**CURNEU ASSESSMENT**

**PROJECT DOCUMENTATION**

**SD03Q017**

**Name: Priyadharshini R**

**Roll no: 1833039**

**Course Name: M.Sc. Decision and Computing Sciences**

**Email id:** [**priyadharshinir.my@gmail.com**](mailto:priyadharshinir.my@gmail.com)

**Problem Statement 2: Diabetes prediction**

Predict which patient has diabetes from Diabetes Database.csv and try to understand the dataset attributes and try to figure out type ML model suits and build from scratch.

**SUMMARY:**

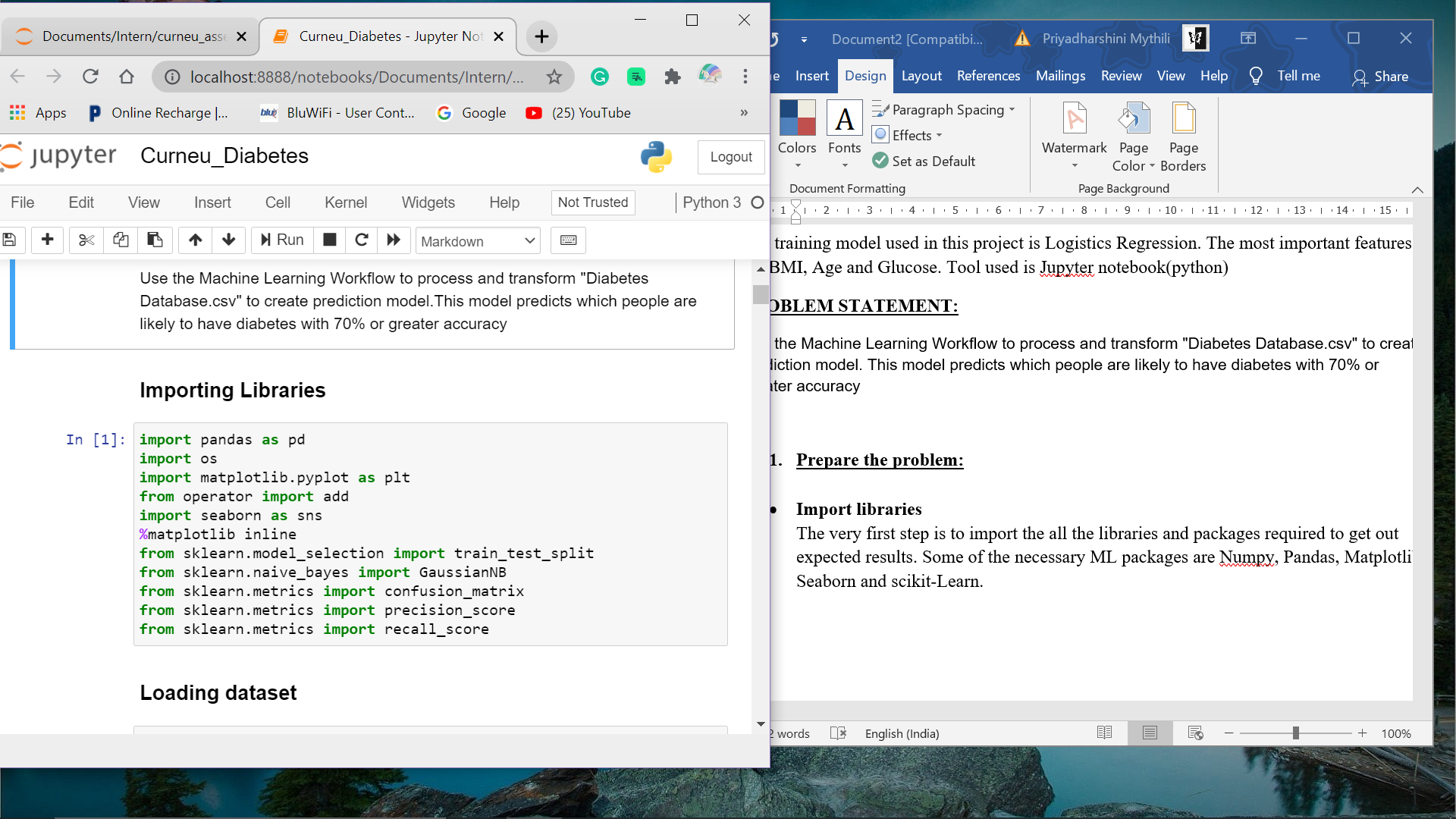
The training model used in this project is Logistics Regression. The most important features are BMI, Age and Glucose. Tool used is Jupyter notebook(python)

Use the Machine Learning Workflow to process and transform "Diabetes Database.csv" to create prediction model. This model predicts which people are likely to have diabetes with 70% or greater accuracy

1. **Prepare the problem:**

* **Import libraries**

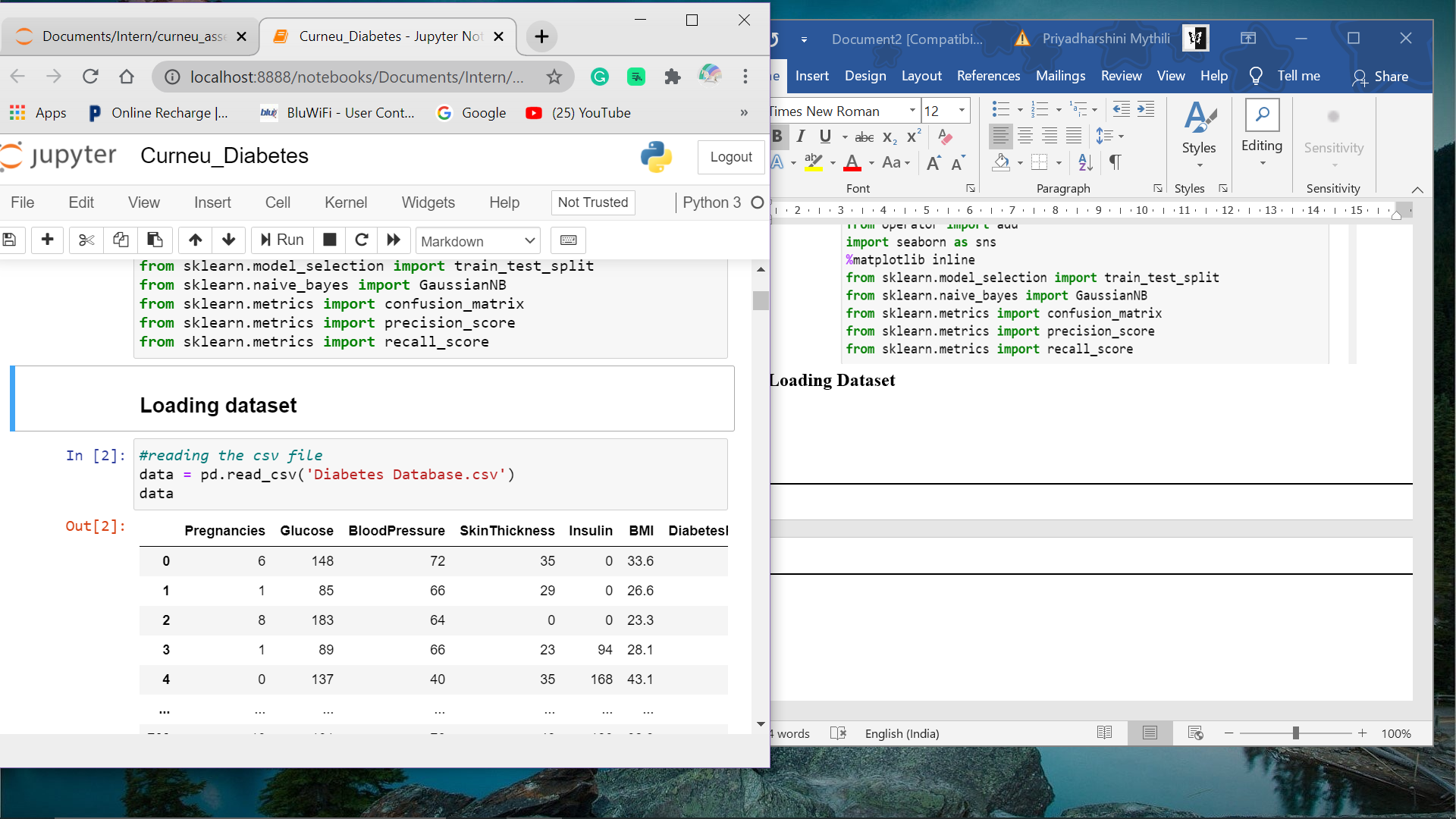
The very first step is to import the all the libraries and packages required to get out expected results. Some of the necessary ML packages are Numpy, Pandas, Matplotlib, Seaborn and scikit-Learn.

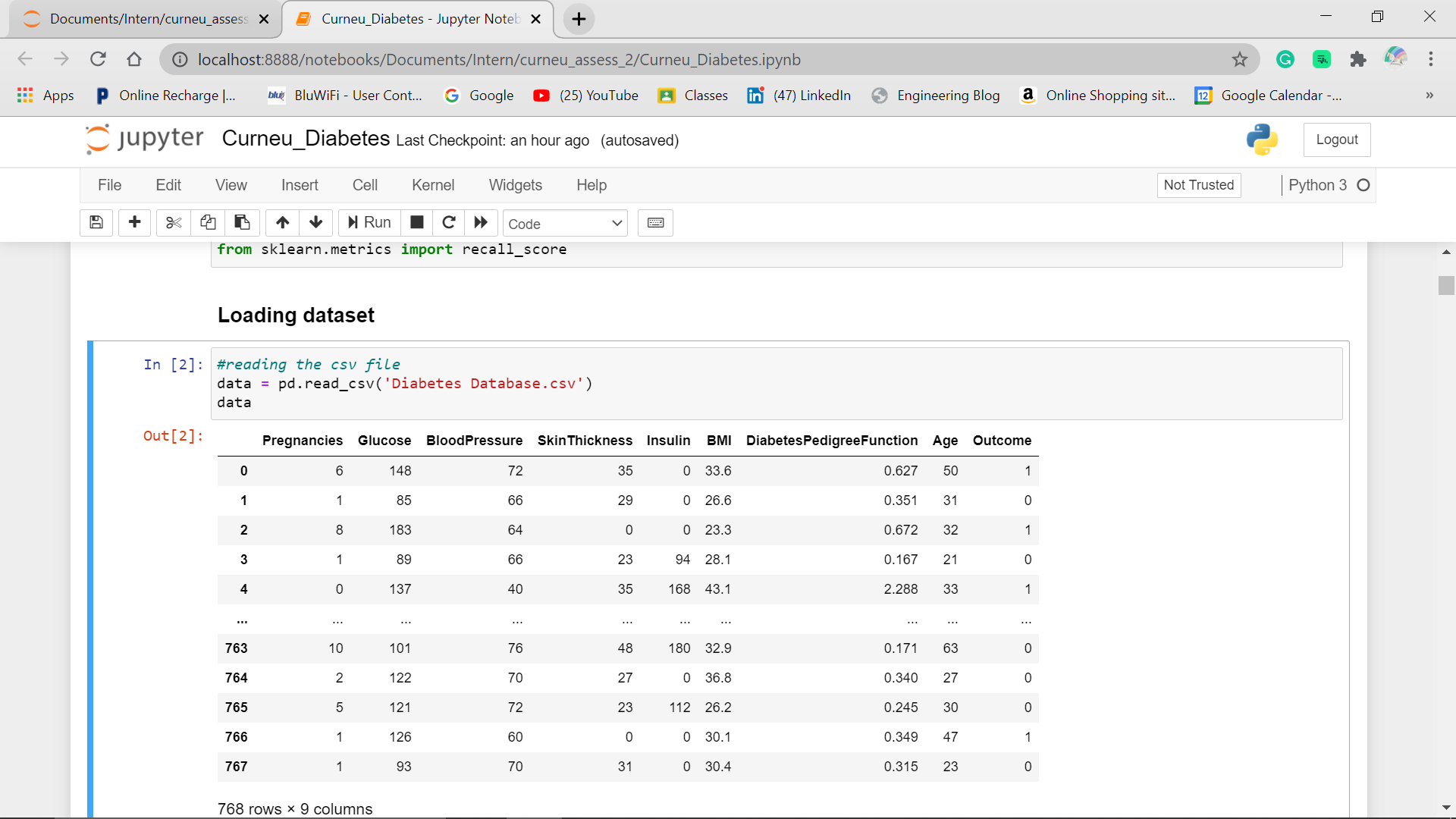


* **Loading Dataset**

Loading the dataset into Jupyter Notebook using Pandas Data Frame

* + - The dataset consists of 768 rows and 9 columns (Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, DiabetesPF, Age, Outcome)





