**LUNG CANCER PREDICTION**

**AIM:** PREDICTION OF LUNG CANCER BY USING SVM ALGORITHM.

**INTRODUCTION**  : Lung Cancer is a most effected disease nowadays , most of the people are effecting due to the cause of Lung Cancer, in this dataset we have a huge information about Lung Cancer and it causes , the dataset contains columns like Gender , Age , Smoking, Anxiety , Alcohol Consuming , Coughing ,Shortness of Breath , Chest Pain , Lung Cancer….

**STEPS TO CLASSIFY LUNG CANCER :**

1. LOAD THE DATA
2. ANALYSIZE AND VISUALIZE THE DATASET
3. MODEL TRAINING
4. MODEL EVALUATION
5. TESTING THE MODEL

**INPUT:** Gender, Age, Smoking, Anxiety, Fatigue, Alcohol Consuming, Coughing, Shortness of breath, Chest Pain.

**OUTPUT:** Whether the Person is effected withLung\_ Cancer or not.

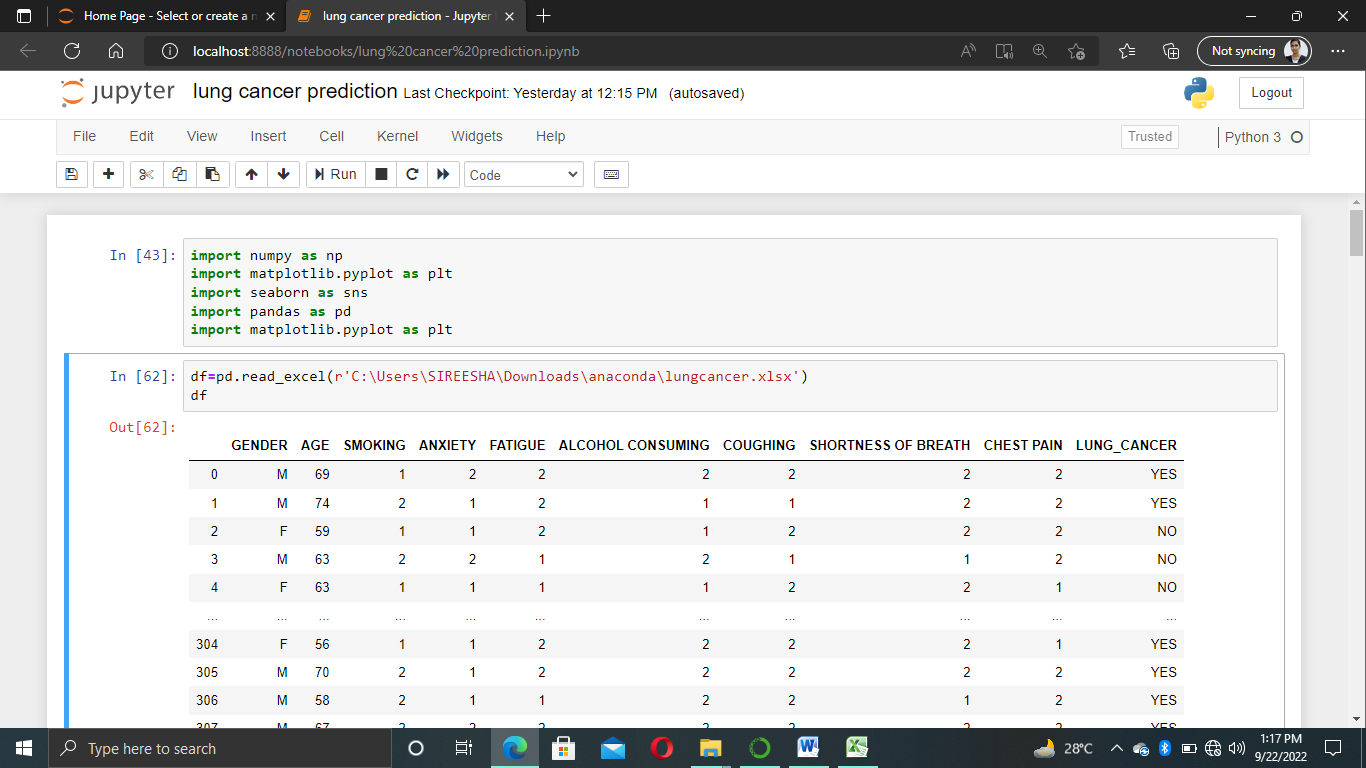
**EXPLANATION :** Will go through step by step

**STEP1 : Loading the data :**

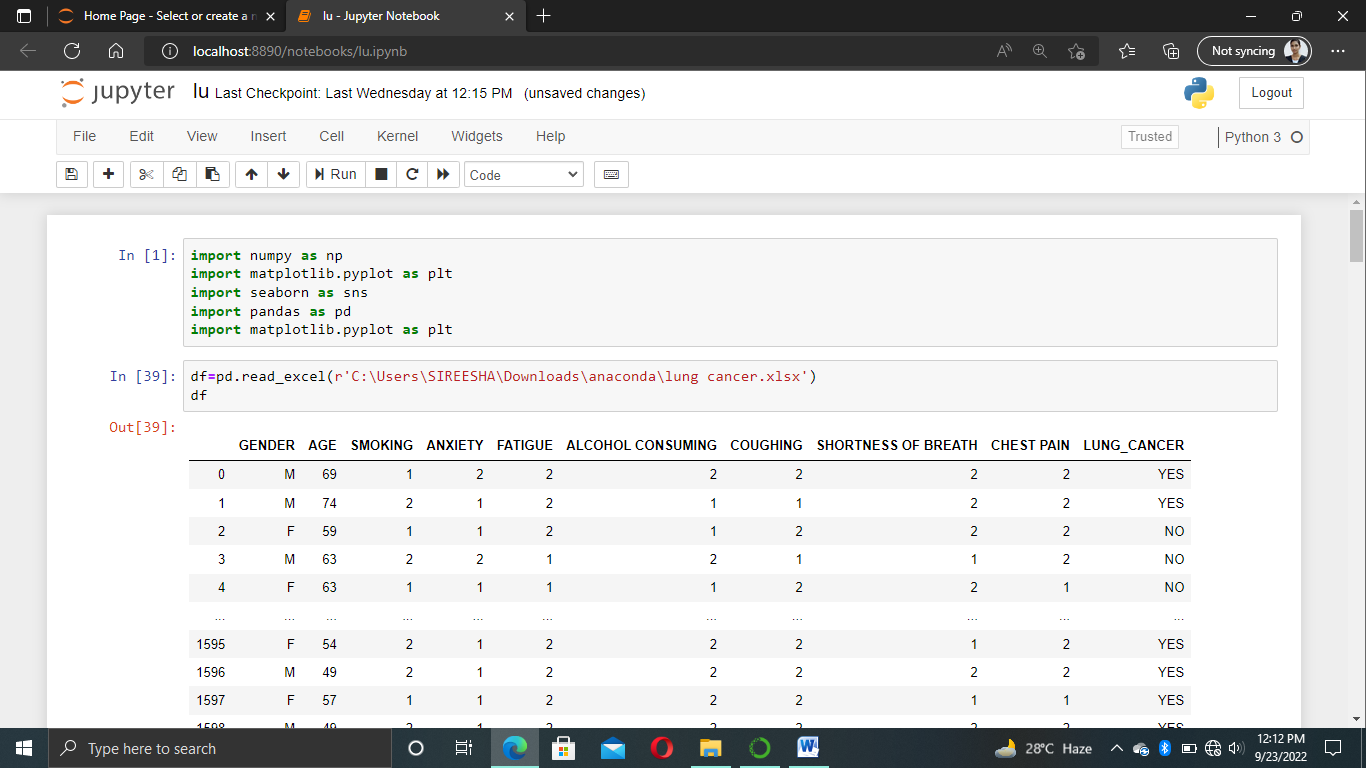
To load the data we have to import necessary packages like numpy ,pandas, seaborn, matplotlib….etc .

