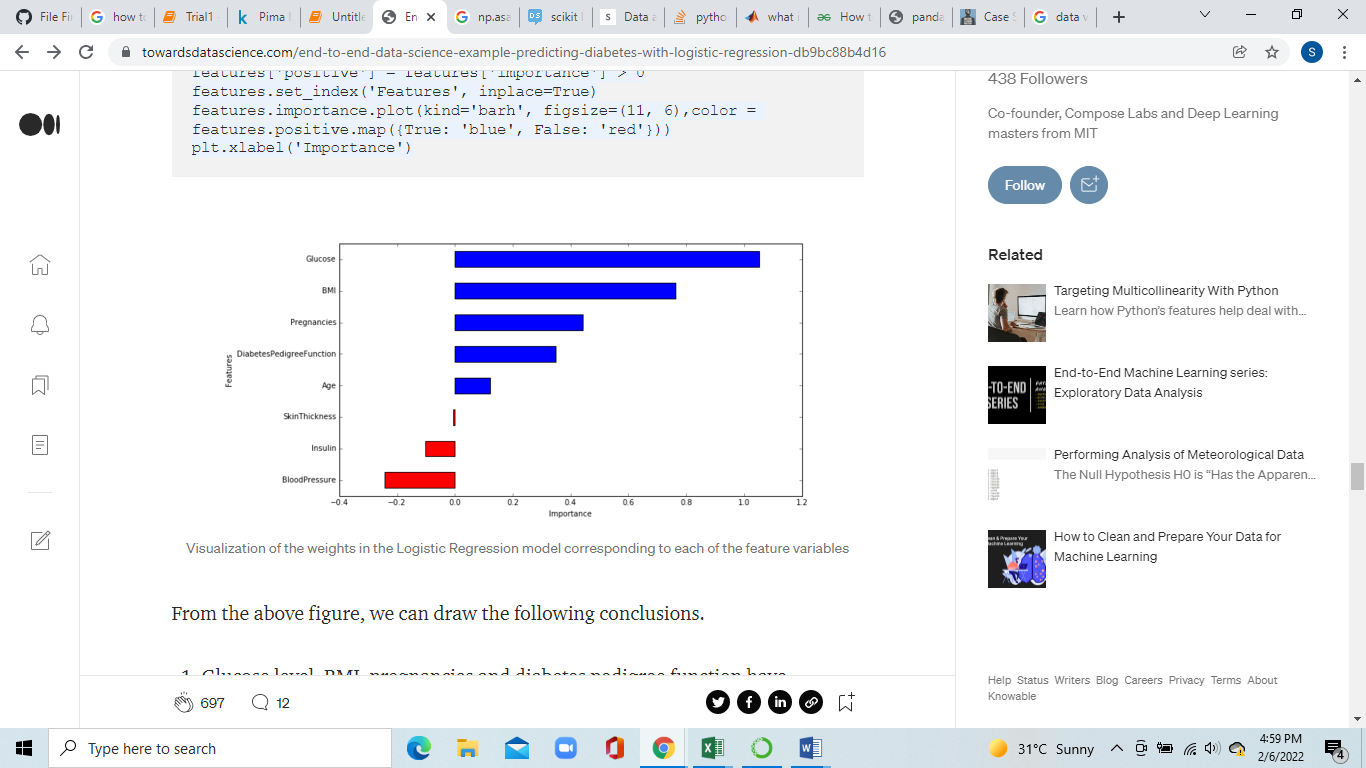


Conclusion – The accuracy of our functions performed on the data is about 78%

# Interpreting the ML Model

To get a better sense of what is going on inside the logistic regression model, we can visualize how our model uses the different features and which features have greater effect.