1. **Summarize the given dataset:**

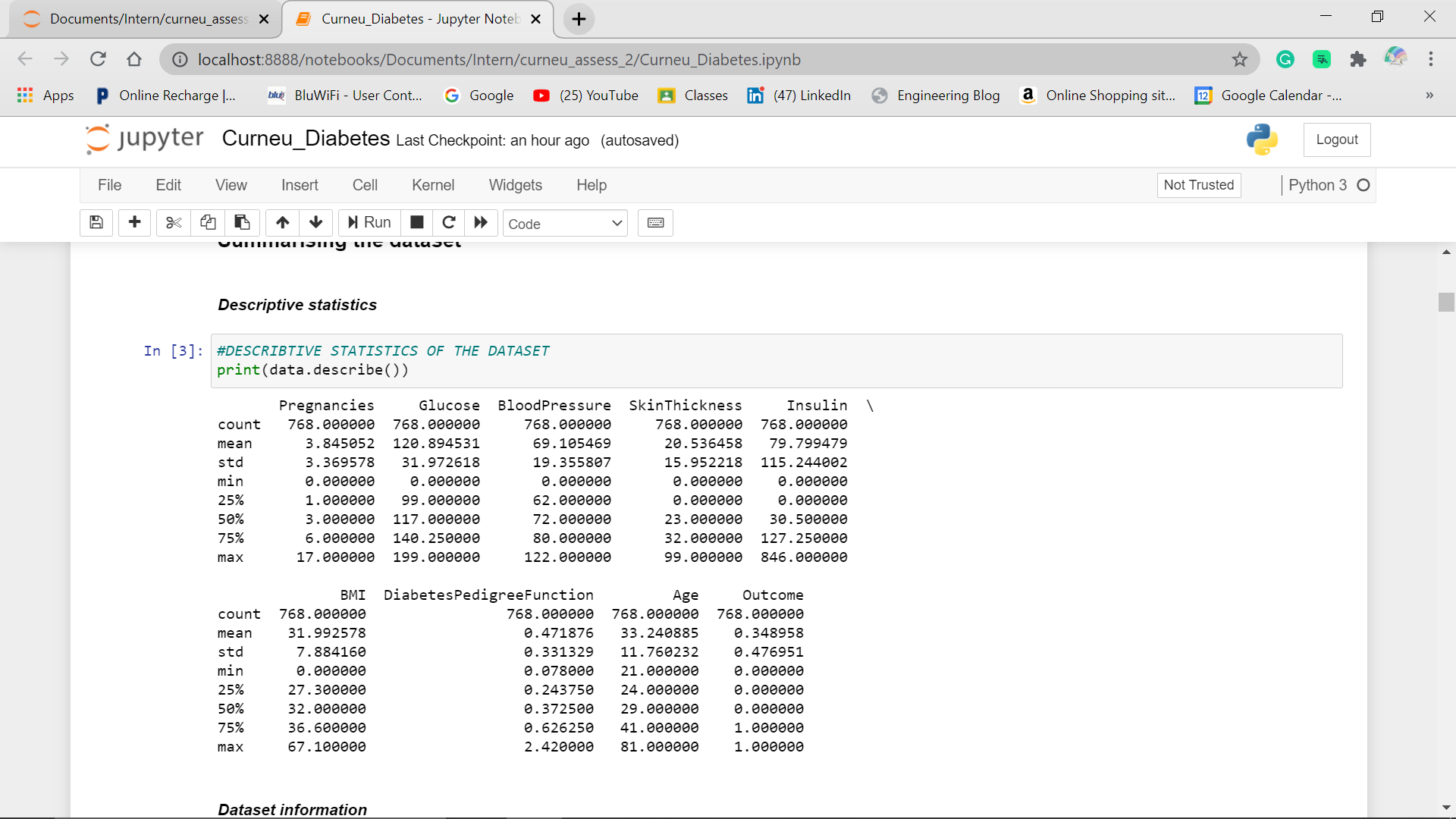
Once the dataset is loaded, we have to see the how many rows and columns are present, what are the data types of each column. To do this Pandas data frame can be used.

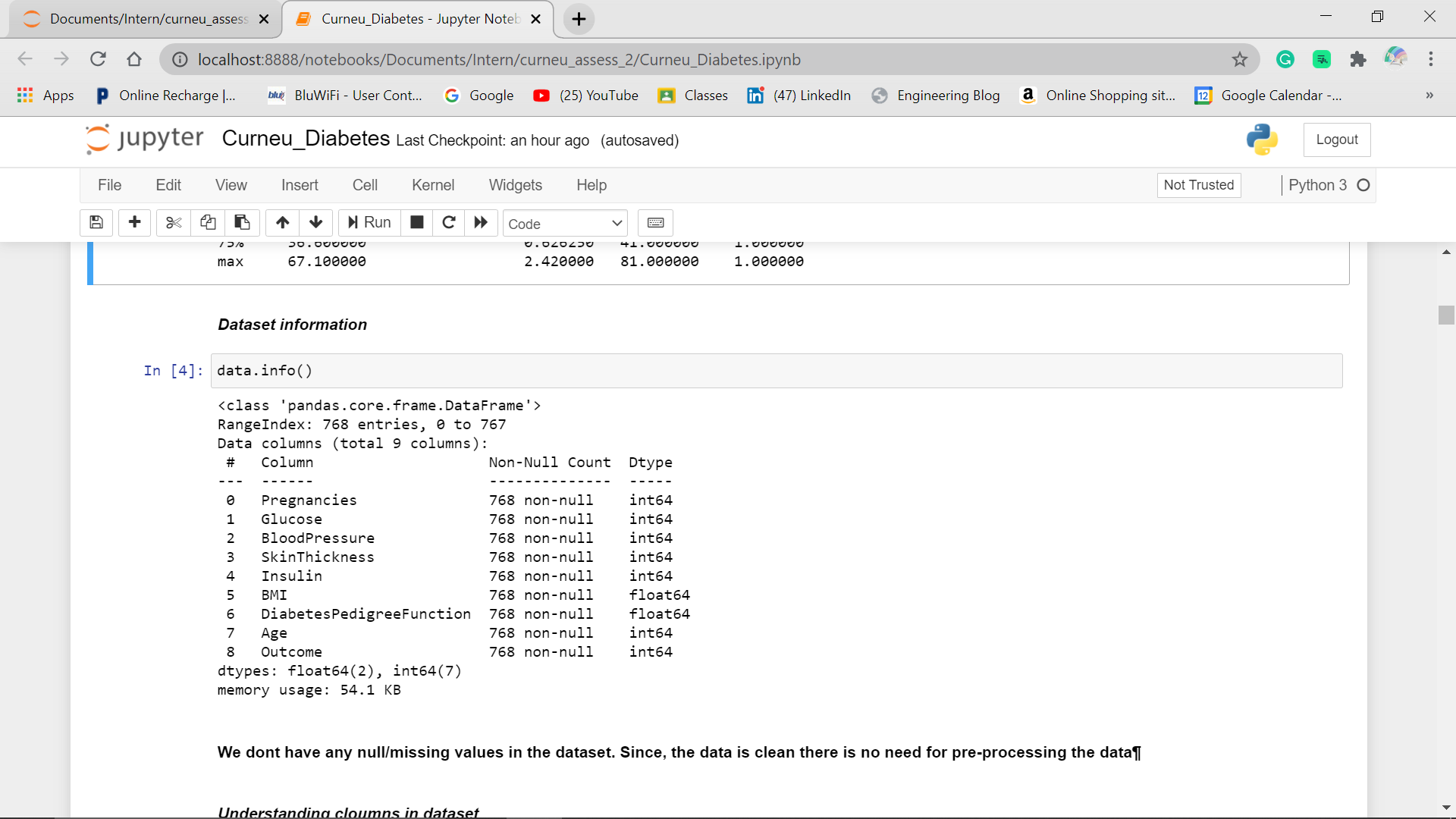
* **Descriptive statistics:**

This describes the statistics like mean, median, percentiles etc, for each column in the dataset.

And the info() function gives us the values, their types in the dataset and presence of missing values can also be identified correctly.

* + - From the dataset information, it is evident that dataset does not have any missing value.



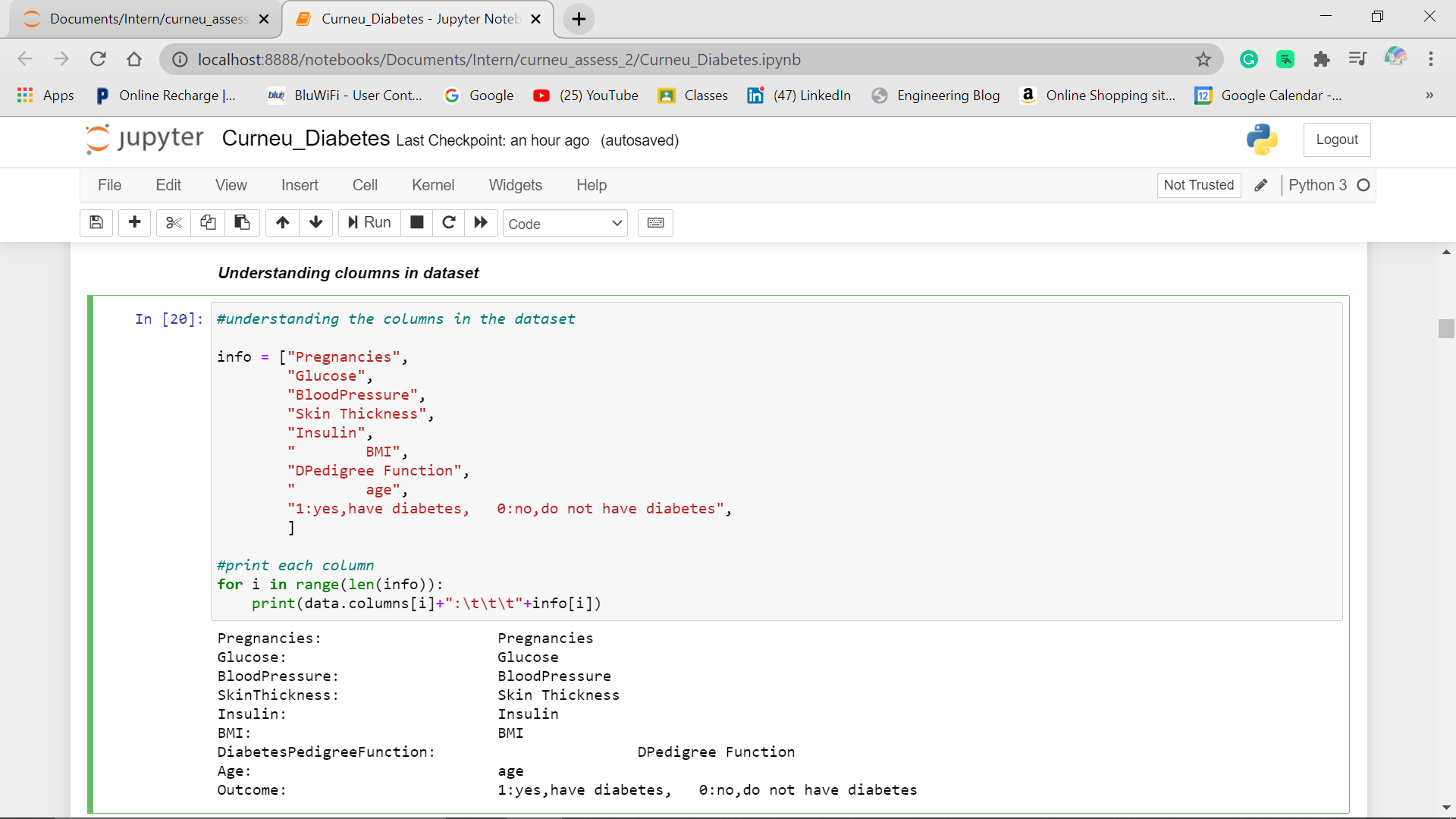


1. **Explanatory DataAnalysis(EDA)**

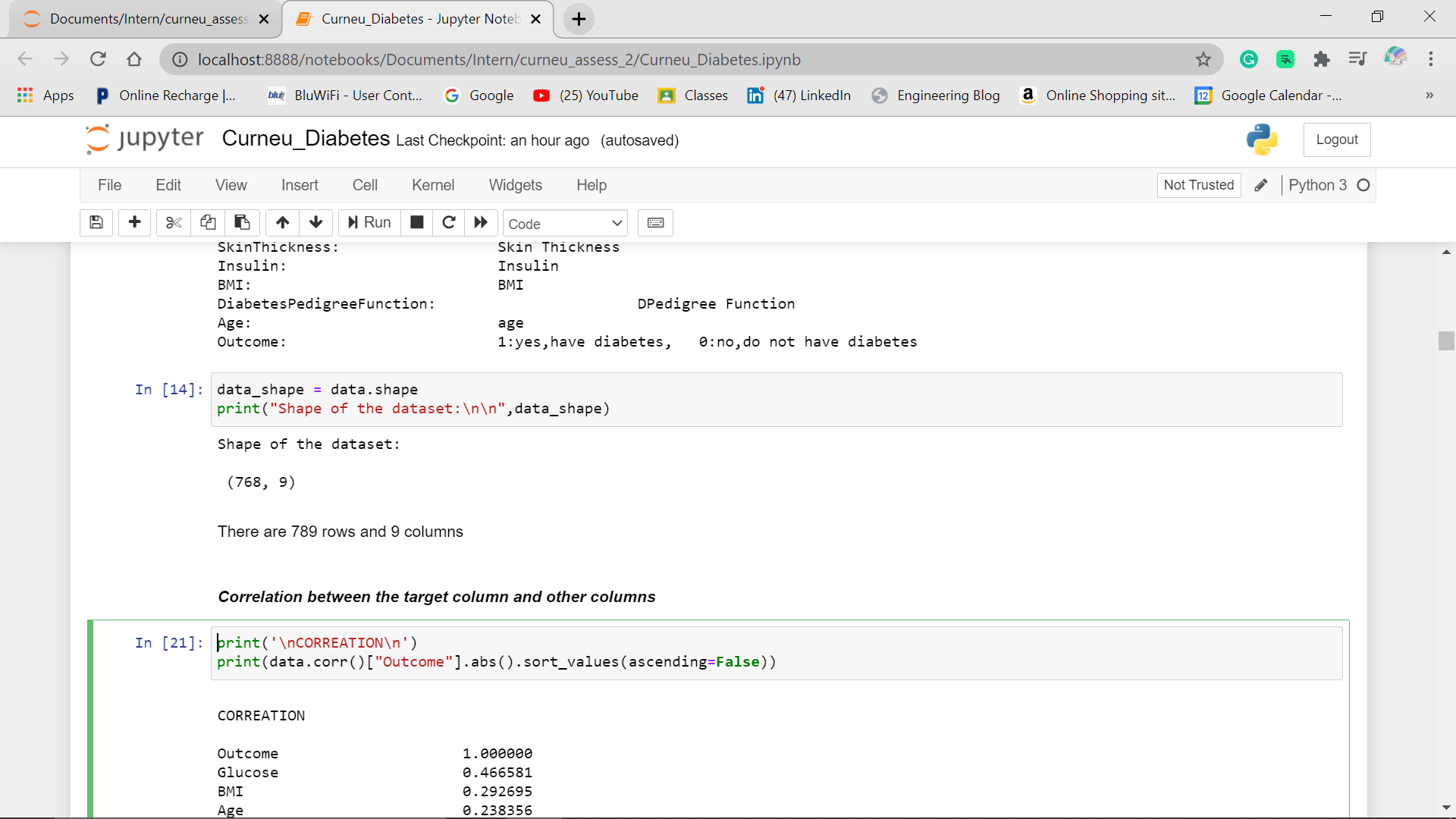
* EDA is the process of analysing and investigating the datasets and summarise their main characteristics, often by data visualisation methods
* Here, we will be analysing every attribute in dataset, along with shape of dataset. Correlation between the target and other attributes also analysed.
* **Understanding Columns:**

