# NAAN MUDHALVAN PROJECT

**Thyroid Disease Classification Using ML**

## BACHELOR OF SCIENCE IN COMPUTER SCIENCE

**TO THE**

THIRUVALLUVAR UNIVERSITY, SERKKADU, VELLLORE-632115

BY

## ARUNKUMAR A --- 35620U09003

## VISHNU K R --- 35620U09031

## MURALI S --- 35620U09013



**APRIL-2023**

## GOVERNMENT ARTS AND SCIENCE COLLEGE, ARAKKONAM-631051

**(AFFILIATED TO THIRUVALLUVAR UNIVERSITY)**

## GUIDED AND HEAD OF DEPARTMENT

**Dr. S. Selvakani**

|  |  |
| --- | --- |
| S.NO | CONTENT |
| 1. | INTRODUCTION |
| 2. | MILESTONE 1: Define Problem /Problem Understanding  \*Specify the business problem  \*Business requirements  \*Literature Survey  \*Social or Business Impact |
| 3. | MILESTONE 2: Data Collection & Preparation   * Download the dataset * Read The Dataset * Data Pre-processing * Checking for null values * Splitting the data x and y * Converting the Data Type * Handling Categorical Values * Splitting data into train and test * Handling Imbalanced Data * Applying StandardScaler * Performing Feature Importance * Selecting Output Columns |

|  |  |
| --- | --- |
| 4. | MILESTONE 3: Exploratory Data Analysis   * Descriptive analysis * Visual analysis * Checking Correlation |
| 5. | MILESTONE 4: Model Building  \*Training the model in  multiple algorithms  \*Random Forest Classifier Model  \*XGBClassifier model  \*SVC model  \* ANN Model  \*Testing the model |
| 6. | MILESTONE 5: Performance Testing & Hyperparameter Tuning  \*Testing model with multiple evaluation metrics  \*Compare the model  \*Comparing model accuracy before & after applying hyperparameter tuning |
| 7. | MILESTONE 6: Model Deployment   * Save the best model * Integrate with Web Framework * Build Python code: * Run the application |
| 8. | MILESTONE 7: Project Demonstration & Documentation  \*Record explanation video for the project end to solution  \*Project Documentation-Step by step project development |

INTRODUCTION

Campus recruitment is a strategy for sourcing, engaging and hiring young talent for internship and entry-level positions. College recruiting is typically a tactic for medium- to large-sized companies with high-volume recruiting needs, but can range from small efforts (like working with university career centers to source potential candidates) to large-scale operations (like visiting a wide array of colleges and attending recruiting events throughout the spring and fall semester).Campus recruitment often involves working with university career services centers and attending career fairs to meet in- person with college students and recent graduates. Our solution revolves around the placement season of a Business School in India. Where it has various factors on candidates getting hired such as work experience, exam percentage etc., Finally it contains the status of recruitment and remuneration details. We will be using algorithms such as KNN, SVM and ANN. We will train and test the data with these algorithms. From this the best model is selected and saved in . pkl format. We will be doing flask integration

# Milestone 1: Define Problem / Problem Understanding

## Activity 1: Specify the business problem

Refer to Project Description

## Activity 2: Business requirements

The business requirements for a machine learning model to predict thyroid disease include the ability to accurately predict thyroid disease based on the scan results, Minimise the number of false positives (wrong thyroid disease

confirmations) and false negatives (thyroid is there but got as not thyroid disease). Provide an explanation for the model's decision, to comply with regulations and improve transparency.

## Activity 3: Literature Survey (Student Will Write)

The thyroid gland is one of the body‟s most visible endocrine glands. Its size is determined by the individual‟s age, gender, and physiological states, such as pregnancy or lactation. It is divided into two lobes (right and left) by an isthmus (a band of tissue). It is imperceptible in everyday life yet can be detected when swallowing. The thyroid hormones T4 and T3 are needed for normal thyroid function. These hormones have a direct effect on the body‟s metabolic rate. It

contributes to the stimulation of glucose, fatty acid, and other molecule consumption. Additionally, it enhances oxygen consumption in

the majority of the body‟s cells by assisting in the processing of uncoupling proteins, which

contributes to an improvement in the rate of cellular respiration.

Thyroid

conditions are difficult to detect in test results, and only trained professionals can do so. However, reading such extensive reports and predicting future results is difficult. Assume a machine learning model can detect the thyroid disease in a patient. The thyroid disease can then be easily identified based on the symptoms in the patient‟s history. Currently, models are evaluated using accuracy metrics ona validation dataset that is accessible.

## Activity 4: Social or Business Impact.

Social Impact:- Untreated/undetected thyroid disease is more dangerous at times it can lead to fatal of the person. So, we can detect it at the earliest then people can get treatment and get cured.

Business Model/Impact:- We can make this application public, offer services as asubscription based or can collaborate with healthcare centres or specialists.

# Milestone 2: Data Collection & Preparation

ML depends heavily on data. It is the most crucial aspect that makes algorithm training possible. So this section allows you to download the required dataset.

## Activity 1: Download the dataset

There are many popular open sources for collecting the data. Eg: kaggle.com, UCI repository, etc.

In this project, we have used drug200.csv data. This data is downloaded fromkaggle.com. Please refer to the link given below to download the dataset.

Link: <https://www.kaggle.com/prathamtripathi/drug-classification>

**Activity 1.1: Importing the libraries**

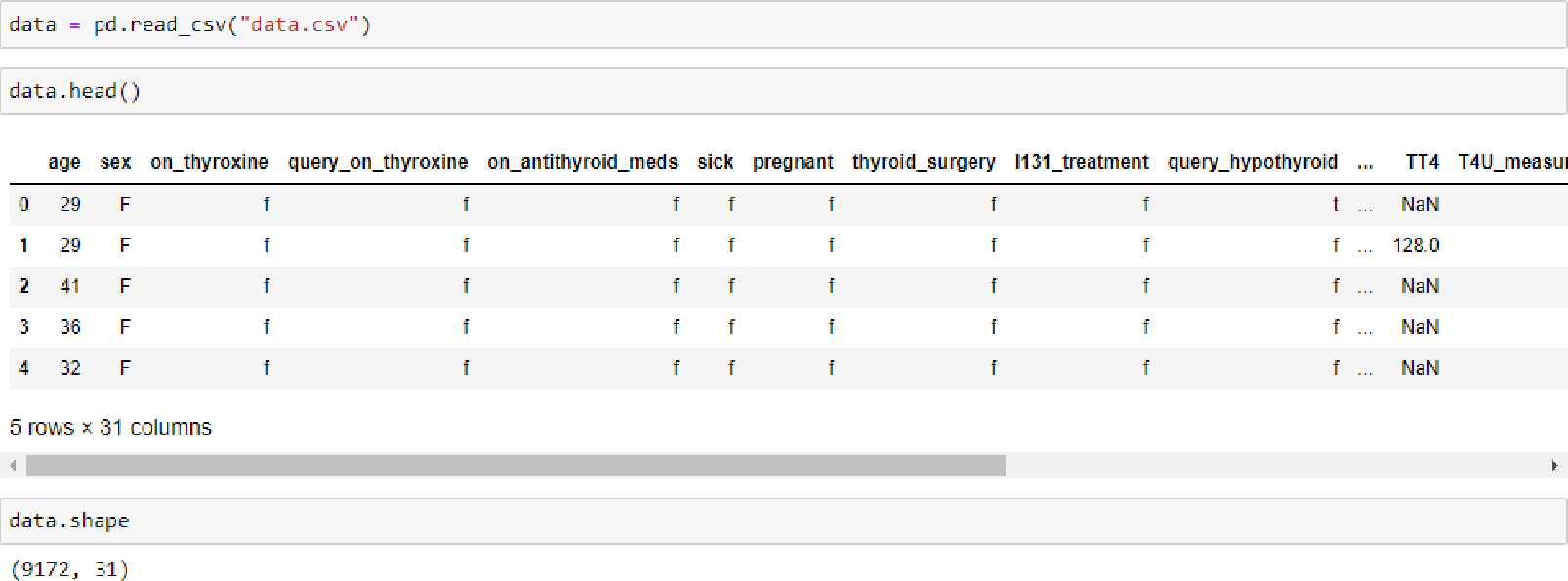
Import the necessary libraries as shown in the image.