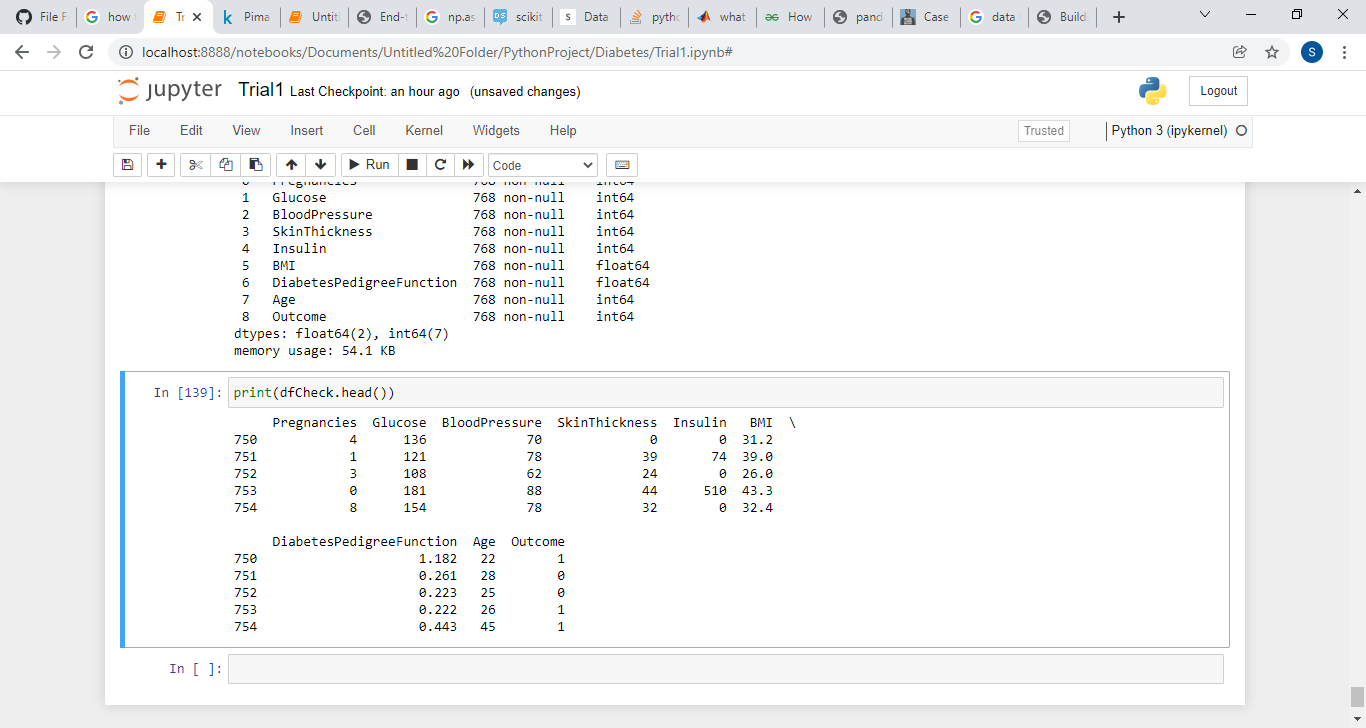


From the above figure, we can draw the following conclusions.