Here, we get to know about the values in each column in the dataset

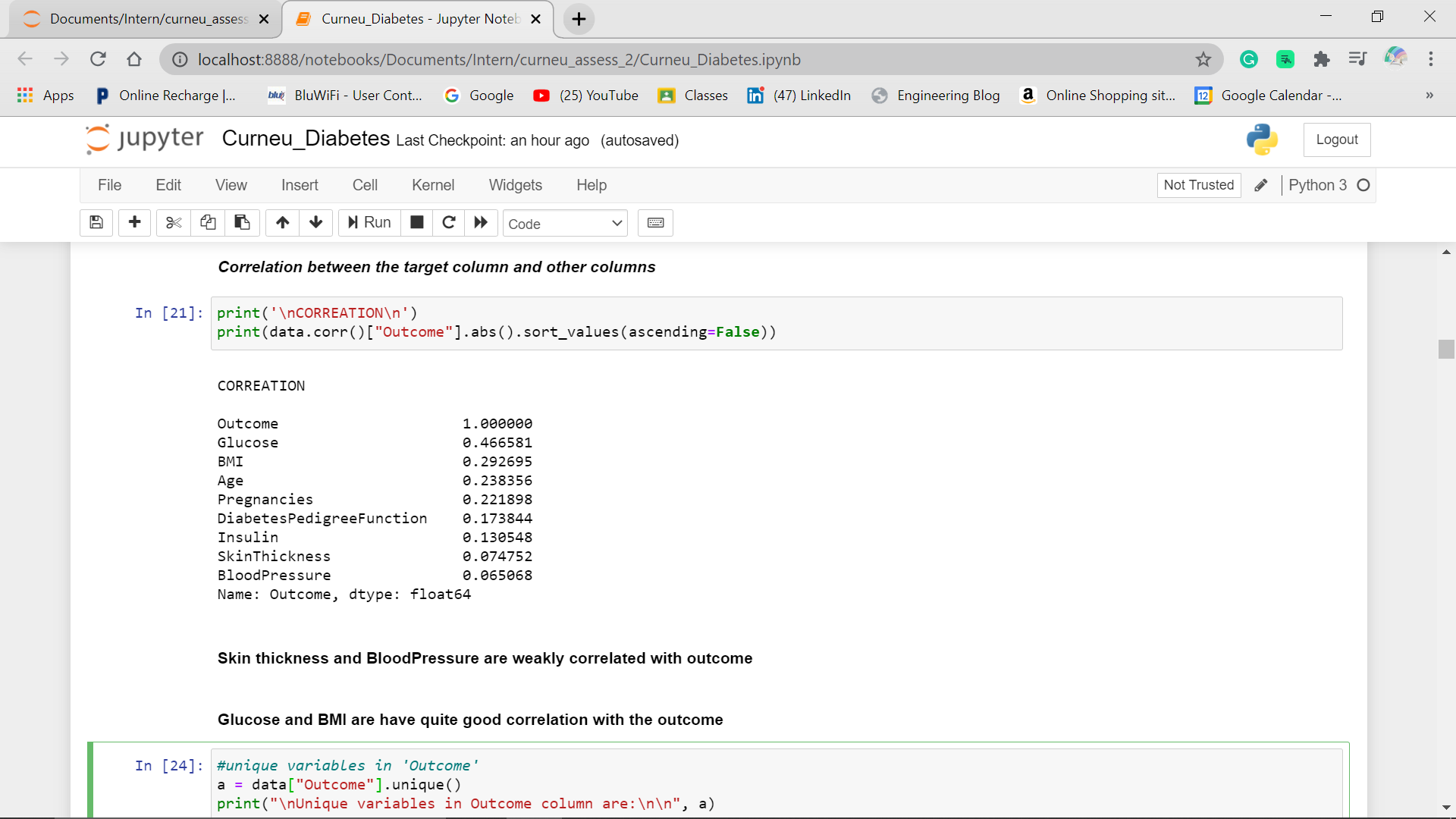
* + - In the outcome column ‘0’ refers to No and ‘1’ refers to Yes



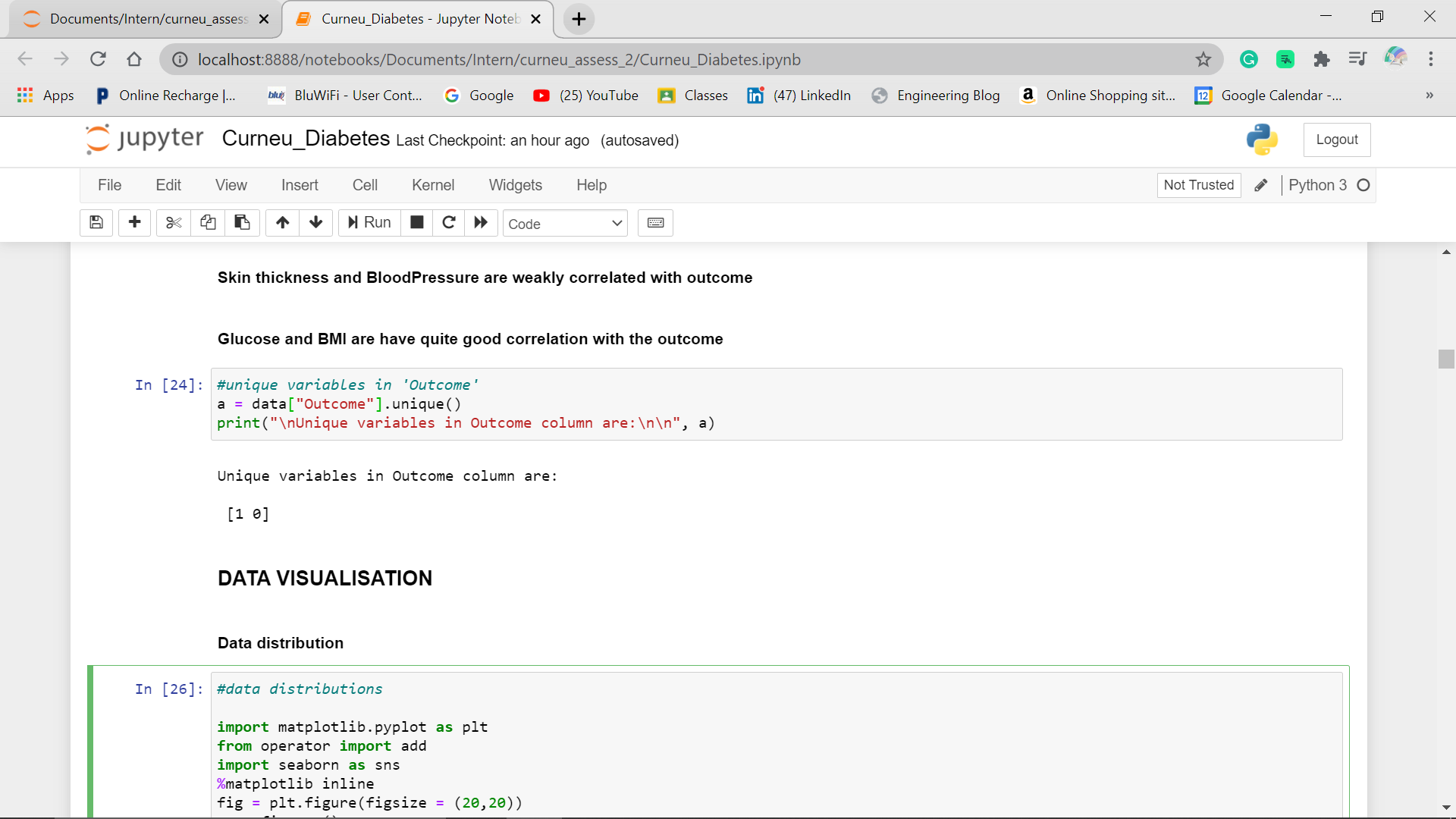
* **Shape of dataset:**
* There are 789 rows and 9 columns present in the dataset
* 9 columns such as Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, BMI, Diabetes Pedigree Function, Age and Outcome



* **Correlation between the target and other attributes:**
  + *Skin Thickness* and *Blood Pressure* are **weakly Correlated** with Outcome
  + *Glucose* and *BMI* are having quite **good correlation** with Outcome