**Activity 1.2: Read the Dataset**

Our dataset format might be in .csv, excel files, .txt, .json, etc. We can read thedataset with the help of pandas.

In pandas, we have a function called read\_csv() to read the dataset. As a parameter, we have to give the directory of the csv file.



## Activity 2: Data Pre-processing

As we have understood how the data is, let's pre-process the collected

data.

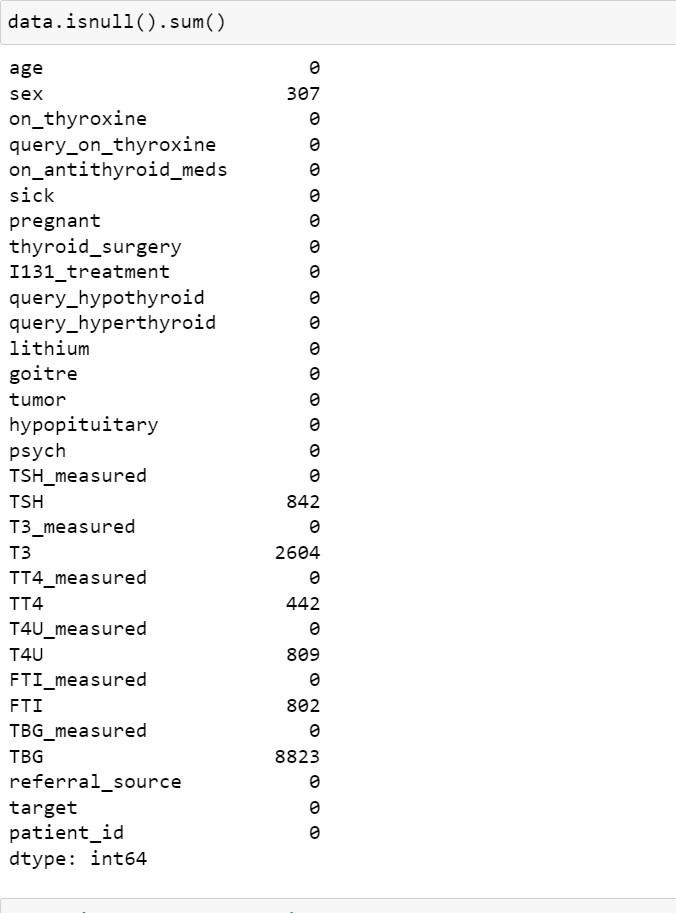
The download data set is not suitable for training the machine learning model as it might have so much randomness so we need to clean the dataset properly in order to fetch good results. This activity includes the following steps.

* Handling missing values
* Descriptive analysis
* Splitting the dataset as x and y
* Handling Categorical Values
* Checking Correlation
* Converting Data Type
* Splitting dataset into training and test set
* Handled Imbalanced Data
* Applying StandardScaler

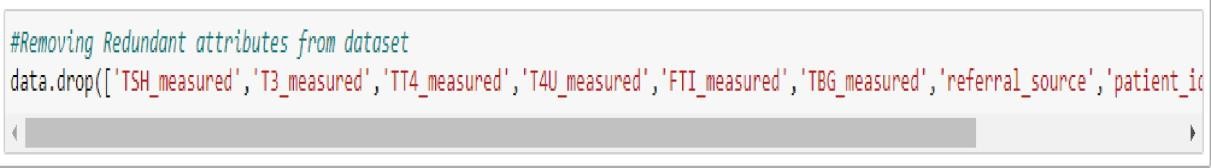
Note: These are the general steps of pre-processing the data before using it for machine learning. Depending on the condition of your dataset, you may or may not have to go through all these steps.

**Activity 2.1: Checking for null values**

* For checking the null values, data.isnull() function is used. To sum thosenull values we use the .sum() function to it. From the below image we found that there are no null values present in our dataset. So we can skiphandling the missing values step.

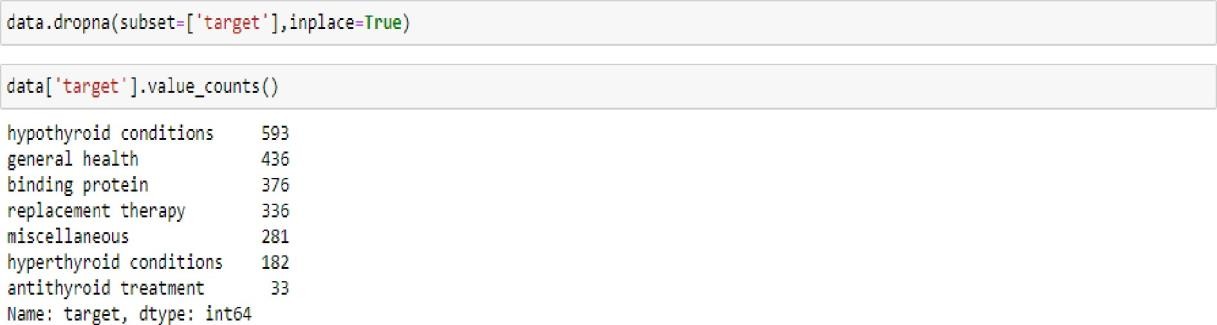


* Removing the Redundant attributes from the dataset.