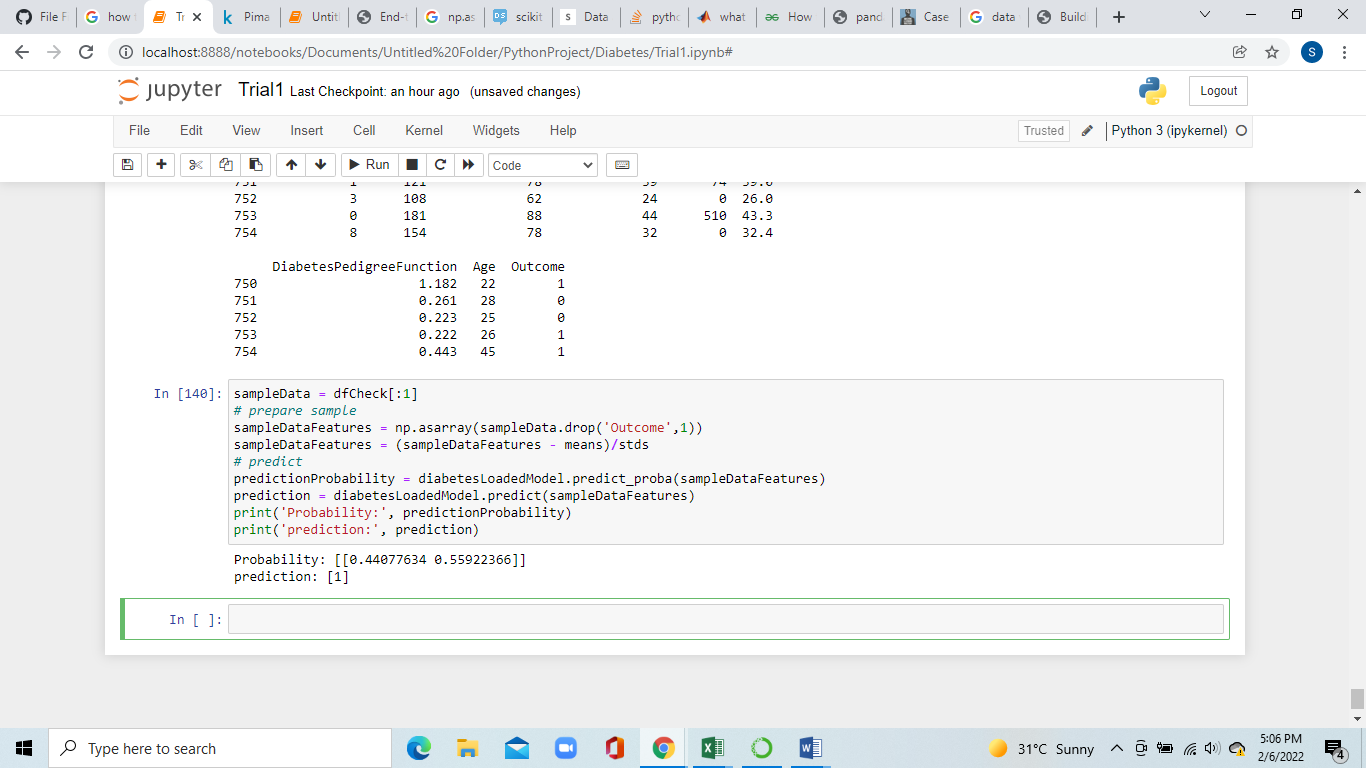
1. Glucose level, BMI, pregnancies and diabetes pedigree function have significant influence on the model, specially glucose level and BMI. It is good to see our machine learning model match what we have been hearing from doctors our entire lives!
2. Blood pressure has a negative influence on the prediction, i.e. higher blood pressure is correlated with a person not being diabetic. (also, note that blood pressure is more important as a feature than age, because the *magnitude* is higher for blood pressure).
3. Although age was more correlated than BMI to the output variables (as we saw during data exploration), the model relies more on BMI. This can happen for several reasons, including the fact that the correlation captured by age is also captured by some other variable, whereas the information captured by BMI is not captured by other variables.

# Making Predictions with the model

We will now use our unused data to see how predictions can be made. We have our unused data in dfCheck.



We will now use the first record to make our prediction.



The first element of array **predictionProbability** 0.438 is the probability of the class being **0** and second element 0.561 is the probability of the class being **1**. The probabilities sum to 1. As we can see that the**1** is more probable class, we get **[1]** as our prediction, which means that the model predicts that the person has diabetes.

Precautions

1. **Exercise regularly -** Exercise increases the insulin sensitivity of your cells, meaning that you need less insulin to manage your blood sugar levels
2. **Reduce your total carb intake -** Eating foods high in refined carbs and sugar increase blood sugar and insulin levels, which may eventually lead to diabetes. Limiting total carbohydrate intake and choosing options that don’t cause blood sugar spikes may help reduce your risk.
3. **Drink water as your primary beverage -** Drinking water instead of sugary beverages may help manage blood sugar and insulin levels, thereby reducing your risk of diabetes.
4. **Reduce your portion sizes -** Avoiding large portion sizes may help reduce insulin and blood sugar levels, promote weight loss, and decrease your risk of diabetes.