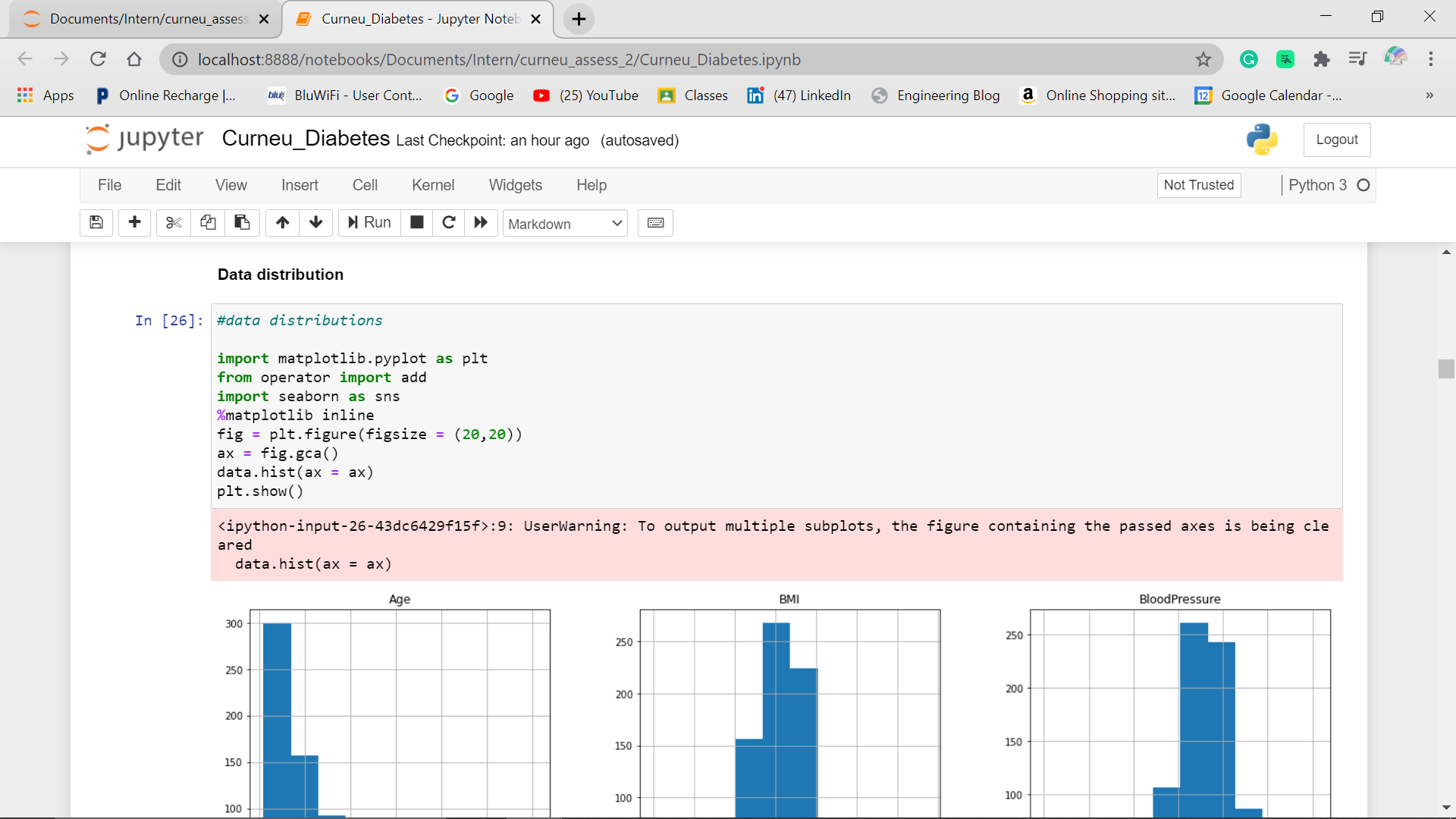


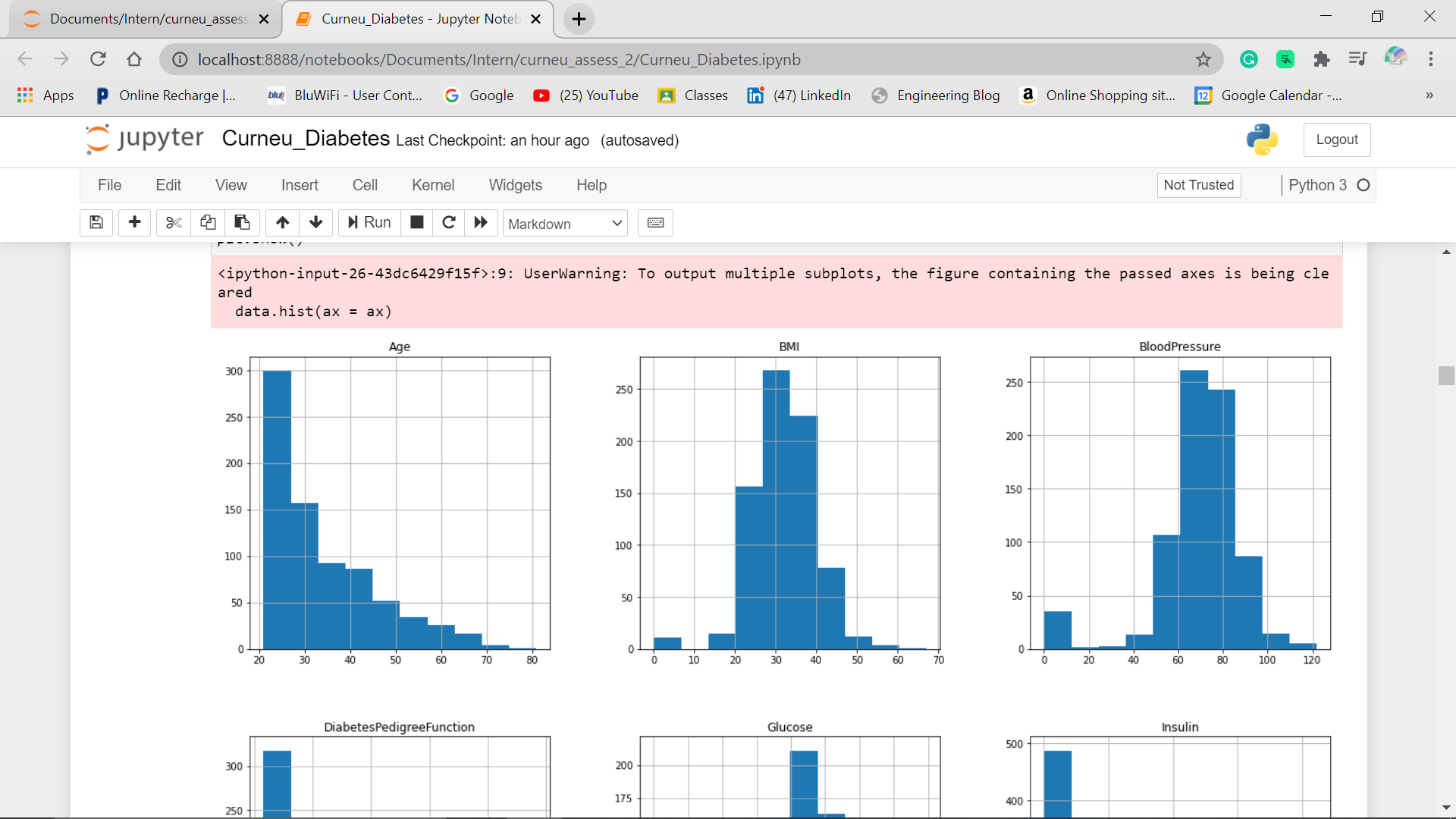
* **Unique variables in Outcome**
  + Unique variables in Outcome Column are 0 and 1

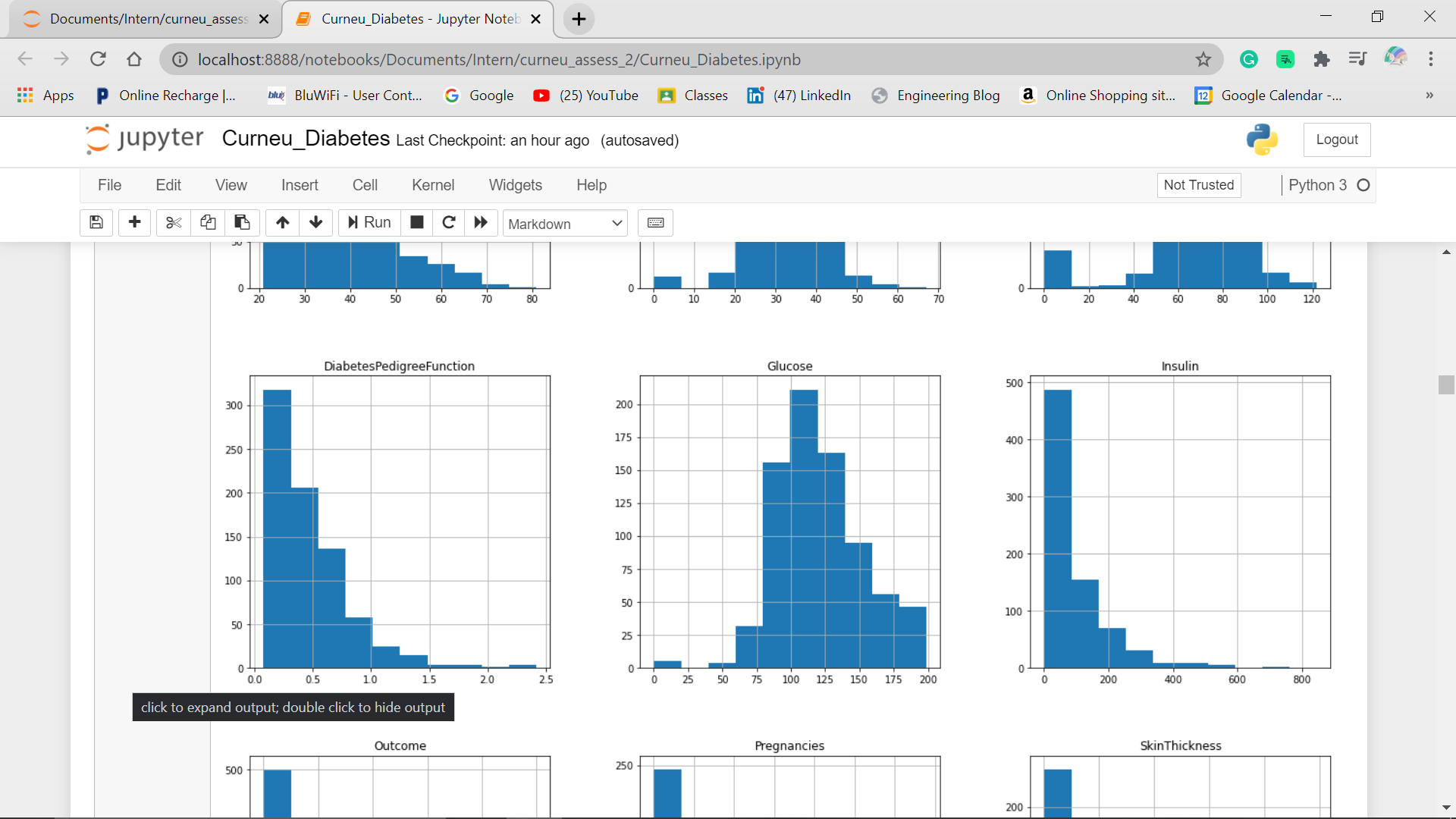


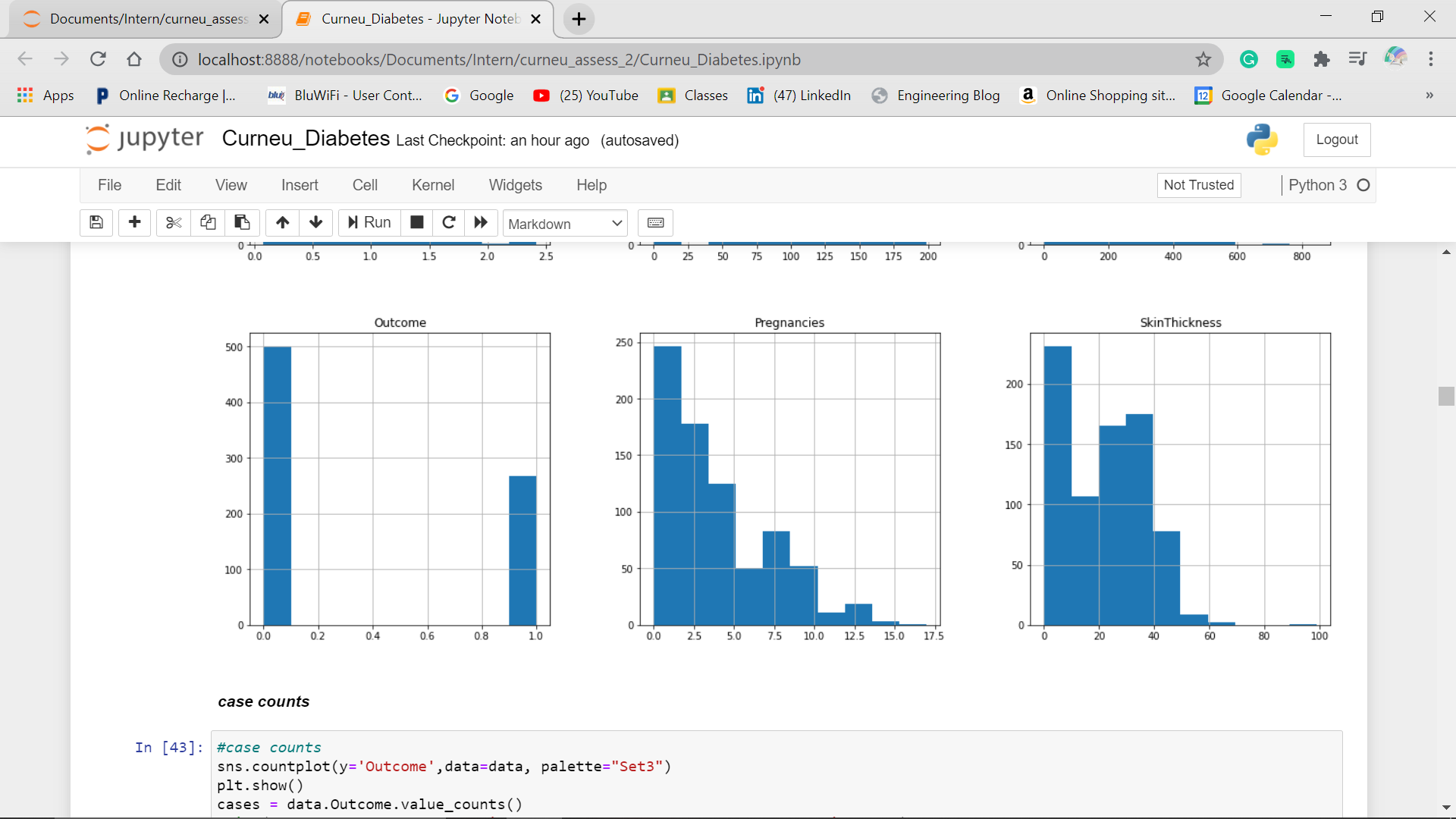
1. **Data Visualisation:**

* Data visualisation are very important as they are quickest way to understand the data patterns.
* Visualisation using Matplotlib, Seaborn can be used to check the correlation within the features and the target.
* Histograms for checking the spread and skewness
* Seaborn to quickly plot the visualisation of the whole data to check multicollinearity etc.
* **Data distribution**
* This helps us to understand the distribution of dataset for each attribute and plots give clear visualisation.
* Here, matplotlib and seaborn libraries are used to plot the graph(histogram)
* From the plot below, the dataset consists of people of age group ranging from 20 to 75
* BMI ranging from 15 to 60
* Blood Pressure from 40 to 120
* Diabetes Pedigree Function from 0.1 to 1.5
* Glucose level from 40 to 200
* Insulin from 0 to 600
* Pregnancies ranging from 0 to 16 and Skin Thickness from 0 to 70