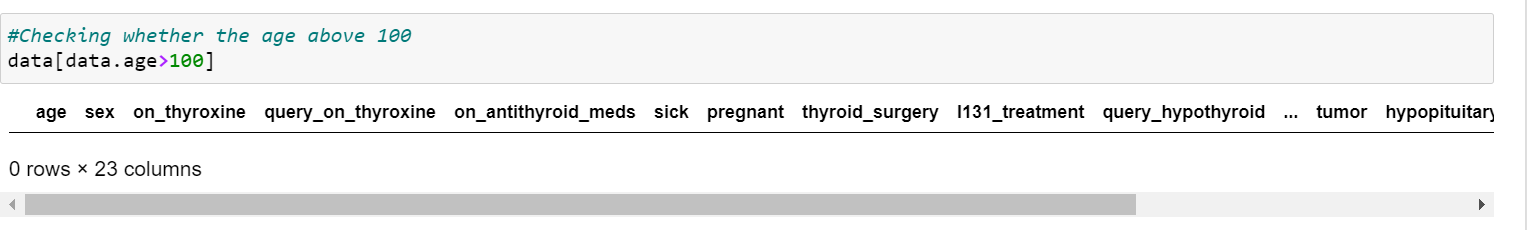
• 

* Re-mapping the 'target' values to the diagnostic Group



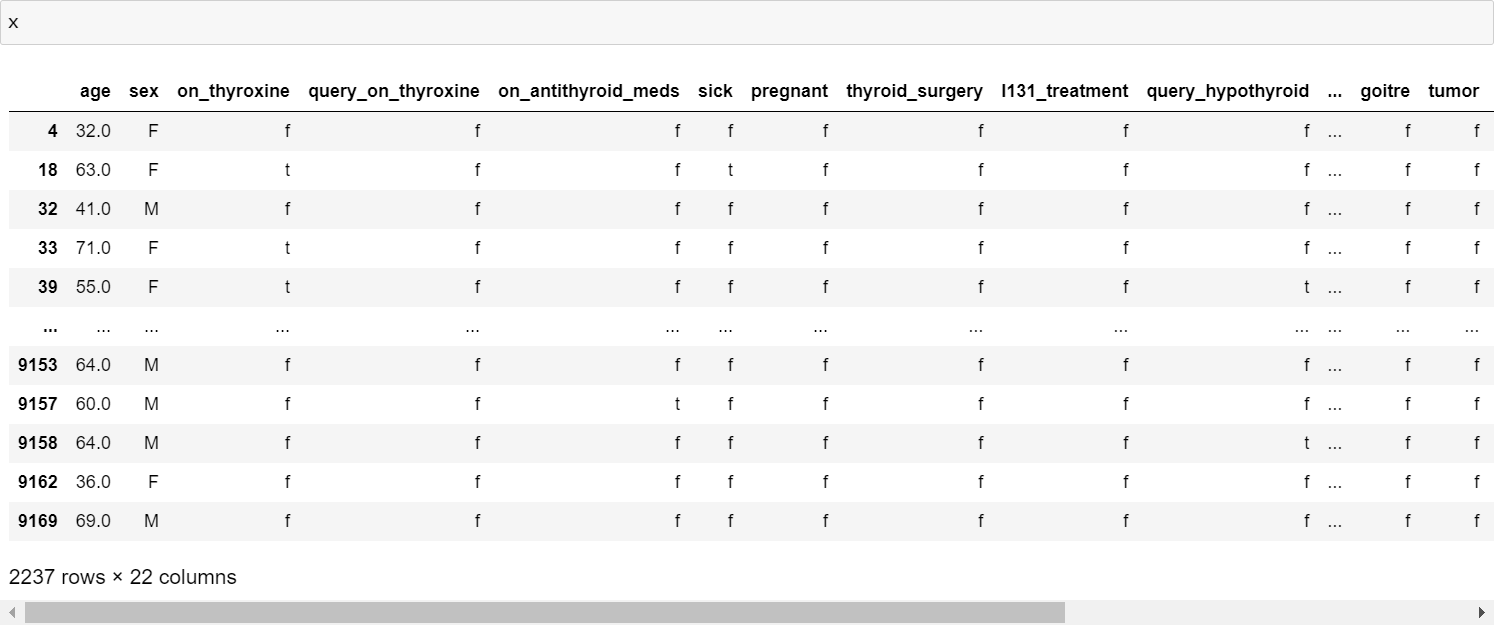
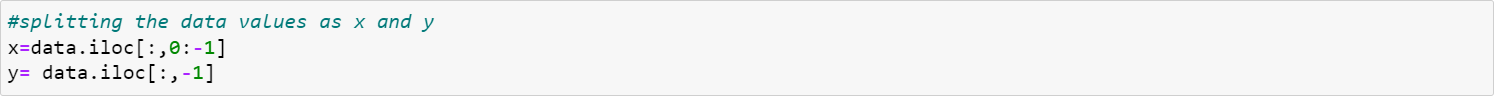
Dropping Null Values

Checking the 'age' is there any above 100 and we drop the age>100.



**Activity 2.2: Splitting the data x and y**

Splitting the data x and y



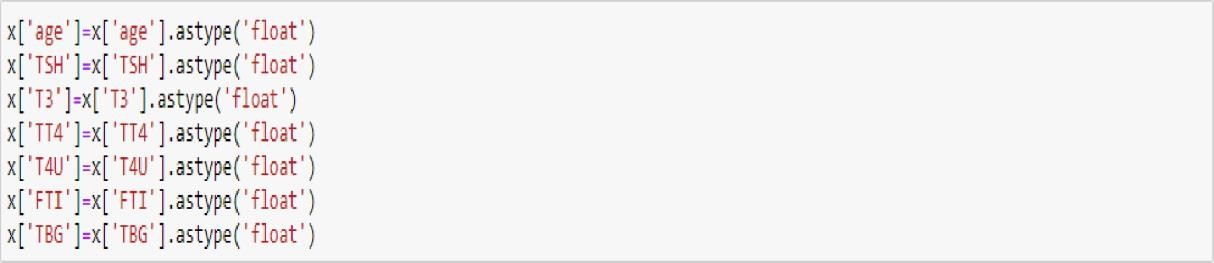
* Making 'F' on wherever we have the 'nan' values on data.



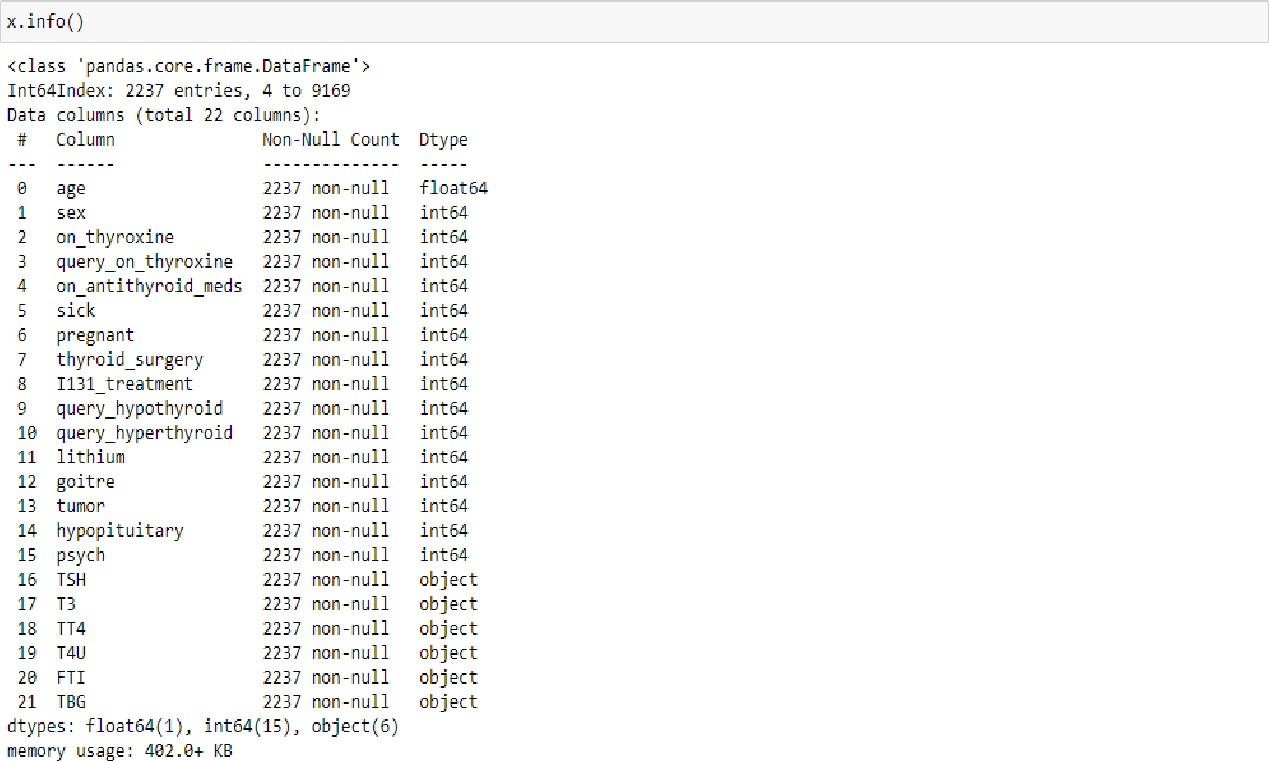
**Activity 2.3: Converting the Data Type**

Converting the data type from object to float. So that we will get output properly. And Checking info about data.