* Numpy will be used for any computational operations.
* we’ll use matplotlib and seaborn for data visualization .
* we use pandas for loading data from various sources like local storages , databases , excel , csv files ……etc.
* Next we load the data using pd. read\_csv () . and set the column name as for the iris data information .
* pd.read\_csv reads CSVfiles.csv transfer comma separated value.
* df.head() only shows the first five rows from the dataset table -All the numeric values.

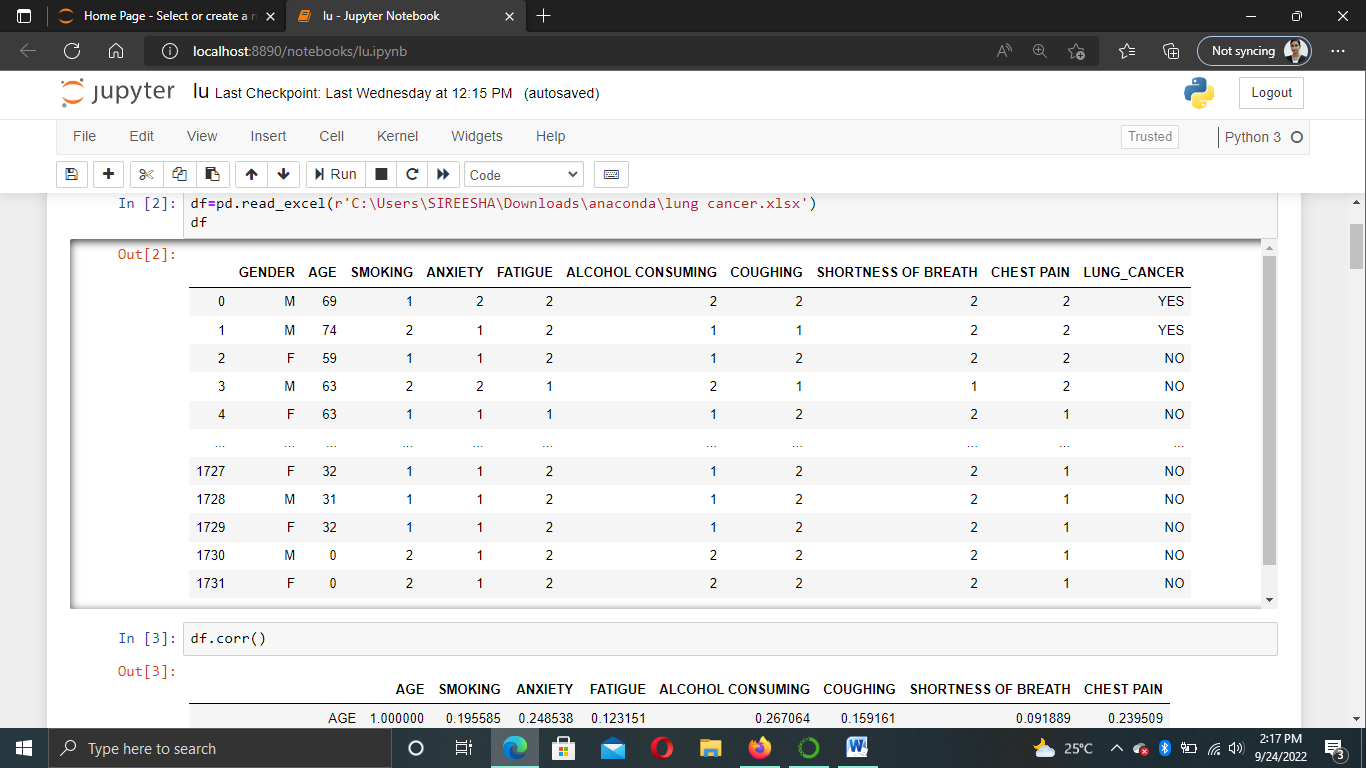
Importing necessary packages:



Loading the data:



OUTPUT:



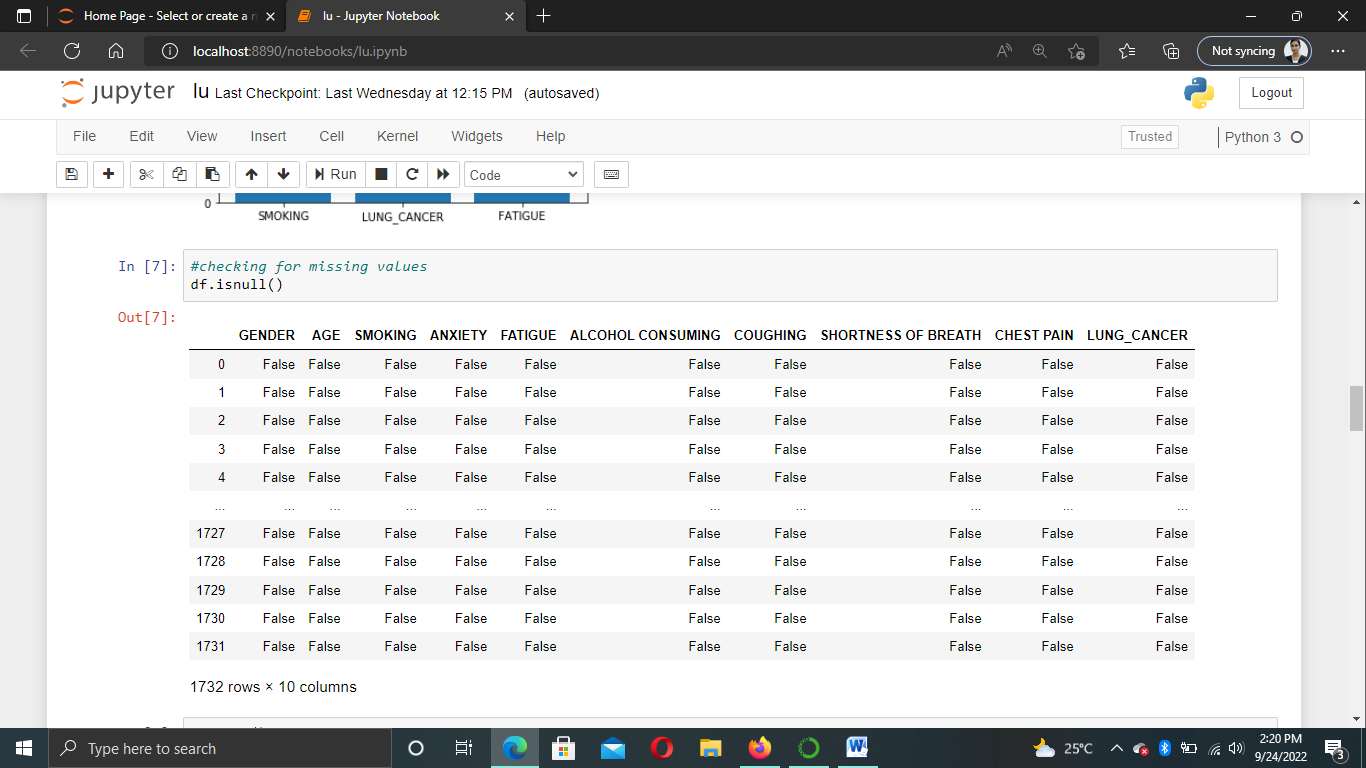
**STEP2 : ANALYSING AND VISUALIZE DATASET**

* Let’s see information…
* Before that we have done some basic operations like describe (). , info () , head functions and tail functions droping function , indexing and slicing ..etc
* Later we prepare the data in a perfect manner without any null values that means we cleared the data
* we have built some model for the dataset based on that data we do some basic operations and analyzed the data
* After that based on analization we visualized the data with seaborn pairplot, matplotlib ….etc

HANDLING THE DATA:

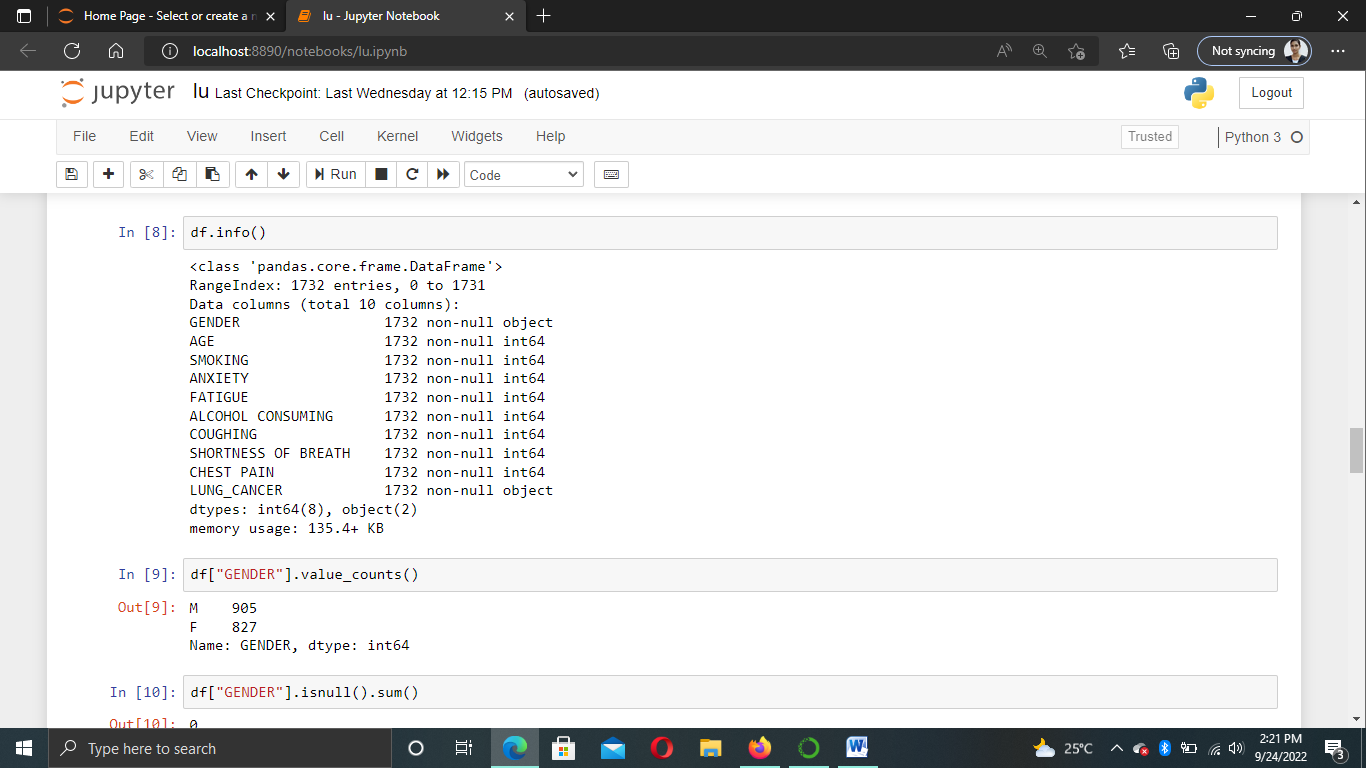
* Checking the missing values amd filling are removing the empty values
* First we have to check it the dataset contains empty values or not

Code for checking the missing values:



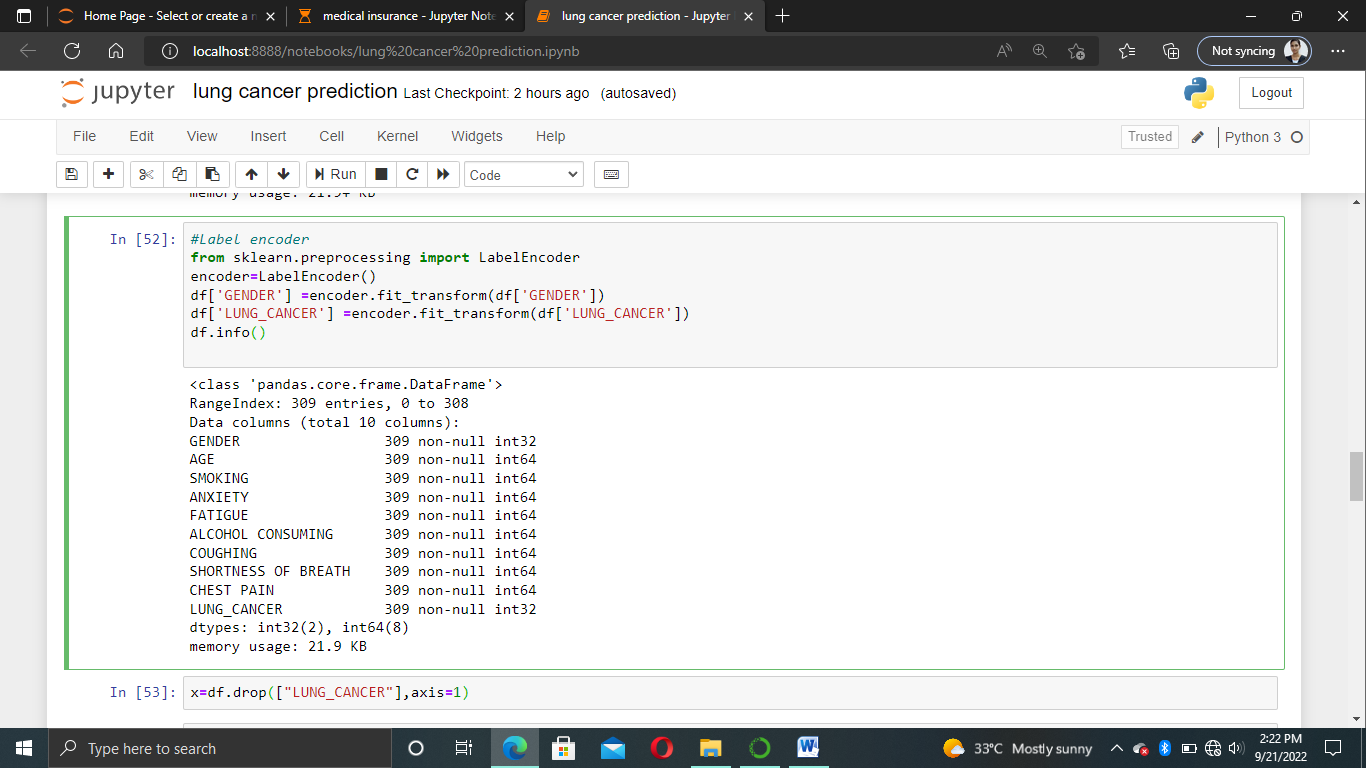
* There is no null values in the given dataset, so we don’t need to fill the values.
* Next we have check the info of the dataset. To find if there is any object type dataset.
* Code for checking info of the dataset.

INFO OF DATASET:

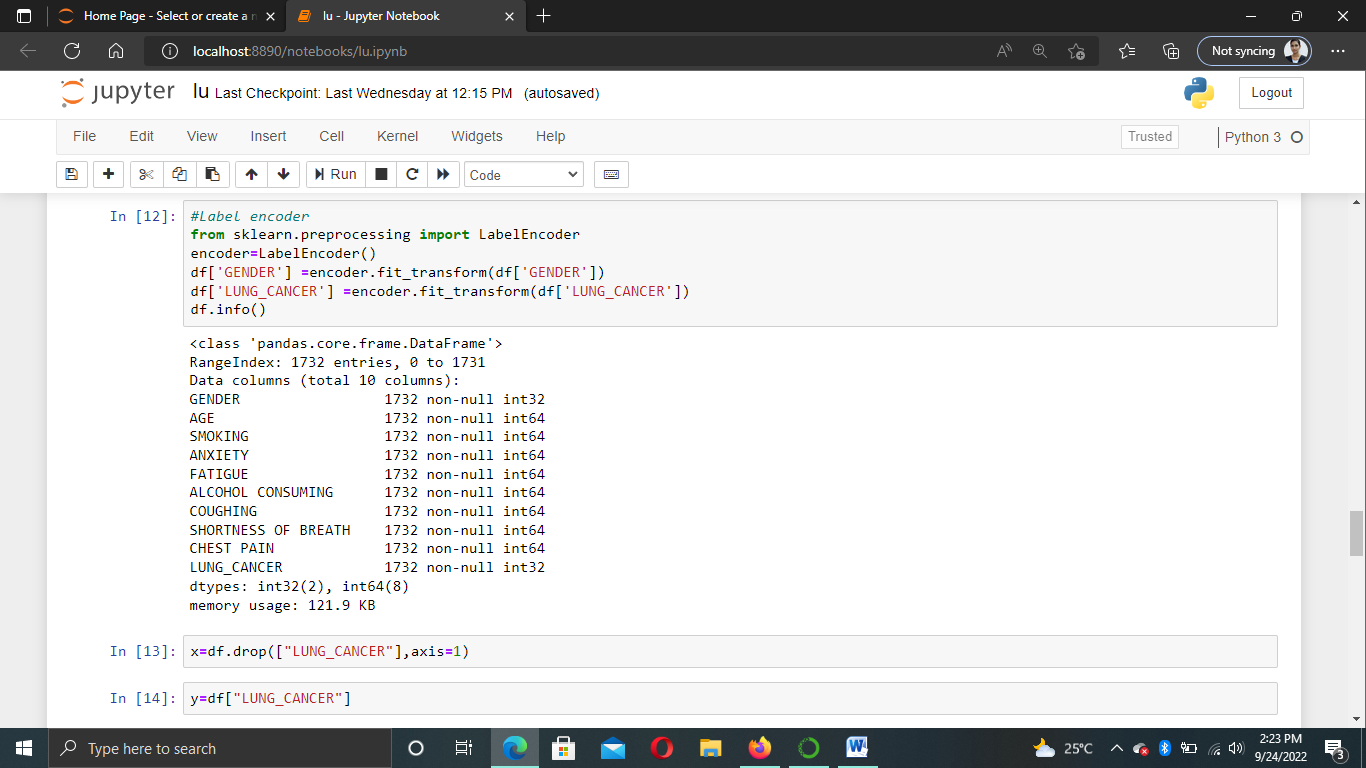


* We have two columns with string type which is in the type of object.
* We have to change string into float or integer values.
* Code for changing string columns into integer:

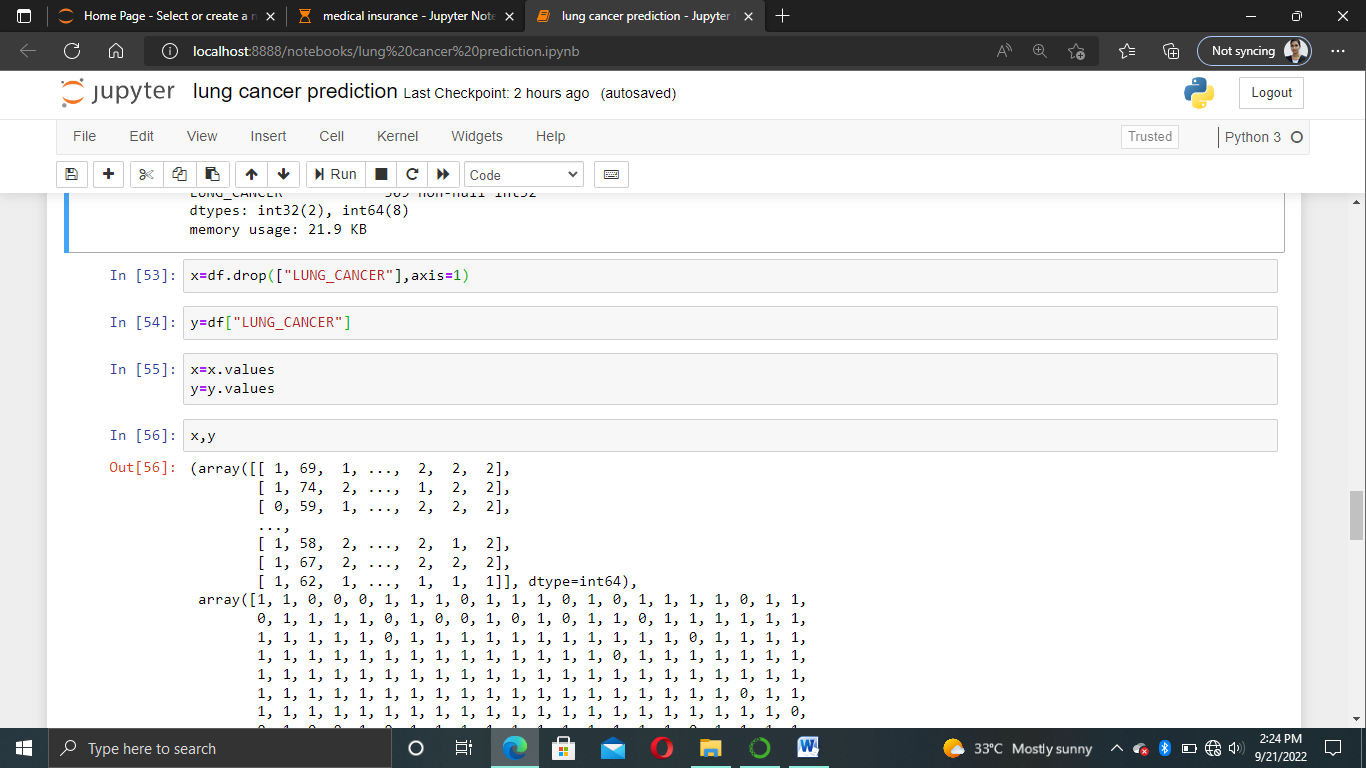
#LabelEncoder: Which is used for assigning numbers for the unique values

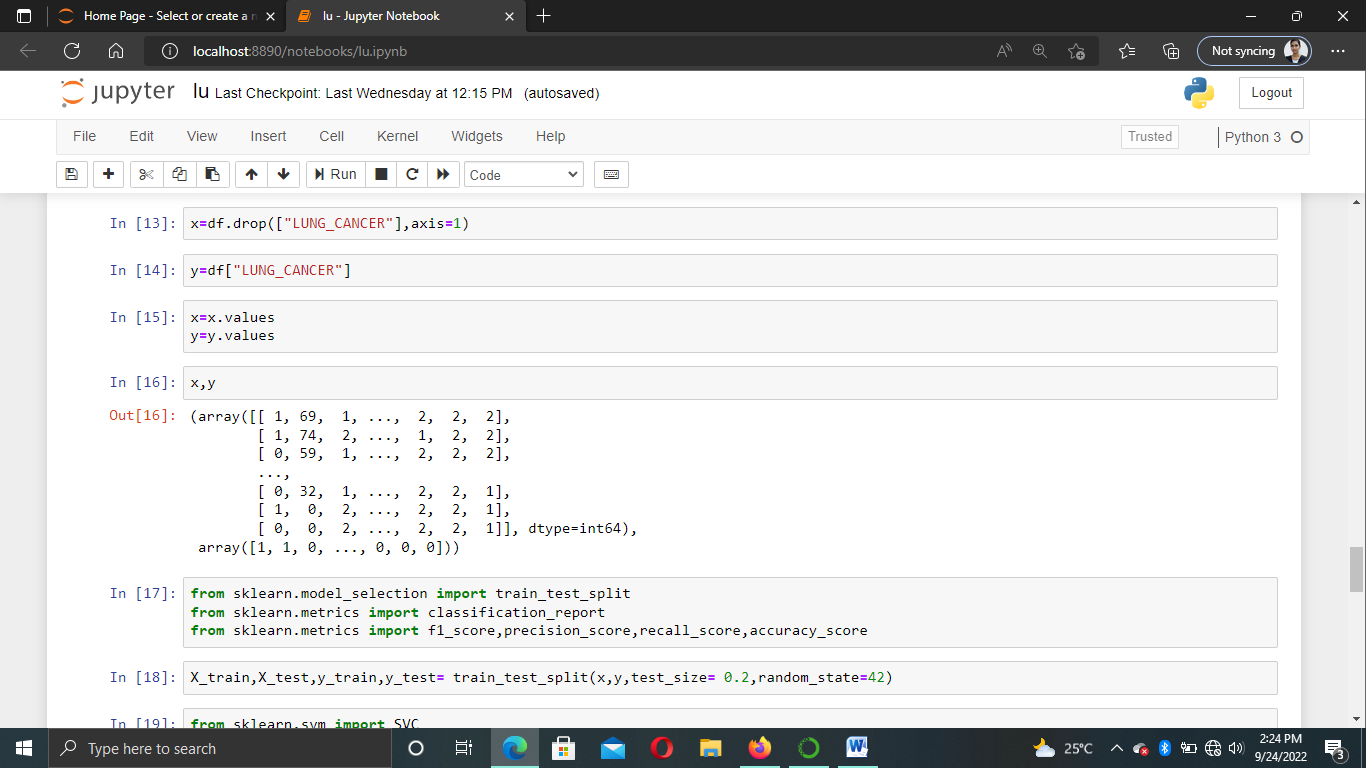


Output:



#Assigning X and Y Values:

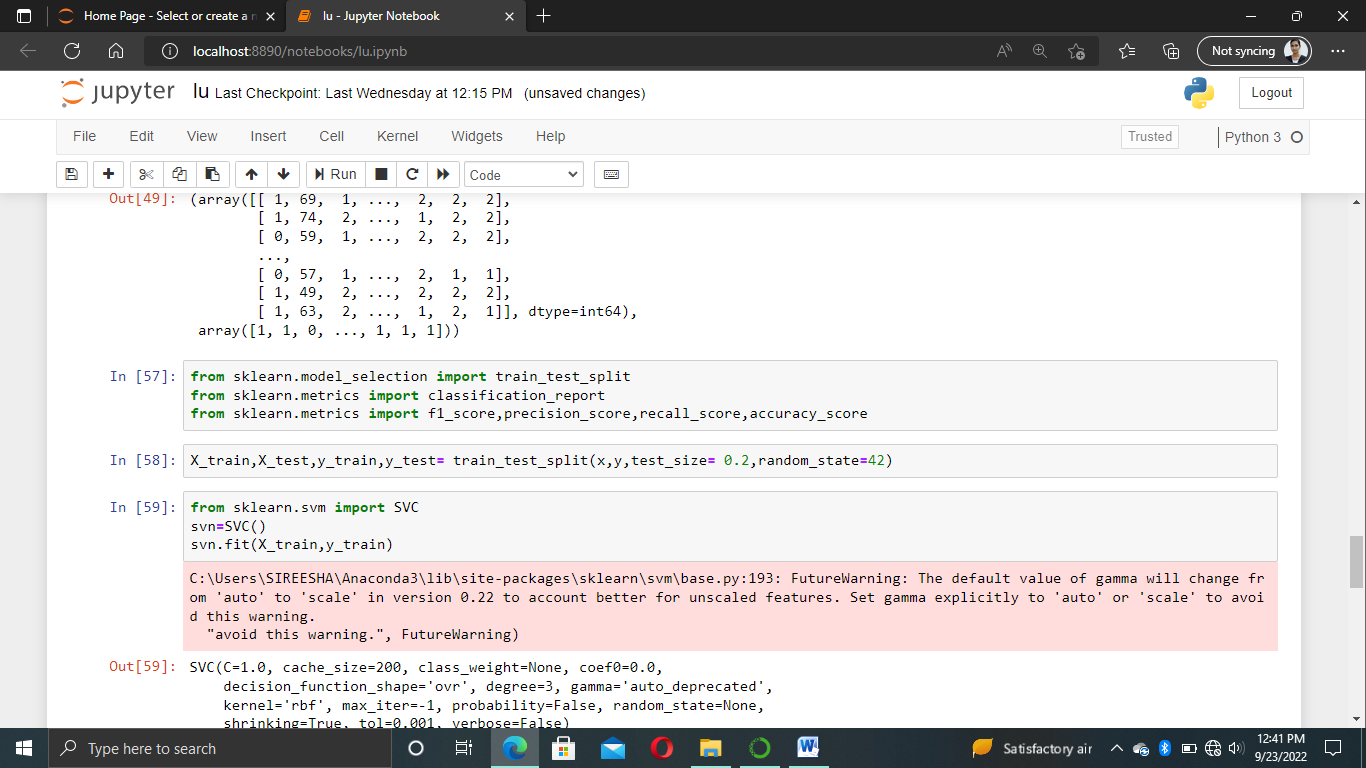


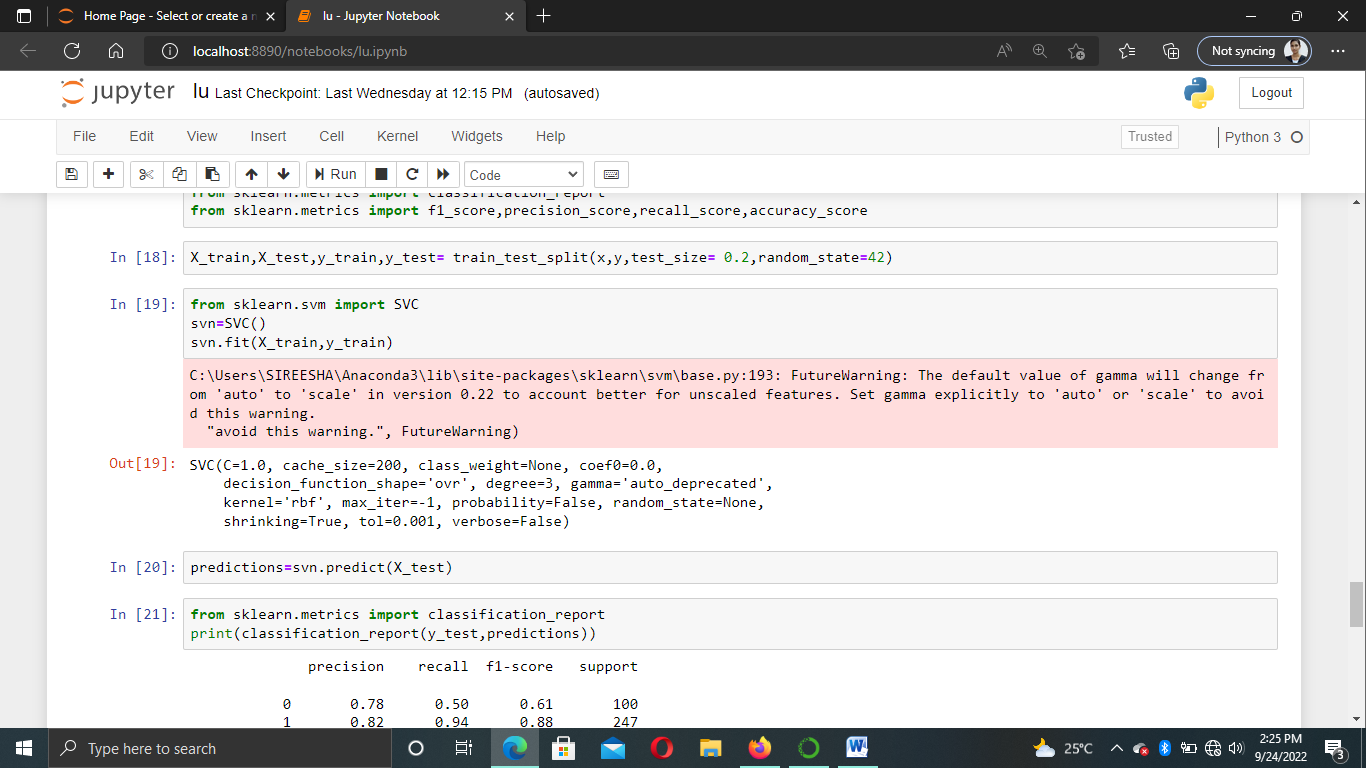


**Step 3 – Model training:**

* Using train\_test\_split we split the whole data into training and testing datasets. Later we’ll use the testing dataset to check the accuracy of the model.
* Here we imported a support vector classifier from the scikit-learn support vector machine.
* Then, we created an object and named it svn.
* After that, we feed the training dataset into the algorithm by using the svn.fit() method.

Splitting the data into training and testing using svm :