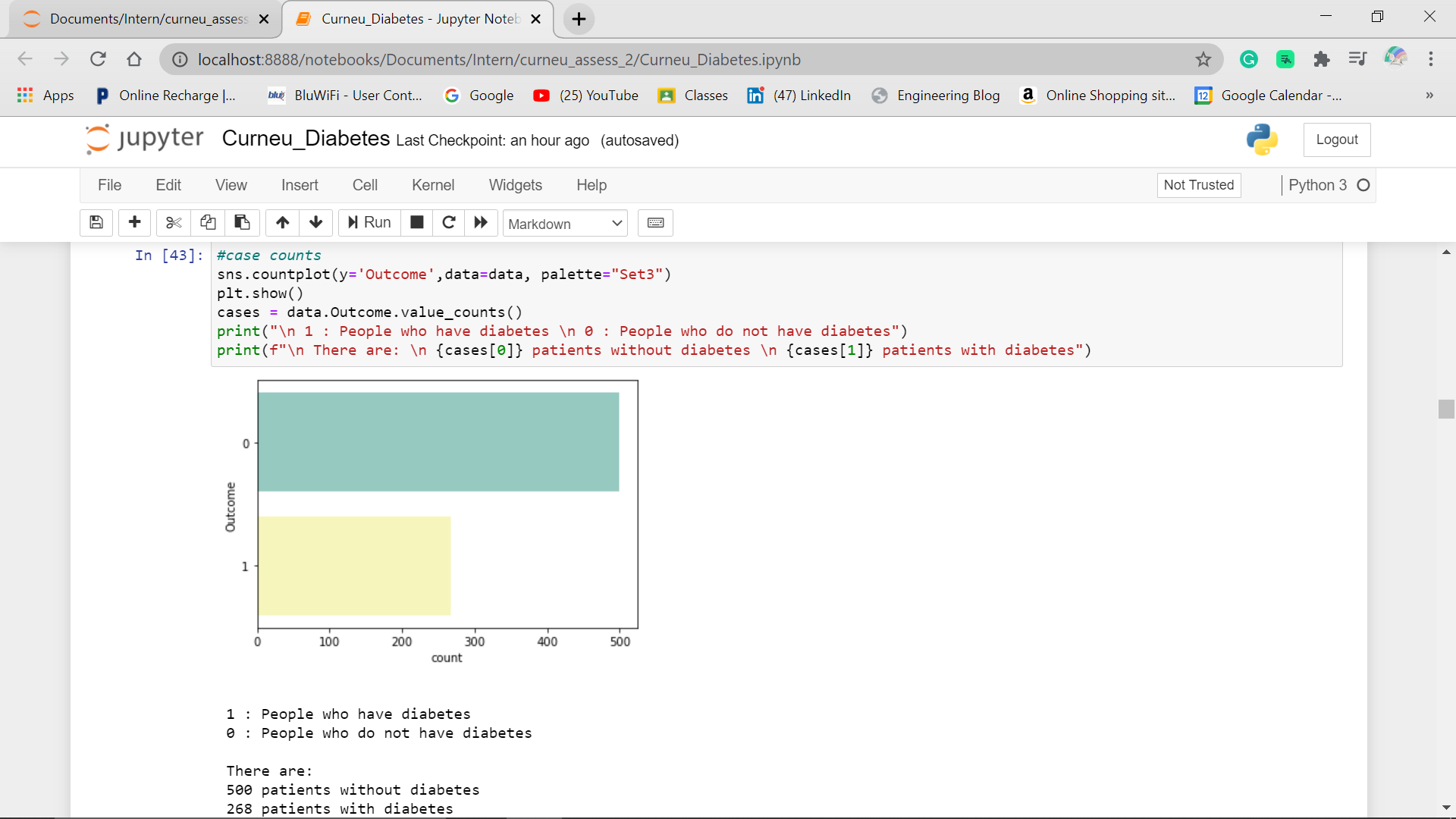




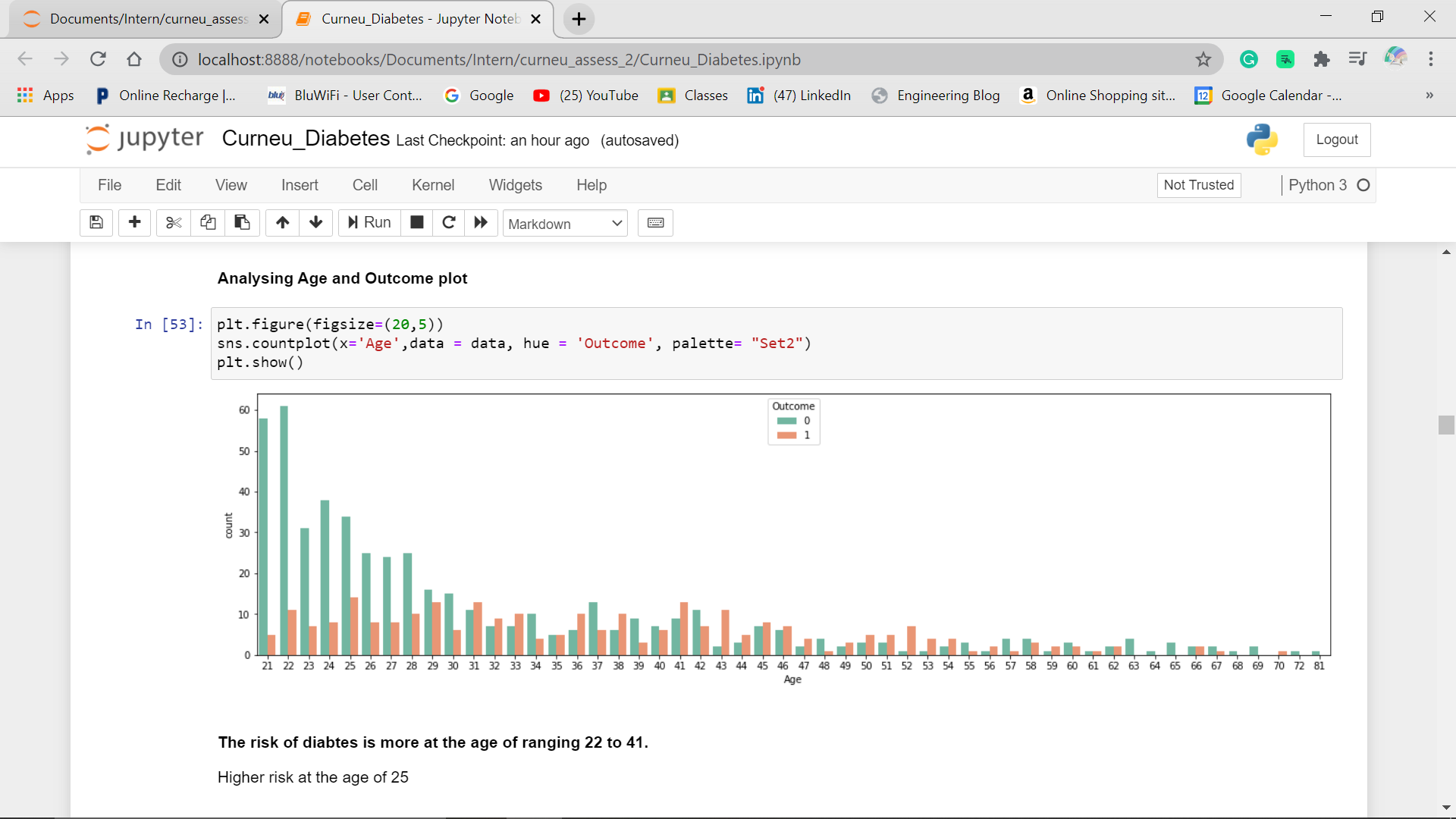




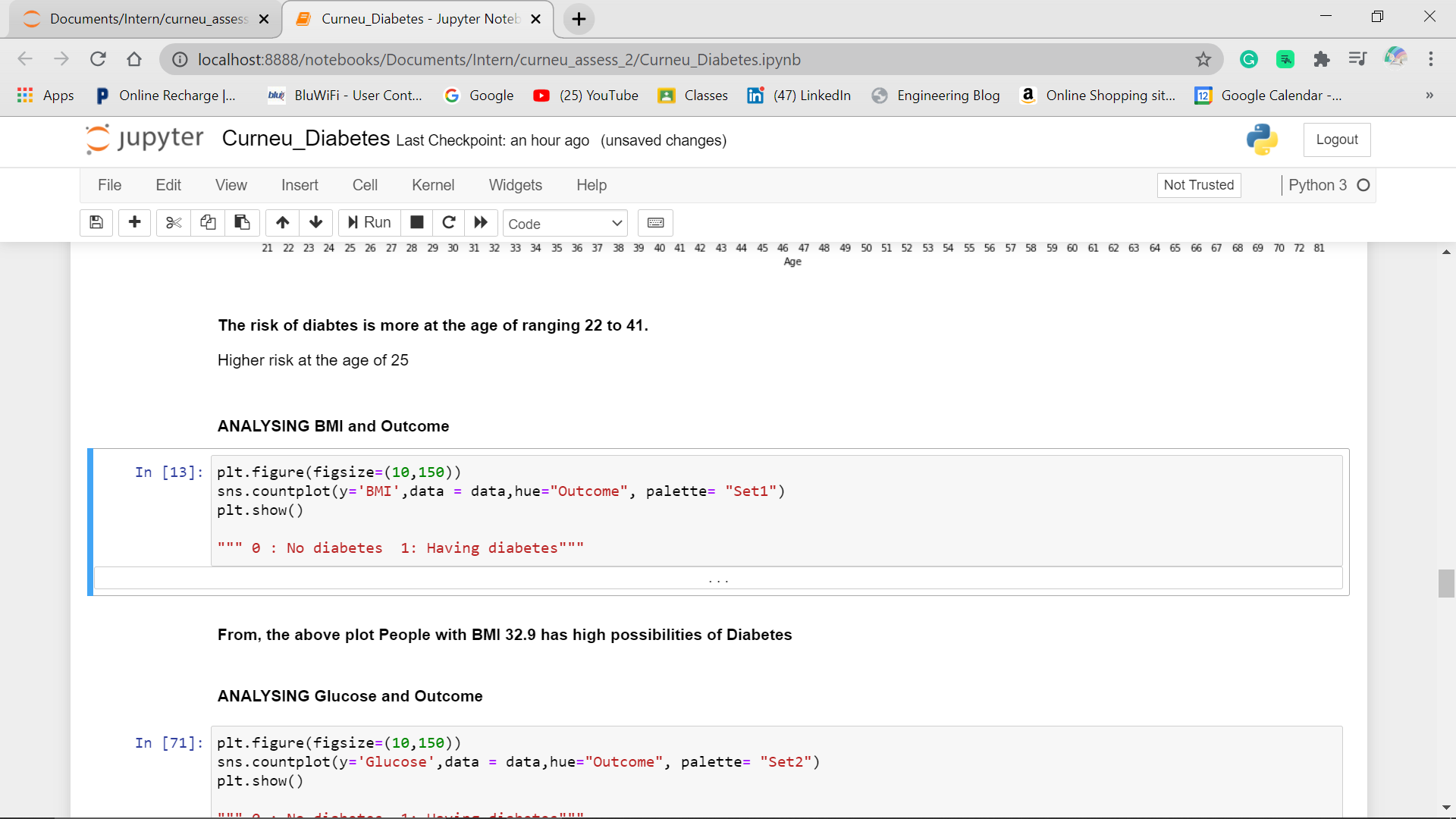
* **Case counts:**
* This helps to analyse the Diabetes patients COUNT from the dataset.
* 1 represents people with Diabetes
* 0 represents people without Diabetes
* 500 patients without heart disease
* 268 patients with heart disease

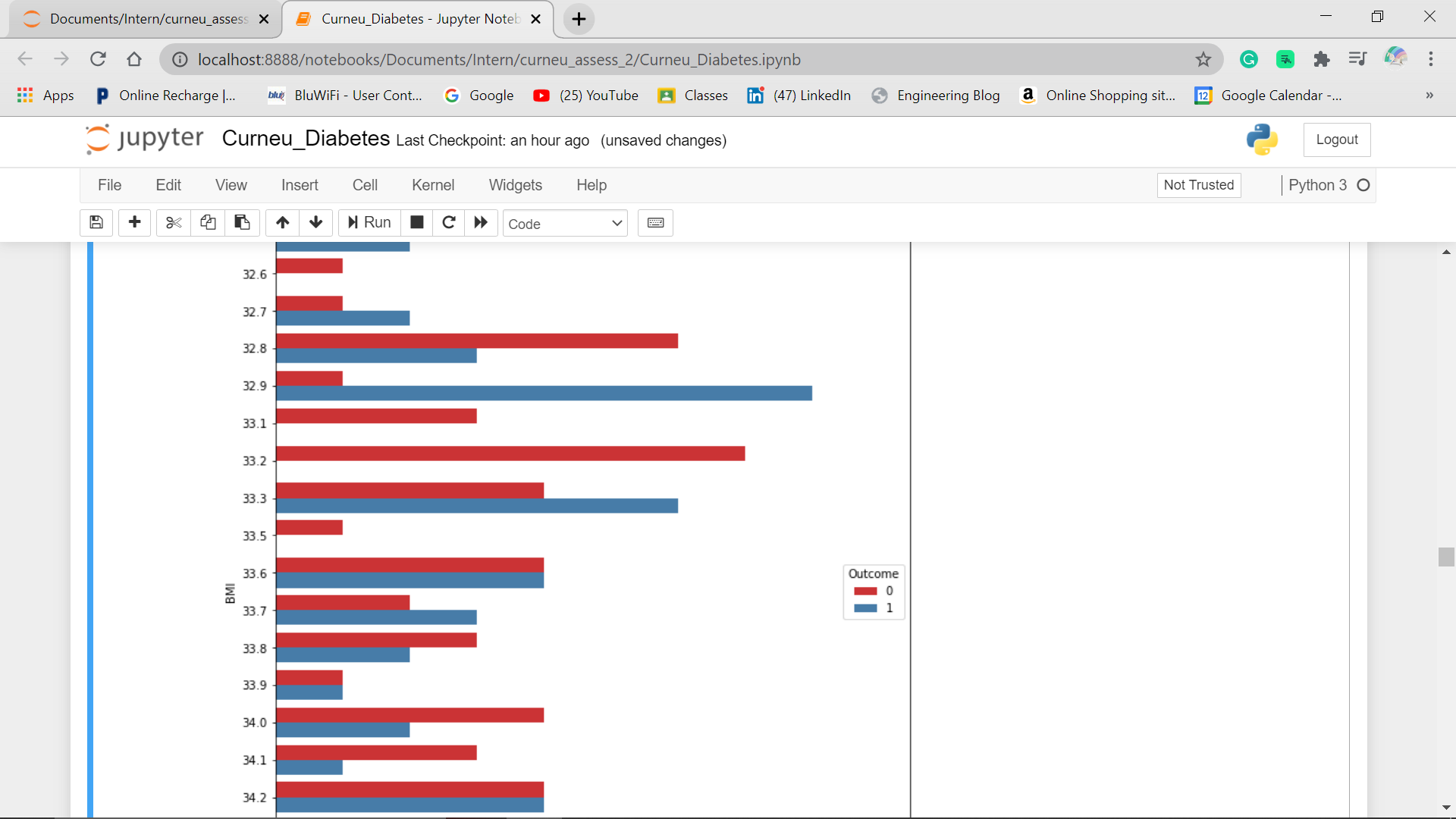


* **Age and Outcome plot:**
* This helps us to visualise that which age group has high risk of Diabetes.
* From the graph, it is evident that patients with Diabetes are at the age ranging from 22 to 41.
* Higher risk the age of 25.