* Here, we have the object values are 'TSH', 'T3', 'TT4', 'T4U', 'FTI', 'TBG'and convert them to float values.



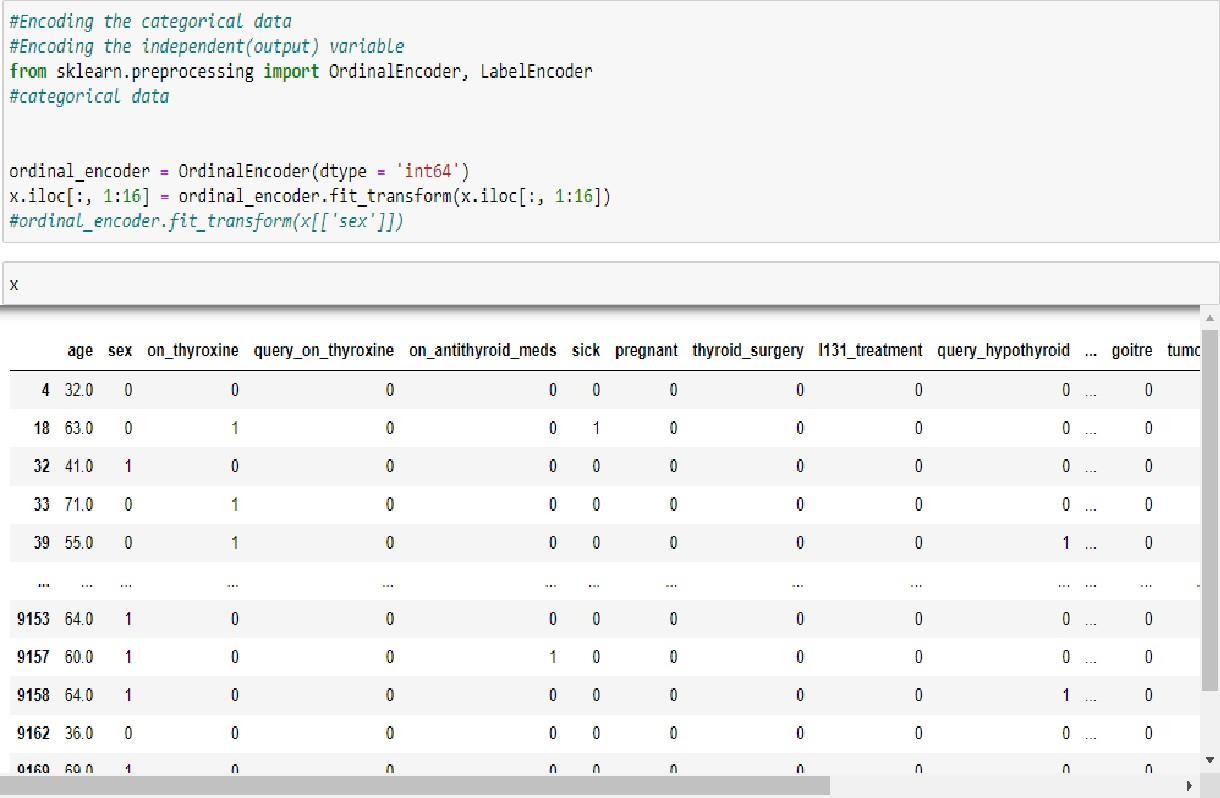
* Then we can check the datatype information about the dataset by code ofx.info()

**Activity 2.4: Handling Categorical Values**

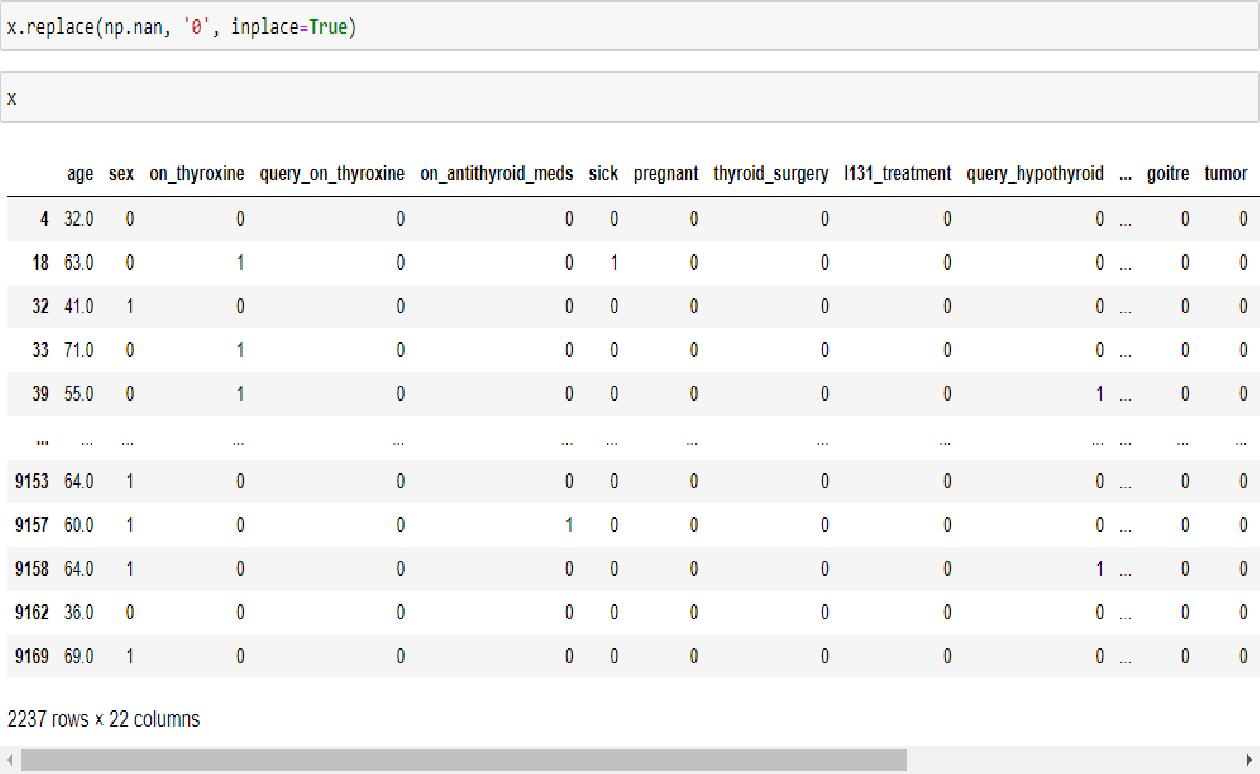
As we can see our dataset has categorical data we must convert the categorical data to integer encoding or binary encoding.