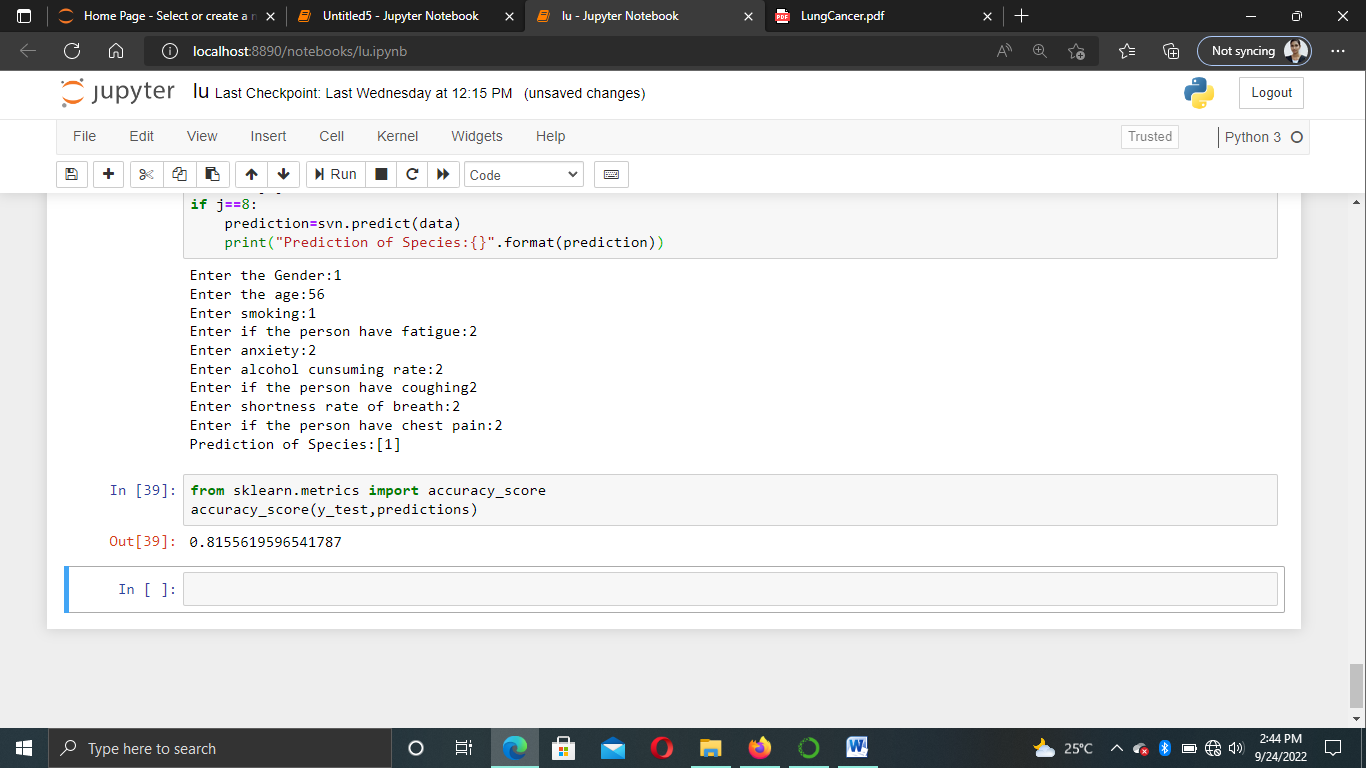
#### Step 4 – Model Evaluation:

* Now we predict the classes from the test dataset using our trained model.
* Then we check the accuracy score of the predicted classes.
* accuracy\_score() takes true values and predicted values and returns the percentage of accuracy.

# Predict from the test dataset

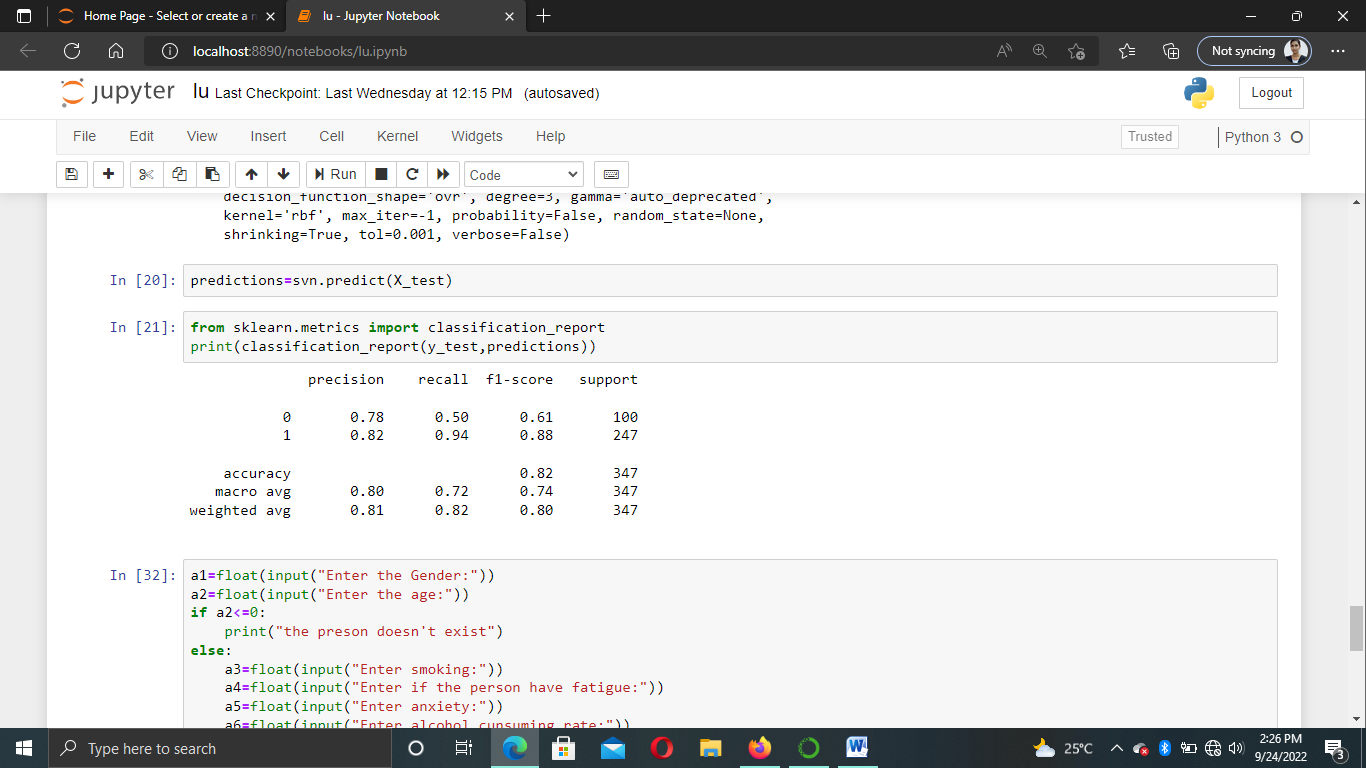
predictions = svn.predict(X\_test)