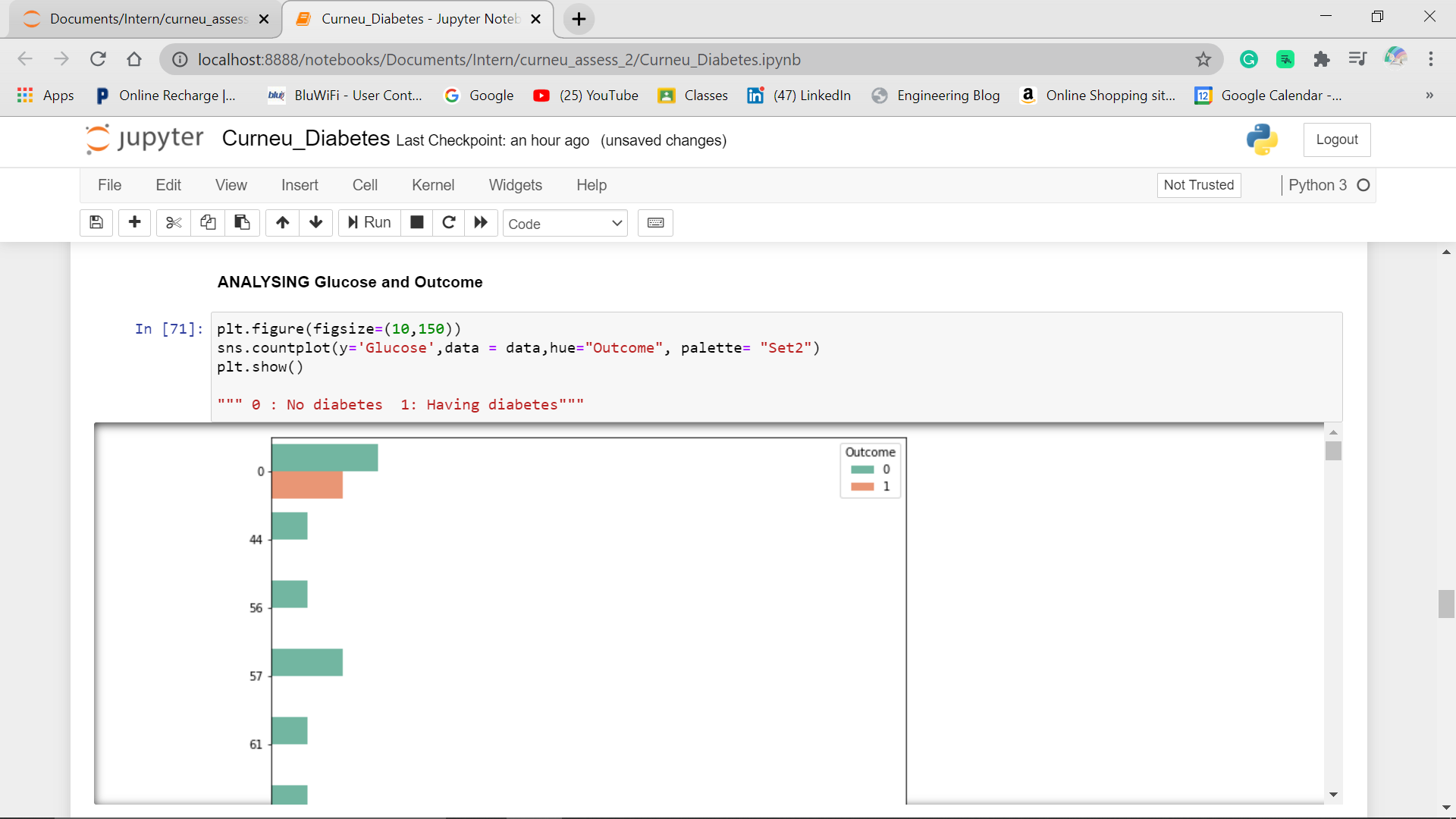


* **Analysing BMI and Outcome:**
  + People with BMI of about 32.9 have high possibility of getting Diabetes





* **Analysing Glucose and Outcome:**
  + People having Glucose level of ranging 90 to 158 have risk of having diabetes
  + People with Glucose level of 125 and 158 have high risk of getting Diabetes



From the above plots the risk factors of Diabetes are:

* BMI
* Glucose

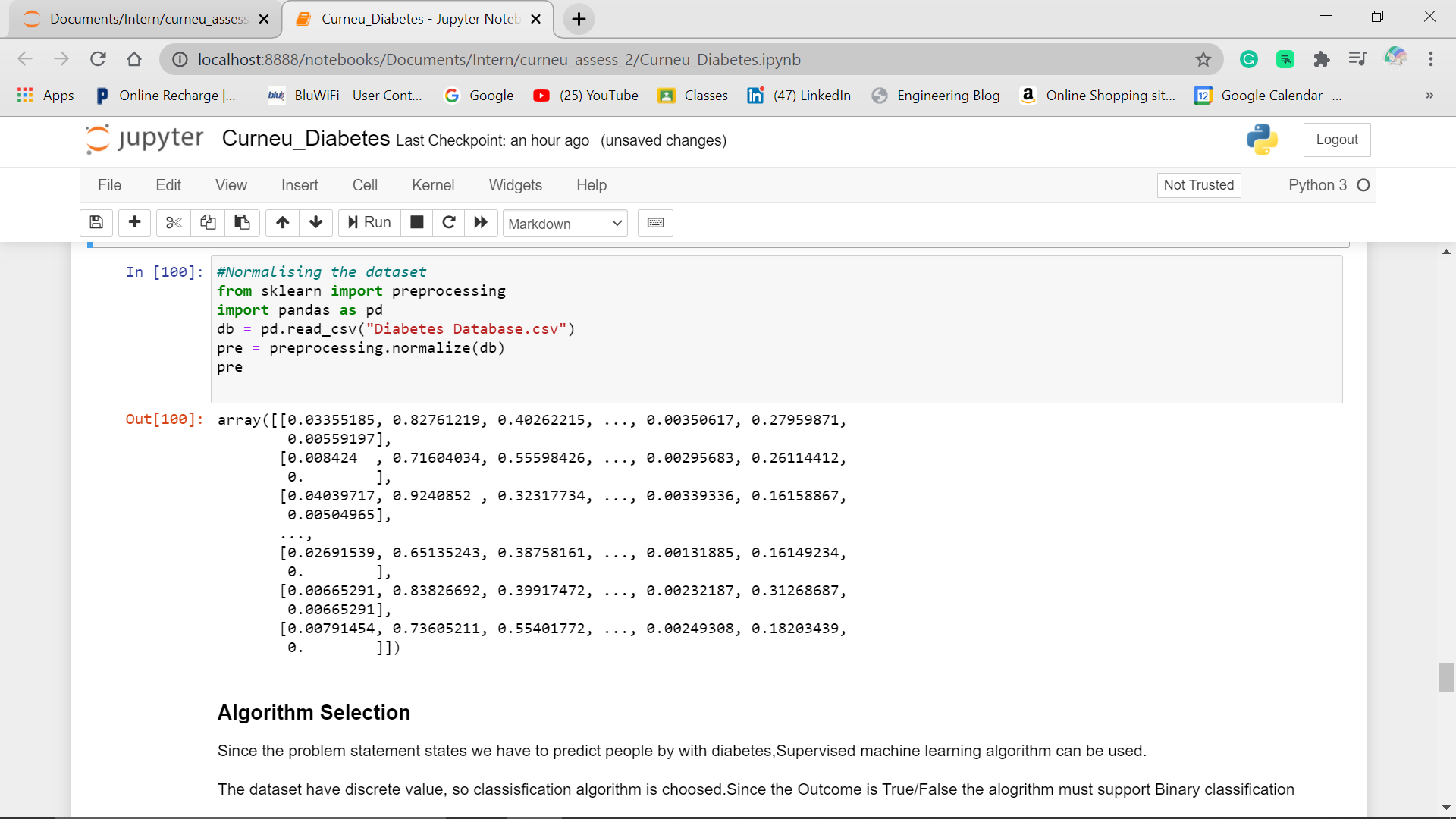
1. **Prepare data:**

Once we understand the data, what it has and what it looks like, we will have to transform the data in order to make it suitable for algorithm process and work more efficiently in order to make accurate results

* **DATA CLEANING:**

Since the dataset has no missing or null values. We can proceed with next step.