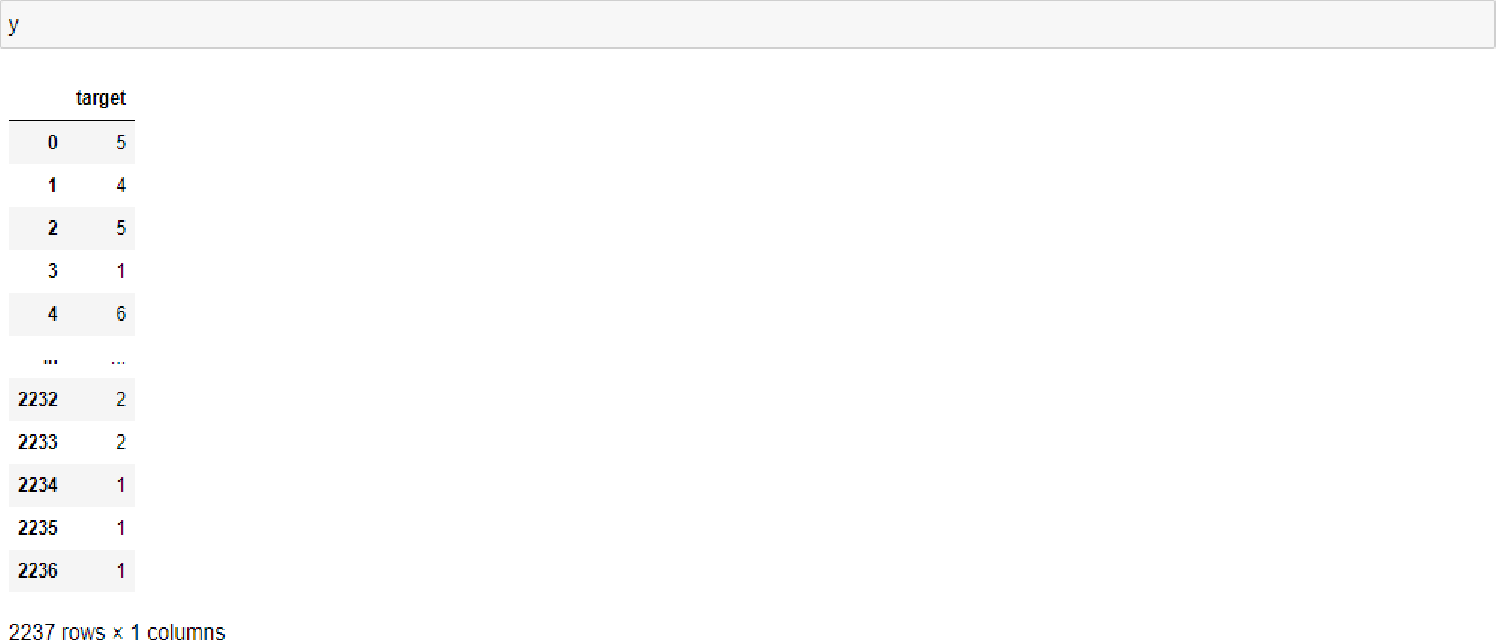
To convert the categorical features into numerical features we use encoding techniques. There are several techniques but in our project we are using Ordinal Encoding and Label Encoding.

* In our project, categorical features are x and y values.
* Here, applying Ordinal Encoding on x values.



* Replacing the nan values with zero (0) values.



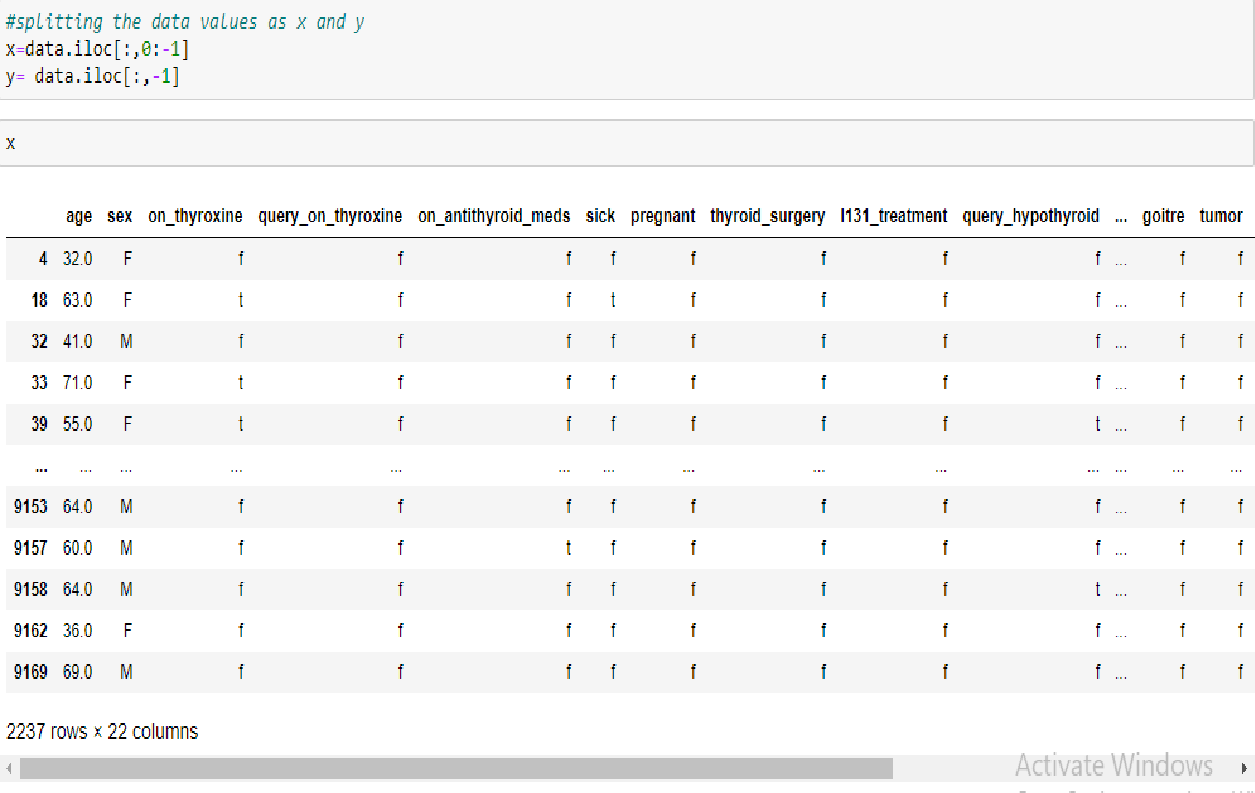
* Now, applying Label Encoding on y(Independent variable) value.

**Activity 2.5: Splitting data into train and test**

Now let‟s split the Dataset into train and test sets

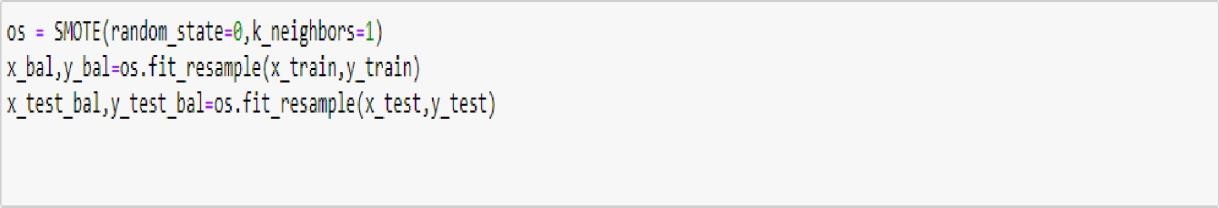
Changes: first split the dataset into x and y and then split the data set Here x and y variables are created. On x variable, data is passed with dropping the target variable. And my target variable is

passed. For splitting training and testing data we are using the

train\_test\_split() function from sklearn. As parameters, we are passing x, y, test\_size, random\_state.