# Calculate the accuracy



The accuracy score is more than 0.80%

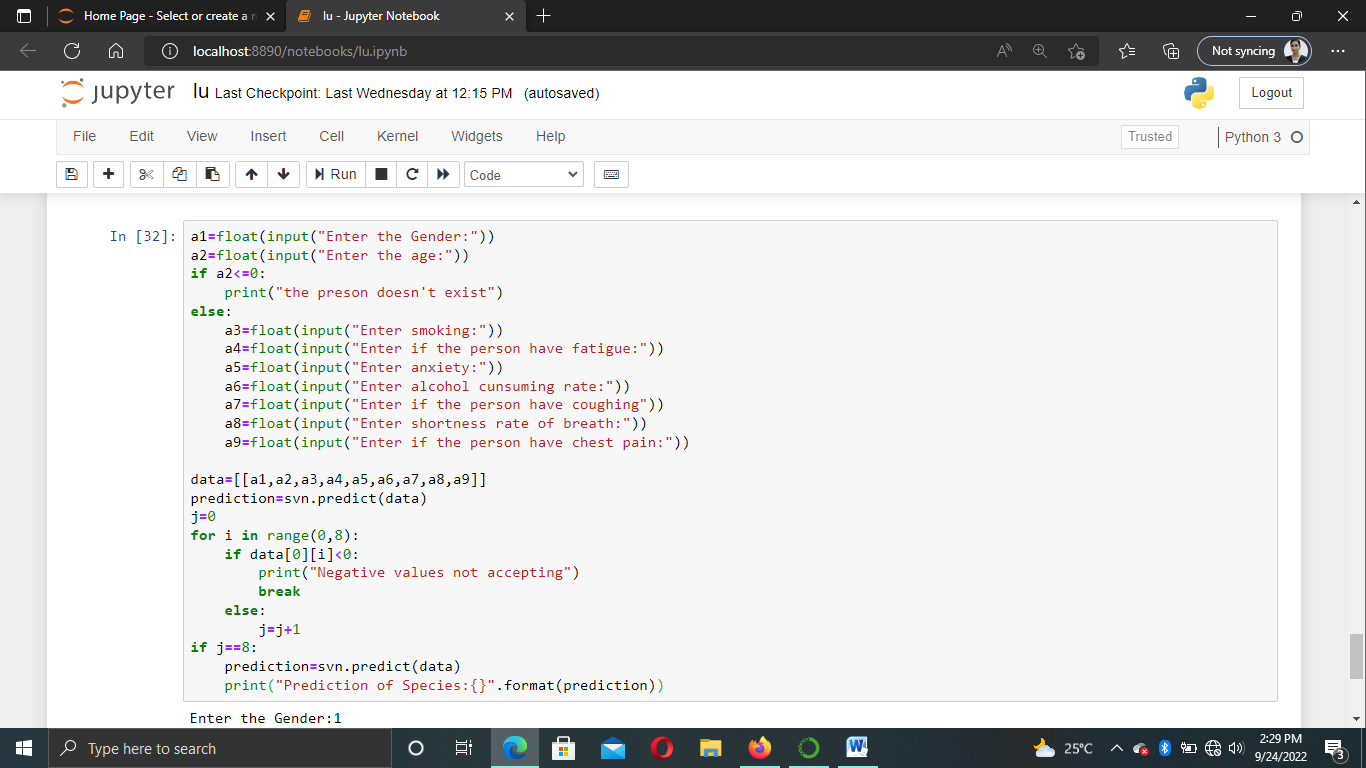
Classification report:



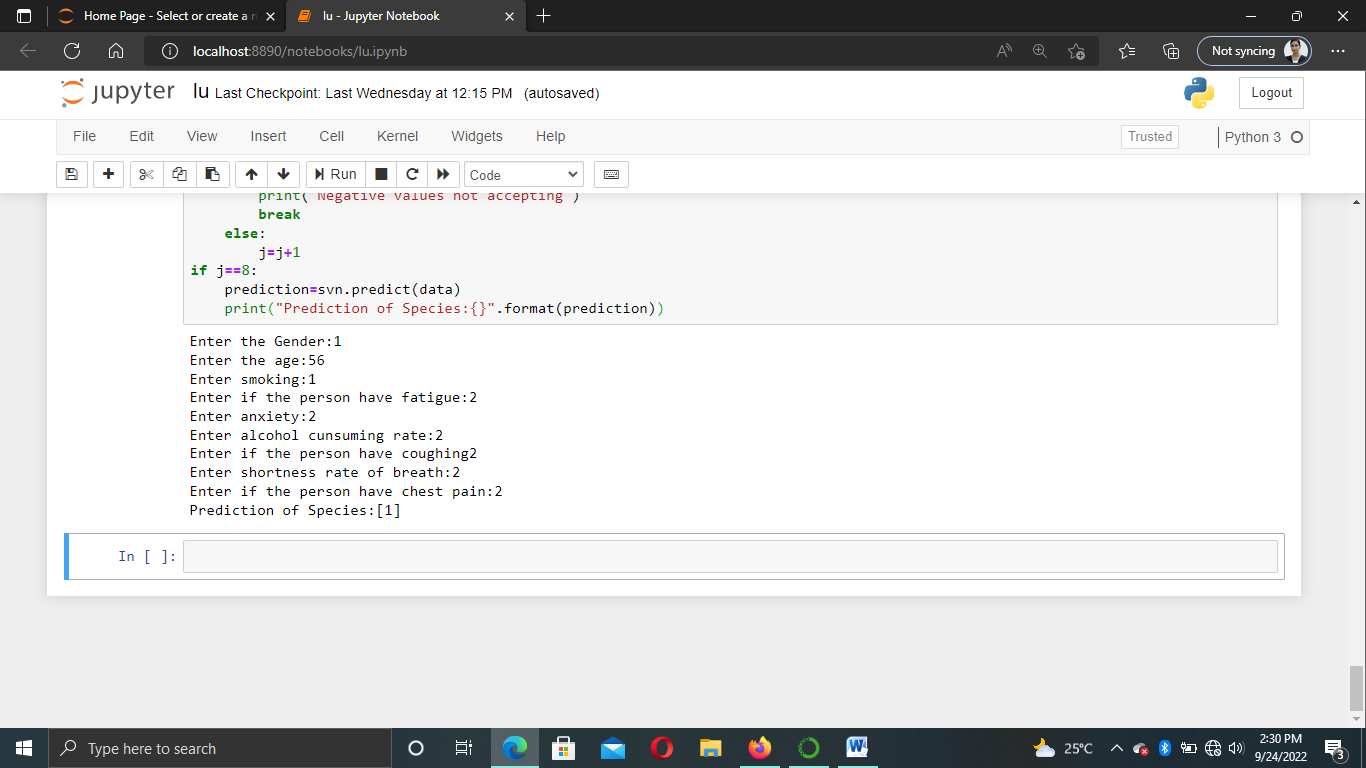
From the classification report we predict how accurate the dataset is

* Precision report tells about ratio of true positives to the sum of true and false positives.
* Recall tells about the ratio of the sum of the true positives and false negatives.
* F1\_score tells about the mean between precision and recall.
* If the f1\_score is 1.0 ,the expected performance of the model is better.
* Support tells about the number of actual occurences of the class in the dataset.

### **Step 5 – Testing the model:**



**Output:**



GIT HUB LINK: <https://github.com/sirirgukt/dsp-project.git>

SUMMARY AND ANALASYS IN MY POINT OF VIEW

I worked on the dataset and the model is successful. We got the accurate value from the model. From the output , It look like the model is predicting correctly.

Here we learned to train our own supervised machine learning model using Lung cancer Dataset, Classification Project with Machine Learning.

It helps us to learn machine learning, data visualization, model creation etc..

Based on our data set when we are visualizing the data with the help of seaborn we observed that we predicted the data how people will have lung cancer based on the some causes.

BY

B.SIREESHA

S180043

CSE-2C