* **Normalisation:**

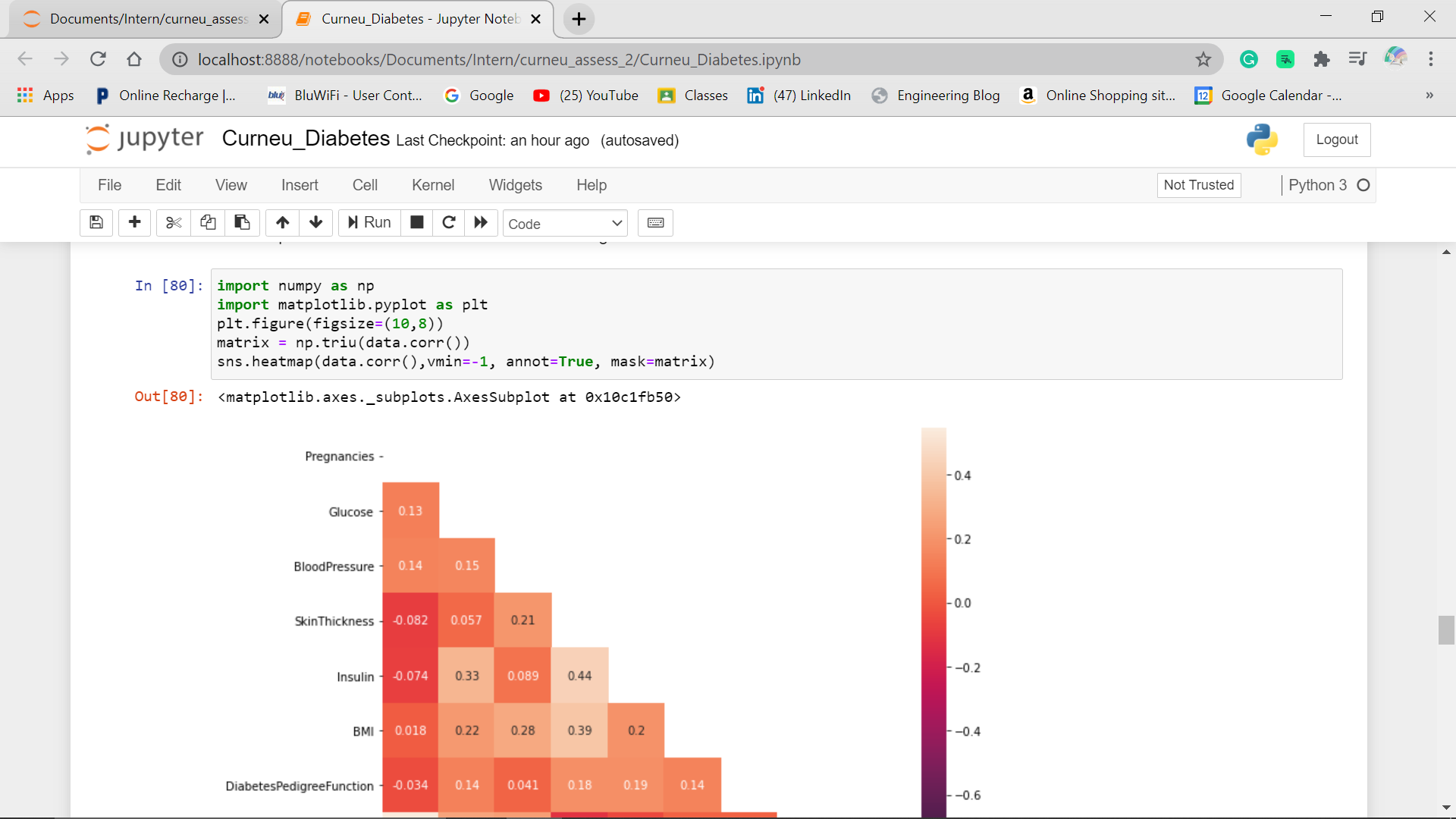


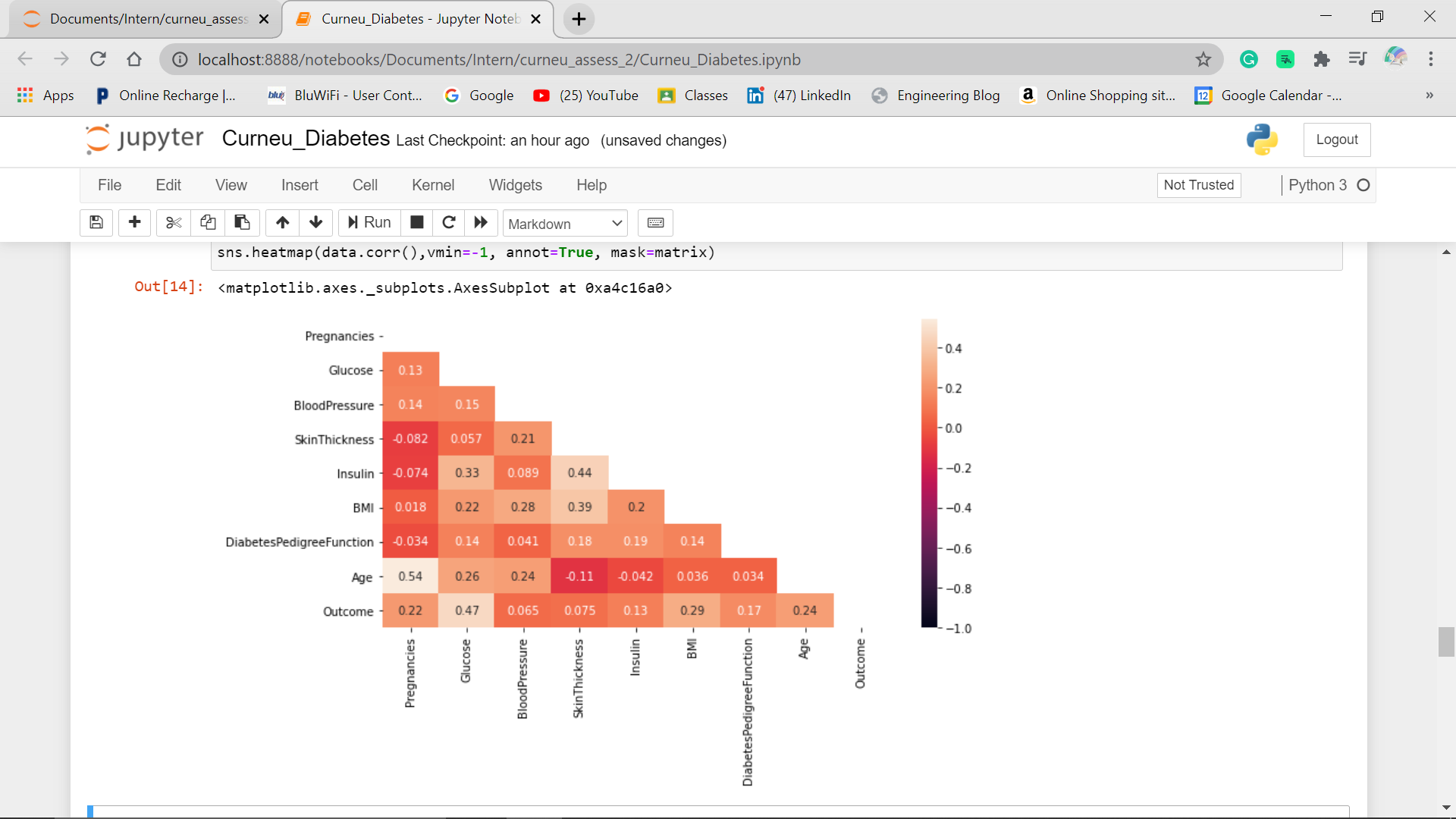
* **FEATURE SELECTION:**

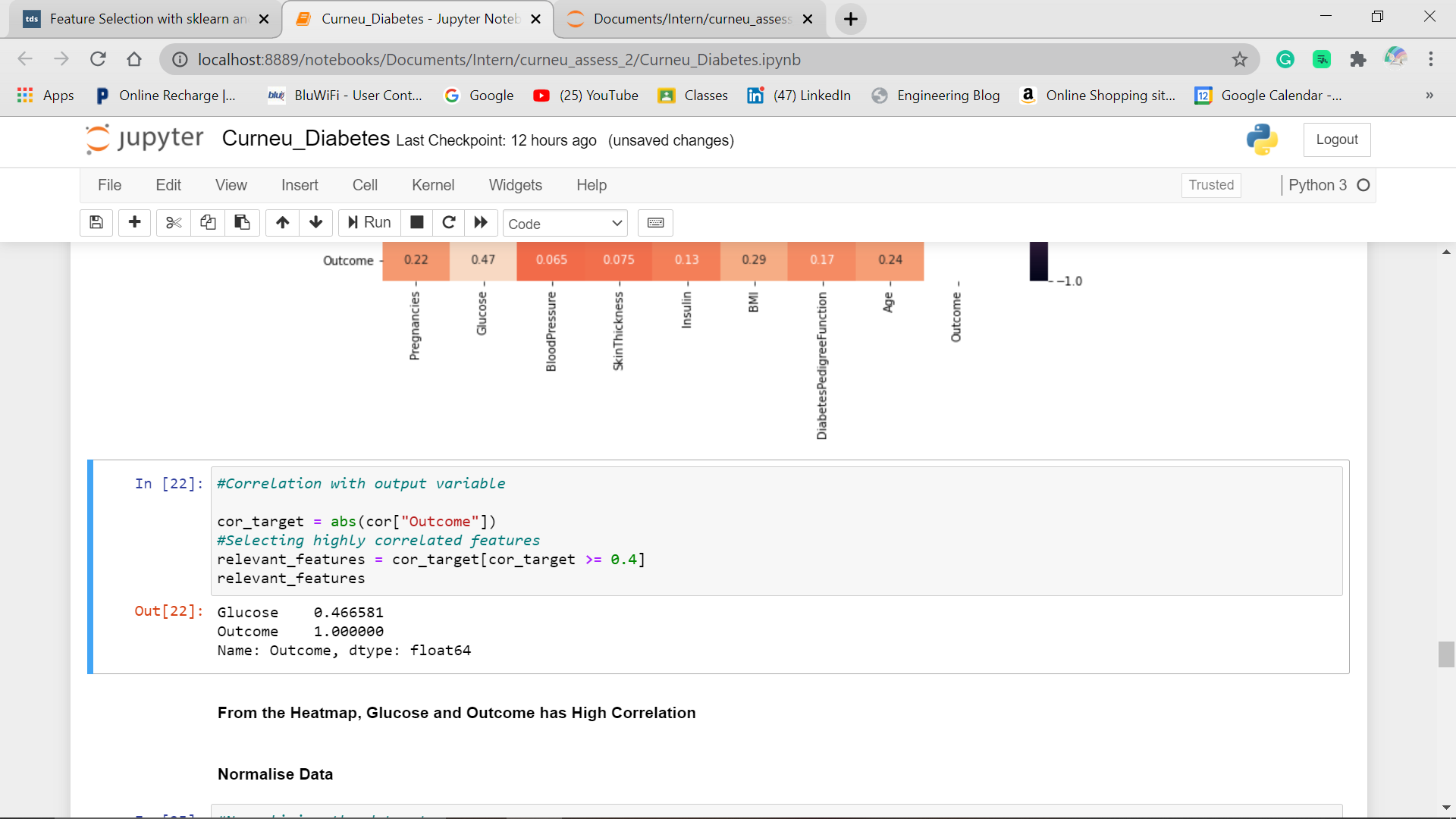
Feature selection is the process of selecting a certain number of most useful features which will be used to train the model. Reducing features saves lot of time.

Here, Feature selection is done by using *Pearson Cor-relation (Filter Method)*

* + **Correlation heat map:**
* Seaborn library is used to plot Heatmap
* Correlation ranges from -1 to +1. Larger the number and darker the colour, the correlation is higher between variable
  + 0 - no correlation
  + +1 - positive correlation
  + -1 - negative correlation
* From the Below Heat map it is evident that**, Glucose has highest correlation with Outcome than the rest**







1. **Evaluate Algorithms:**

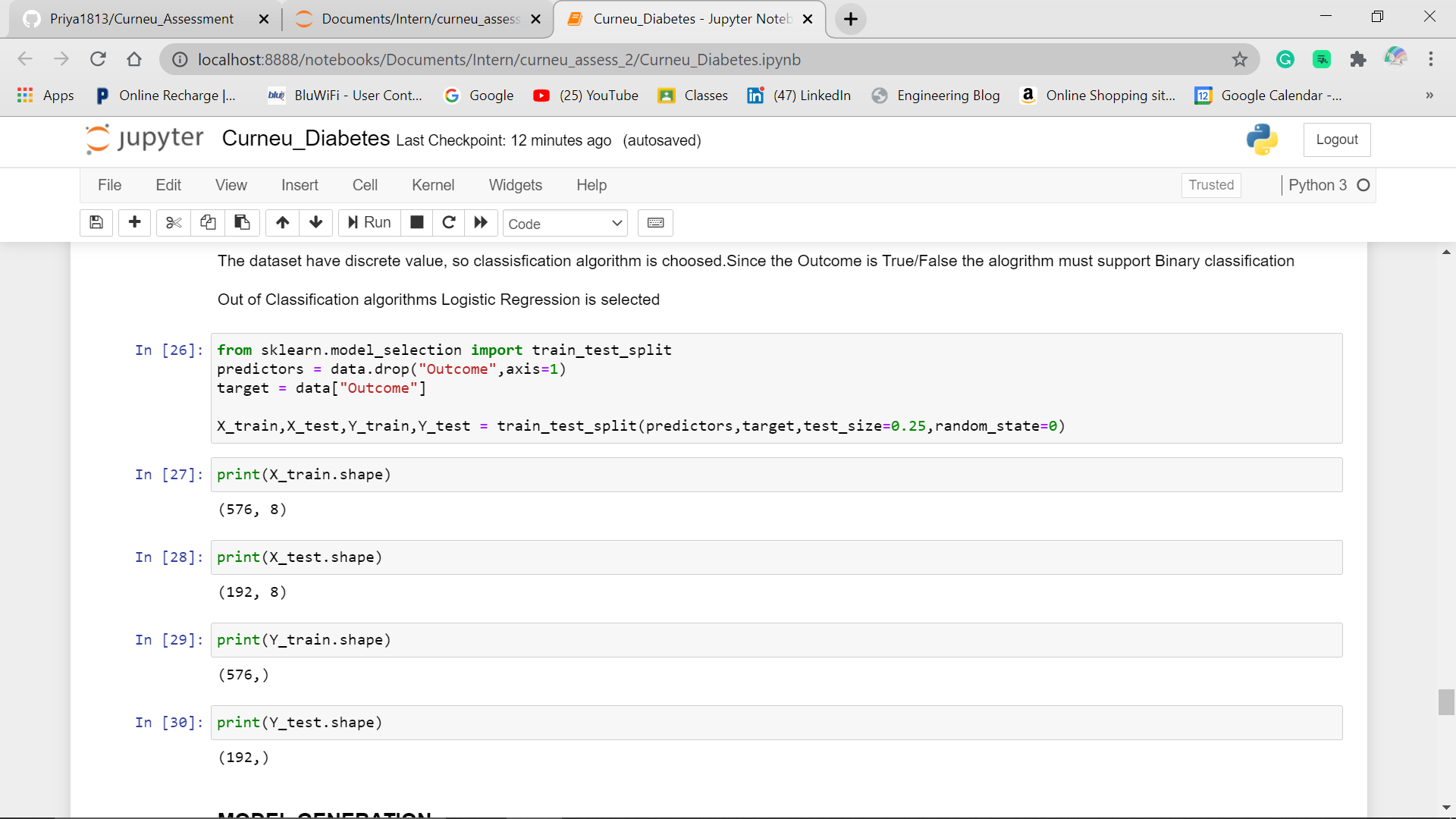
Once our data is pre-processed, we can process to check the performance of various regression/classification algorithm based on the problem statement.

**ALGORITHM SELECTION:**

* Since the problem statement states we have to predict Patients with Diabetes, Supervised machine learning algorithm can be used.
* The dataset has discrete value, so classification algorithm is selected. The algorithm must support Binary classification.
* Out of Classification algorithms Logistic Regression is selected.

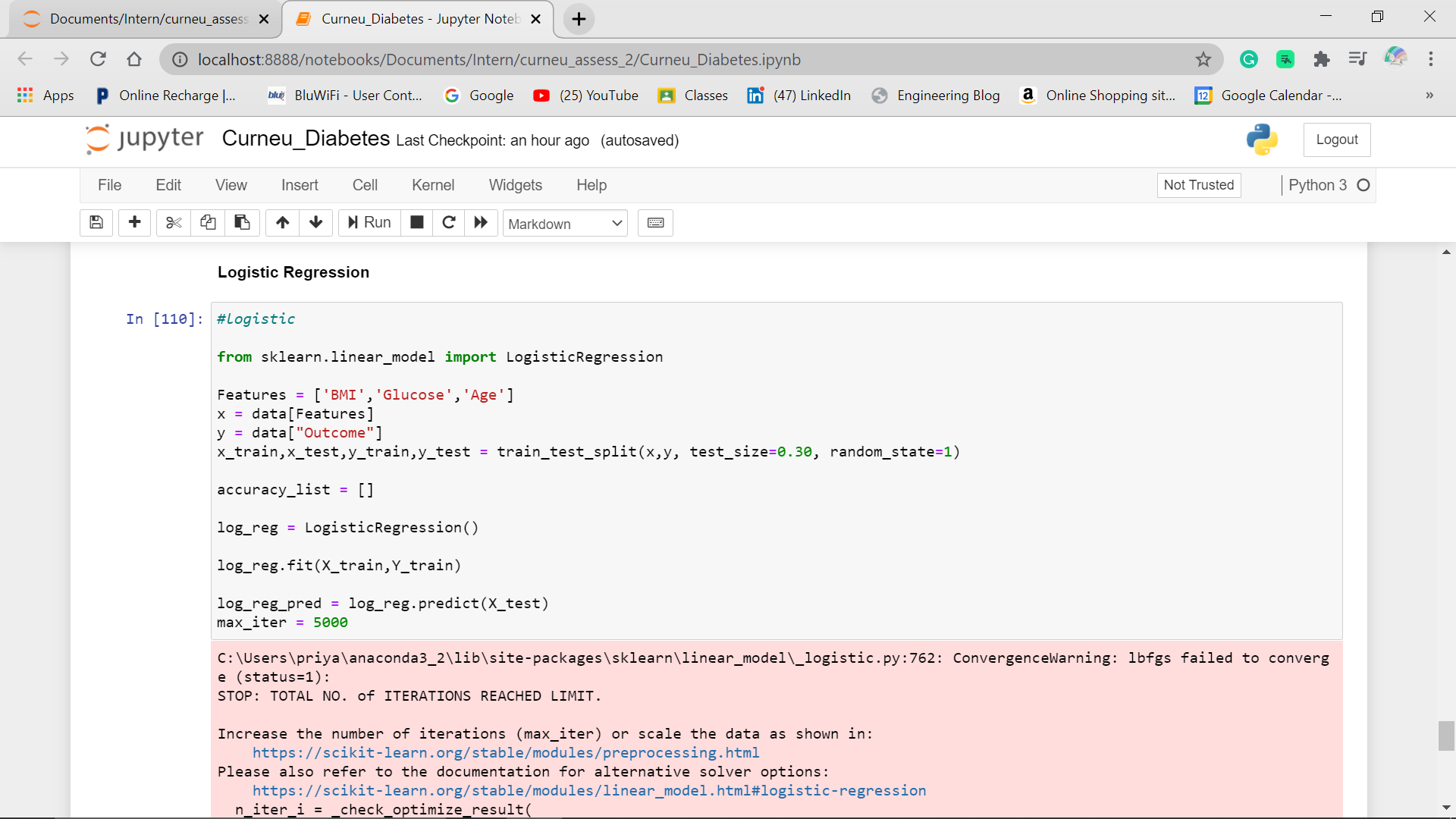
**Training and testing split:**

* We are splitting the dataset into:
  + 75% training (includes 576 rows and 8 columns)
  + 25% testing (includes 192 rows and 8 columns)
* The training model used here is Logistics Regression.