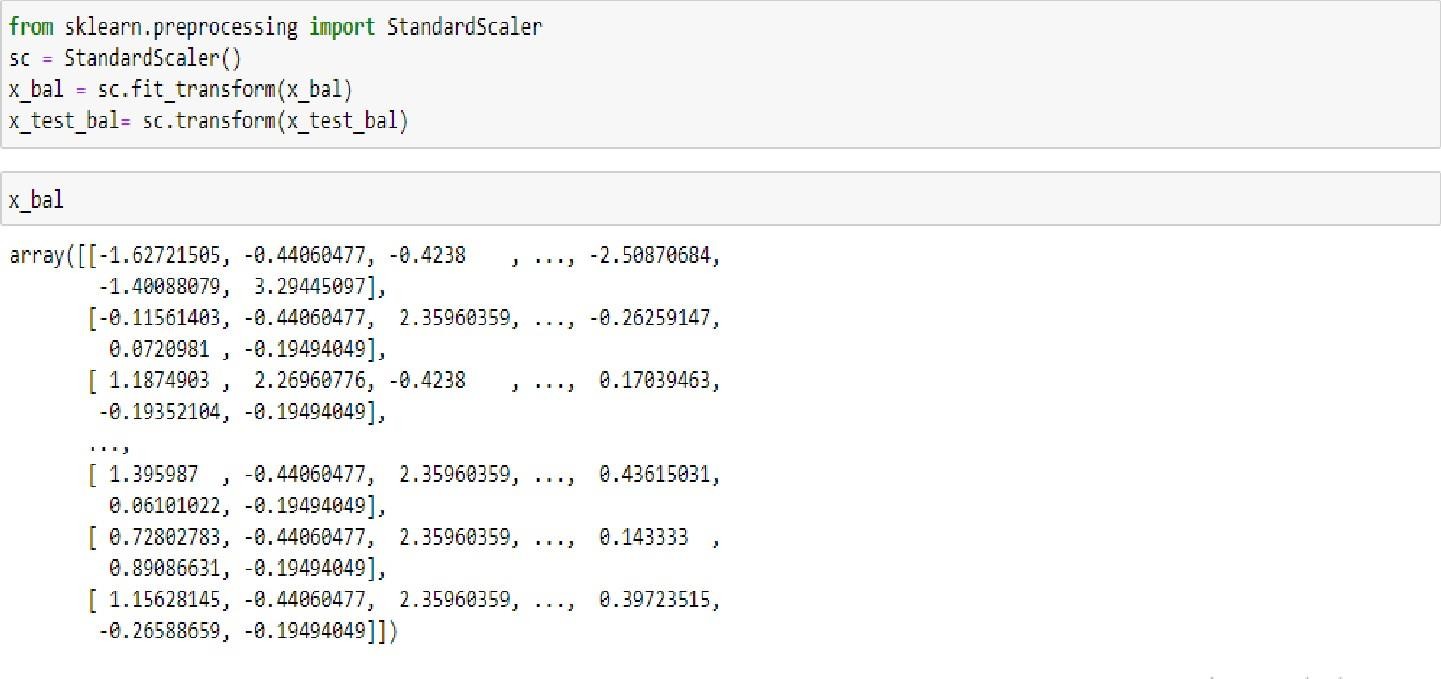


**Activity 2.6: Handling Imbalanced Data**

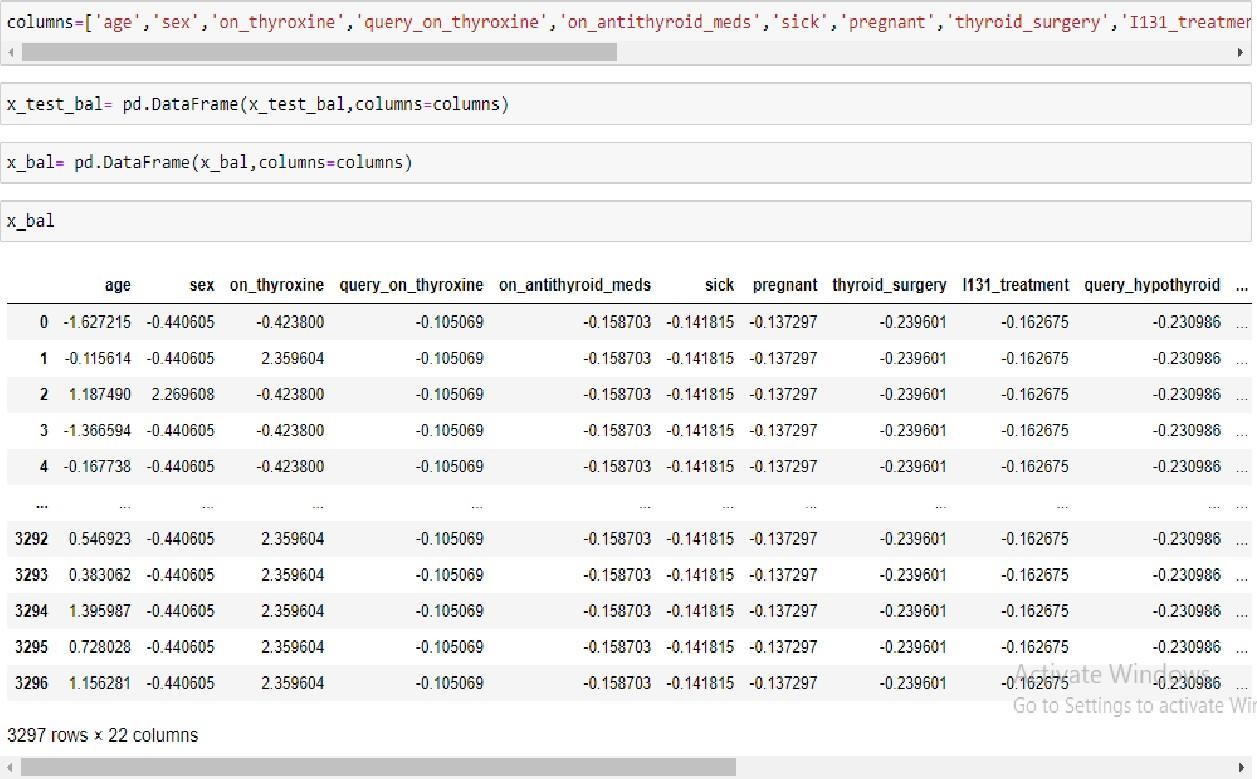


**Activity 2.7: Applying StandardScaler**

* Scaling the features makes the flow of gradient descent smooth and helpsalgorithms quickly reach the minima of the cost function.
* Without scaling features, the algorithm may be biased toward the feature which has values higher in magnitude. it brings every feature in the samerange and the model uses every feature wisely.

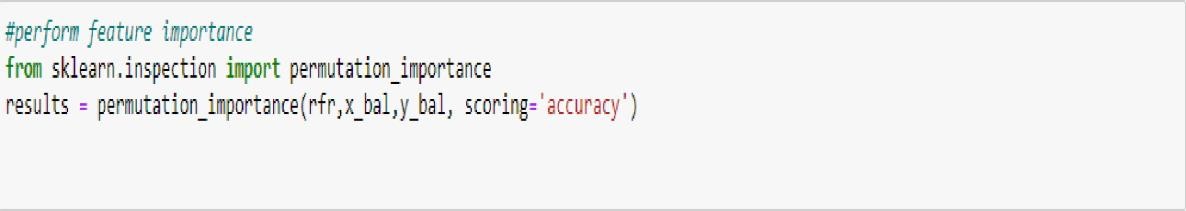


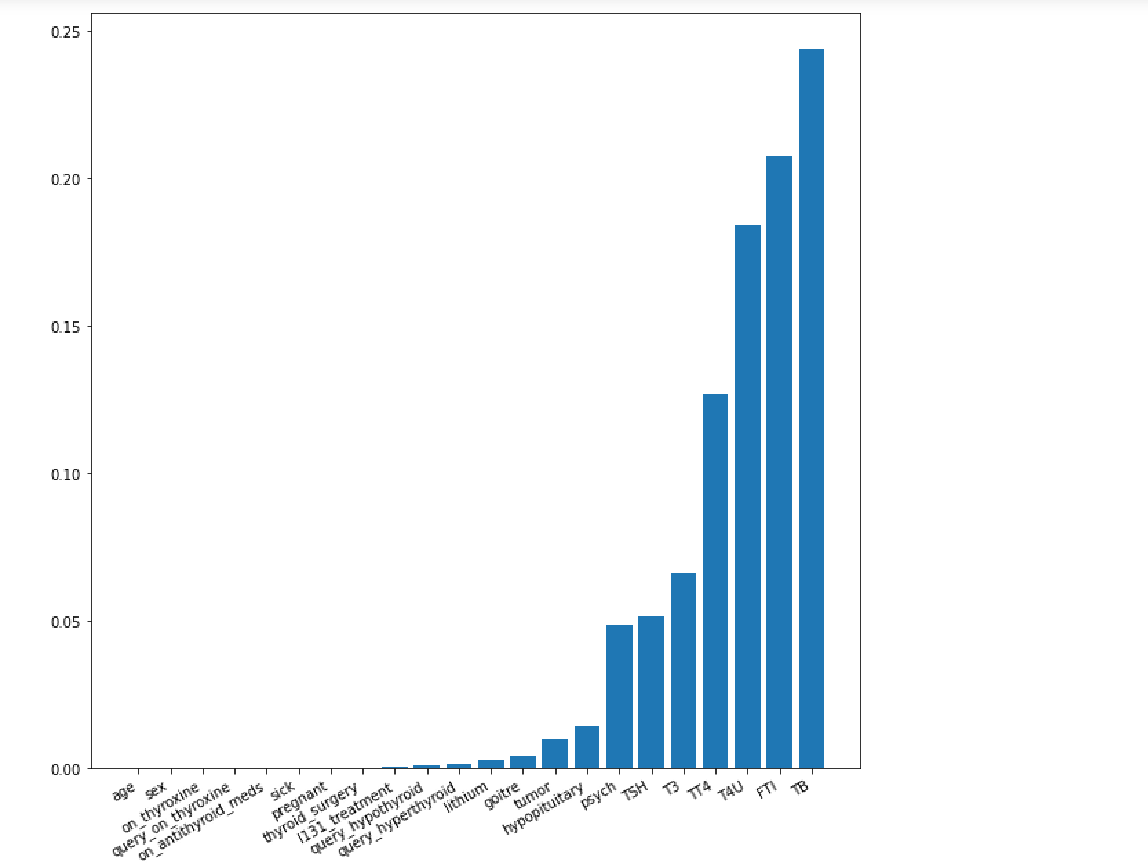
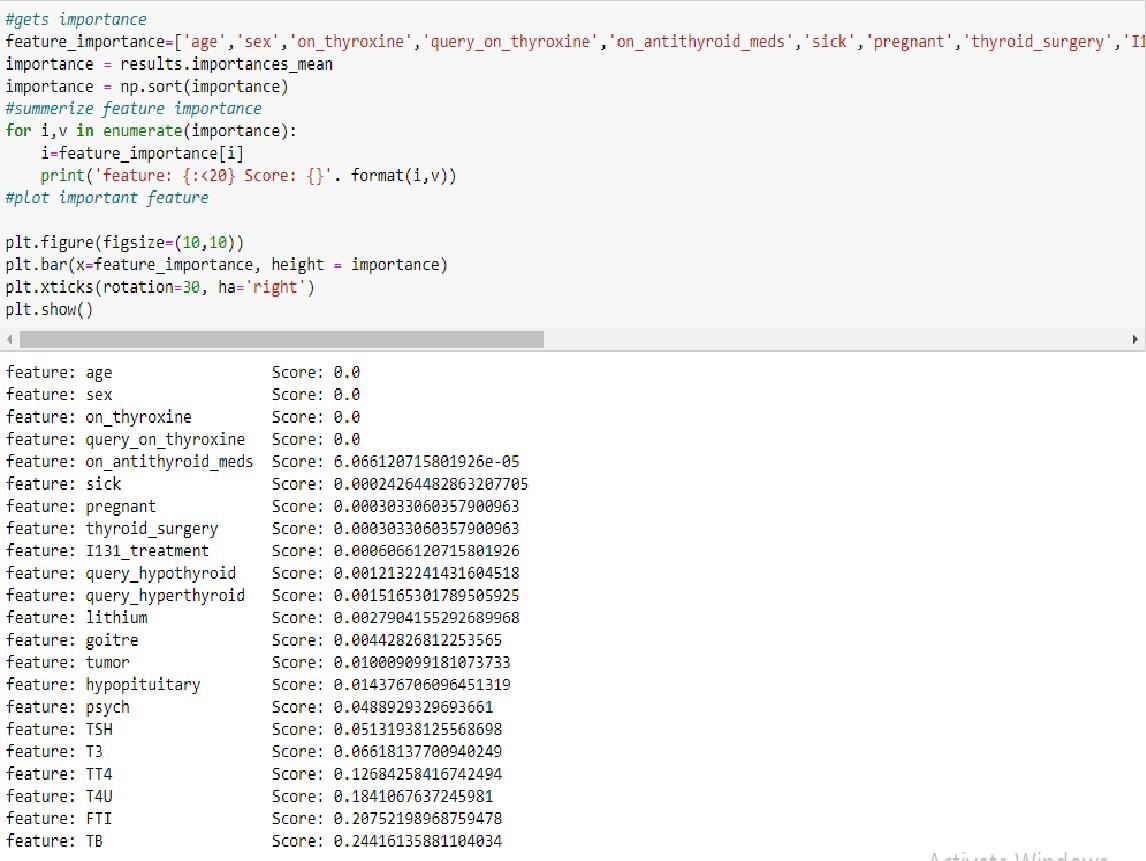
* Here, we have the data in array format and we are making itdataframe(table format).



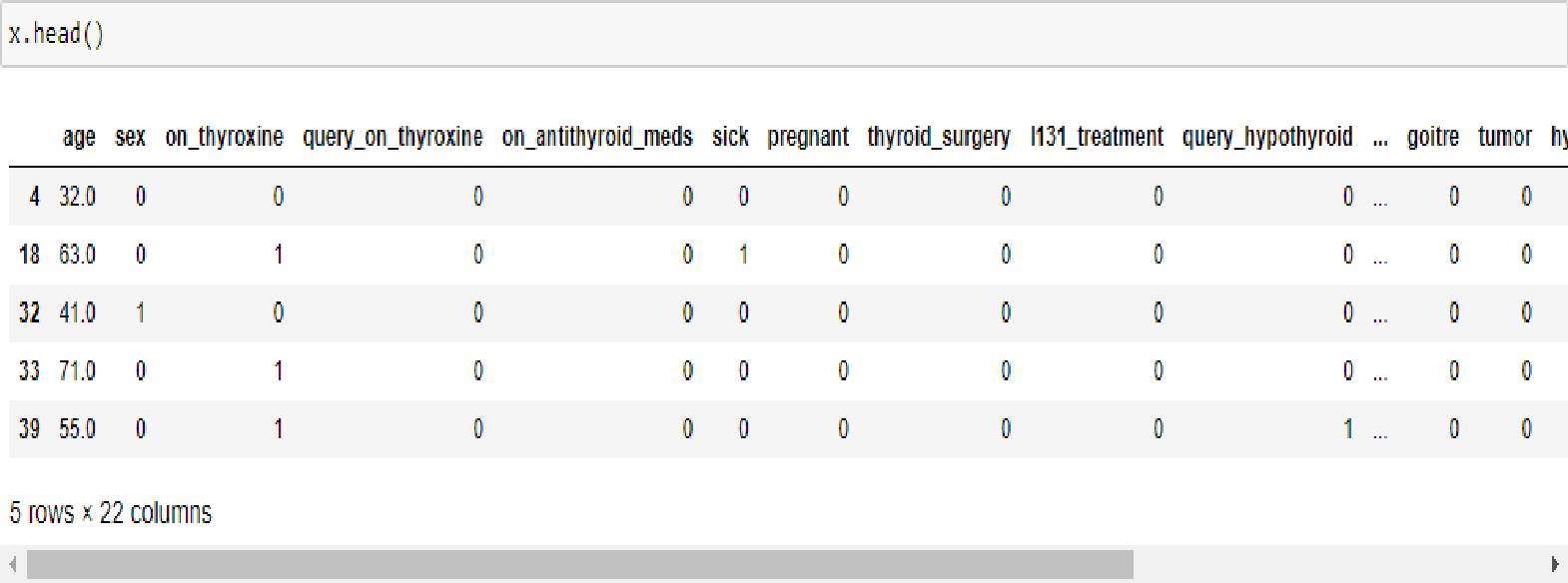
**Activity 2.8: Performing Feature Importance**