**Model generation**

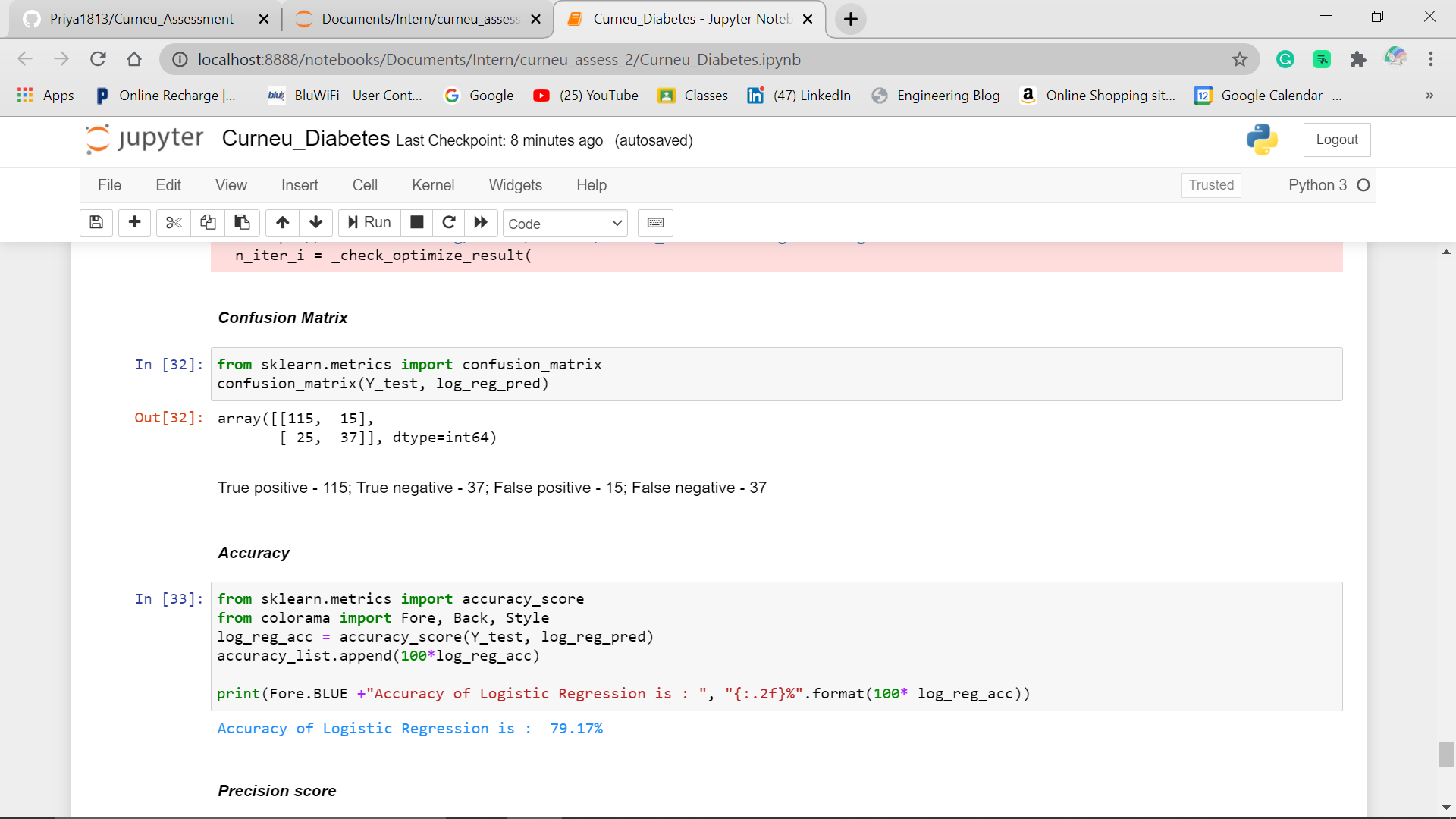
* Once the splitting of data is done, we have to fit the training data
* And then, predict the testing data
* The important features are BMI, Glucose and Age.

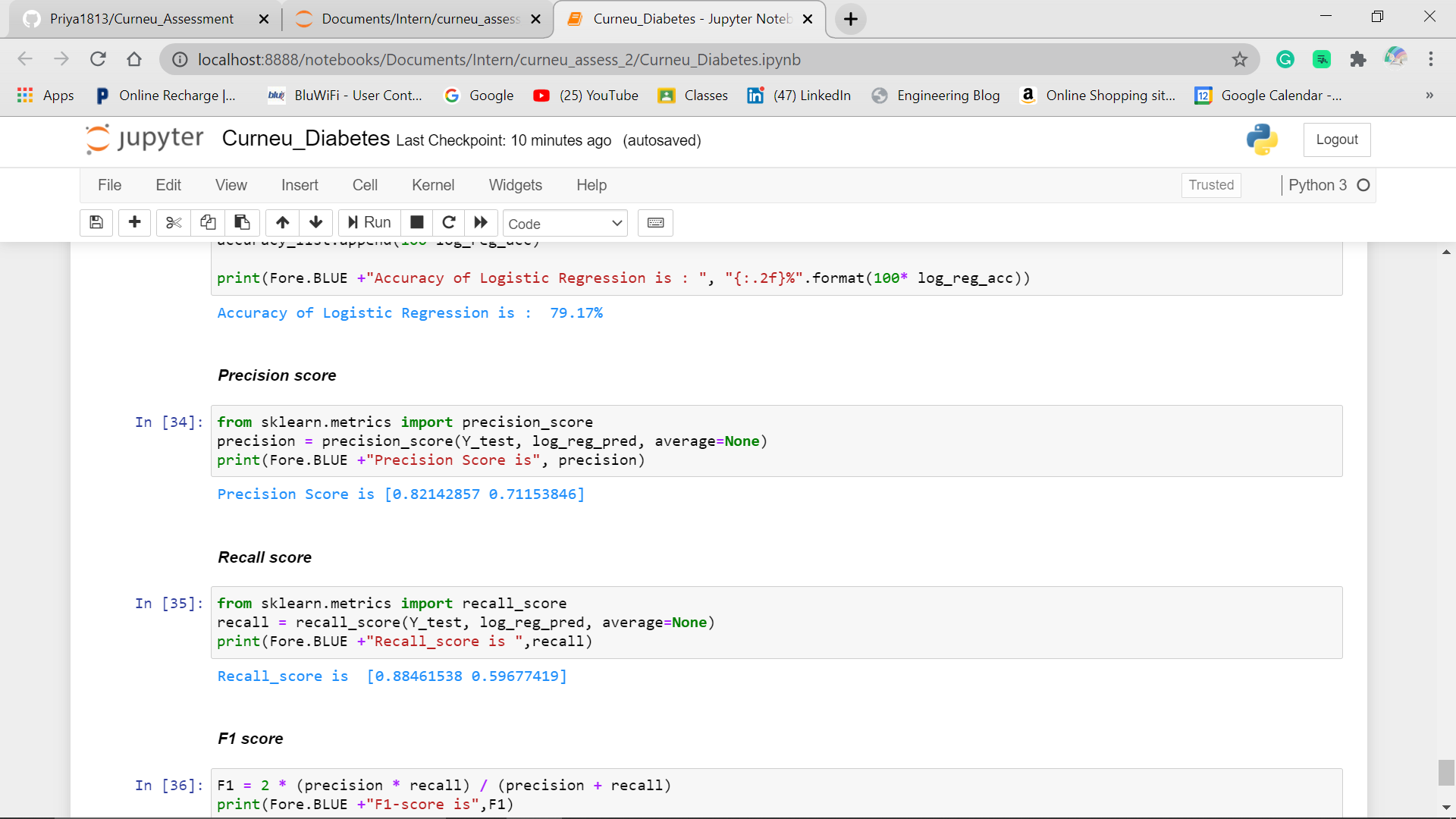


**EVALUATION METRIC:**

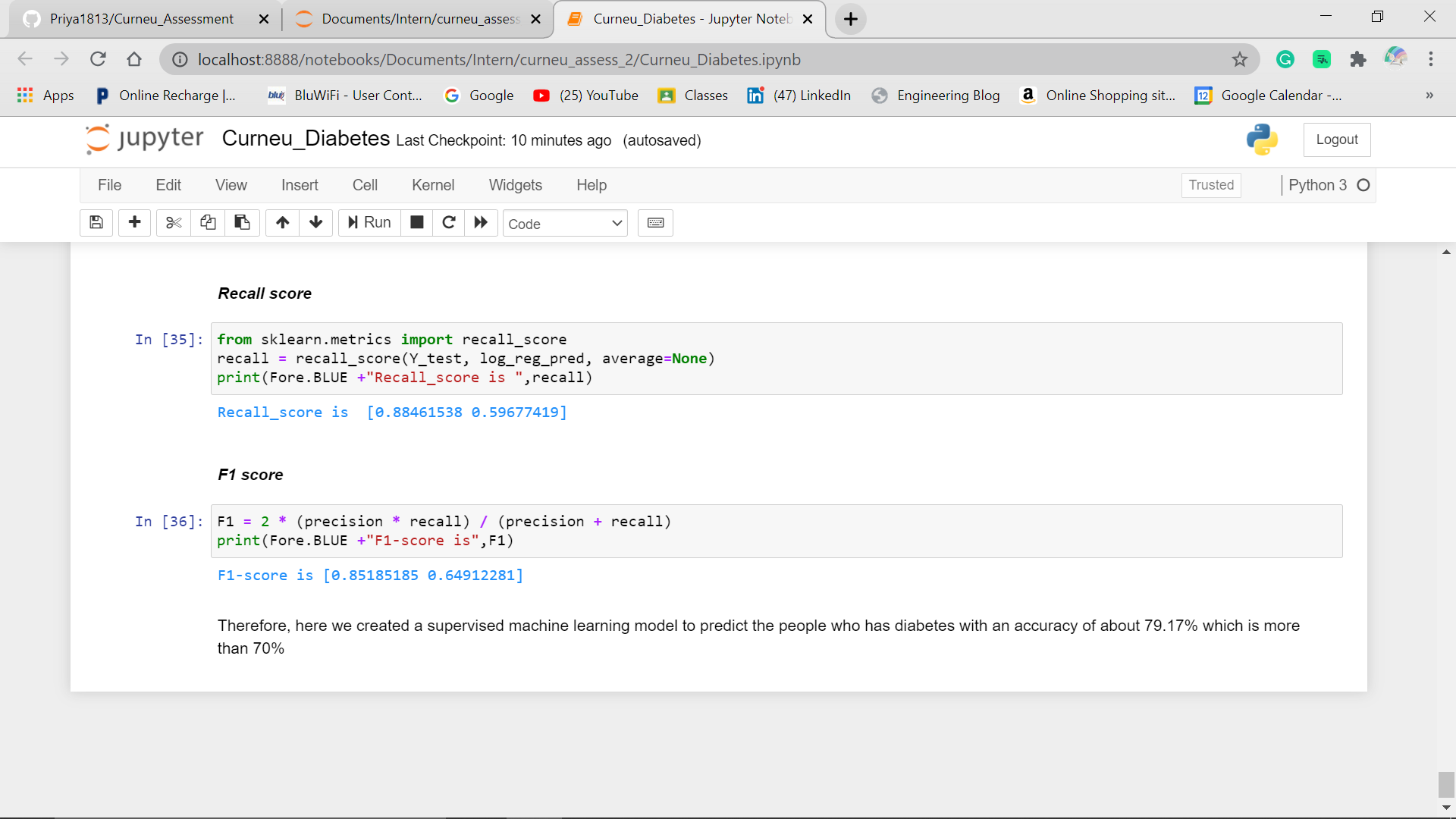
Evaluation is done to check whether your model performed well or not. The following are the most commonly used metrics.

1. **Confusion matrix**
   1. Confusion matrix is used to assess the performance of a Classification model
   2. From the output, the type 2 error (False negative) is risky because it predicts negative when it is actually positive.in our case, predicting no Diabetes when there is a chance.
2. **Accuracy and precision:**
   1. The model has the accuracy score of about **79.17 %** which is quite good.
   2. Precision refers to the percentage of prediction that are correct. The precision for class 0 and class 1 are 0.82 and 0.71 respectively.





1. **Recall and F1 score**
   1. Recall refers to the percentage of total relevant results correctly classified by the algorithm. Since the Recall score is above 0.5 for class 0 and 1, it is considered as good recall score for this model
   2. The F1 score is 0.85 for class 0 and 0.64 for class 1



Therefore, here we created a supervised machine learning model to predict the Patients with Diabetes with an accuracy of about **79.17% which is more than 70%.**

The last step is Deploying the machine learning model and integrate it into existing production environment where it can take input and return an output.