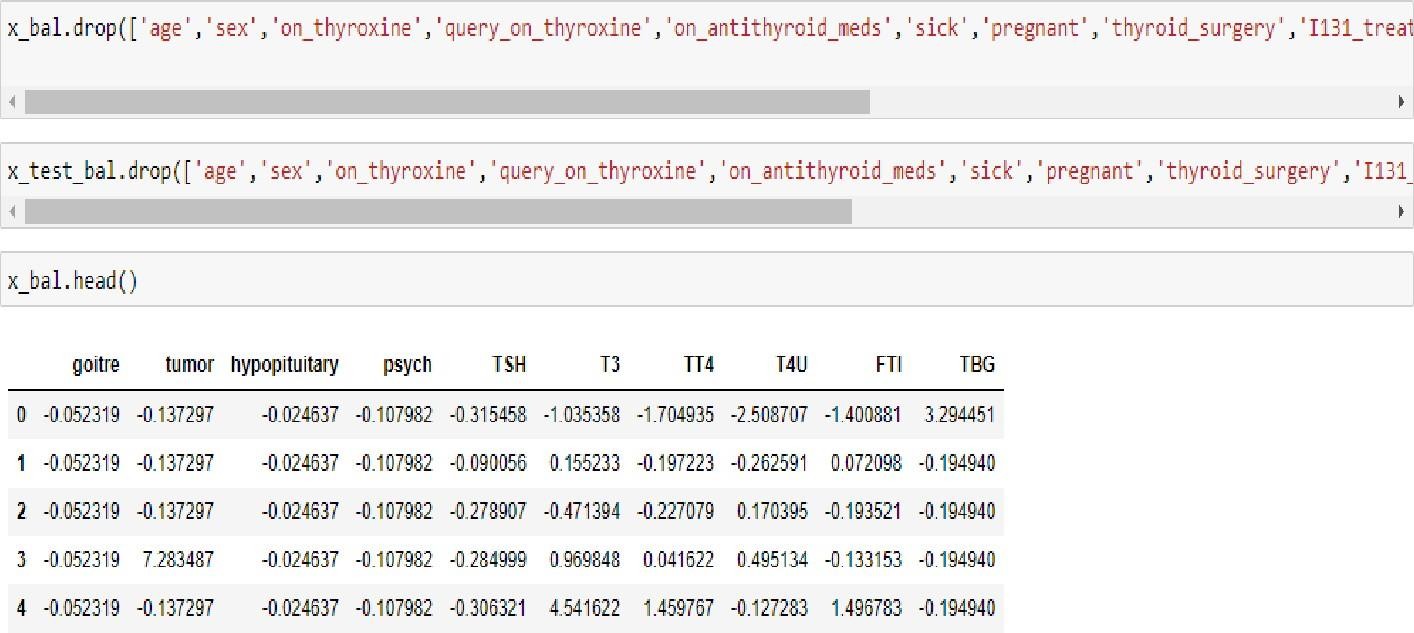
* The idea behind permutation feature importance is simple. The feature importance is calculated by noticing the increase or decrease in error when wepermute the values of a feature.
* If permuting the values causes a huge change in the error, it means the feature isimportant for our model.





**Activity 2.9: Selecting Output Columns**

* Before we have this many columns
* After Performing Feature Importance by using 'Permutation Importance' we aredropping some columns which are not important for 'target'.

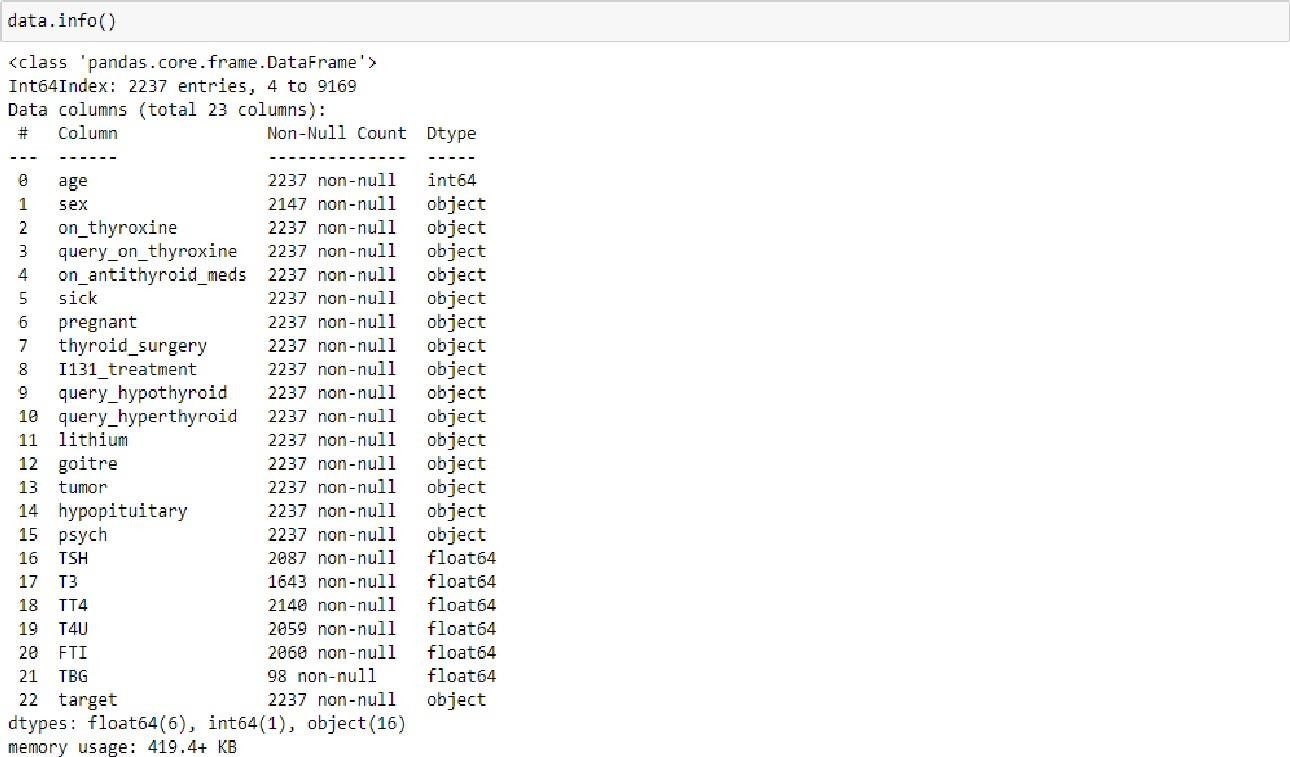


# Milestone 3: Exploratory Data Analysis

## Activity 1: Descriptive analysis

Descriptive analysis is to study the basic features of data with the statistical process. Here pandas have a worthy function called describe. With this described function we can find mean, std, min, max and percentile values of continuous features.

Checking info about data by using data\_info()



## Activity 2: Visual analysis

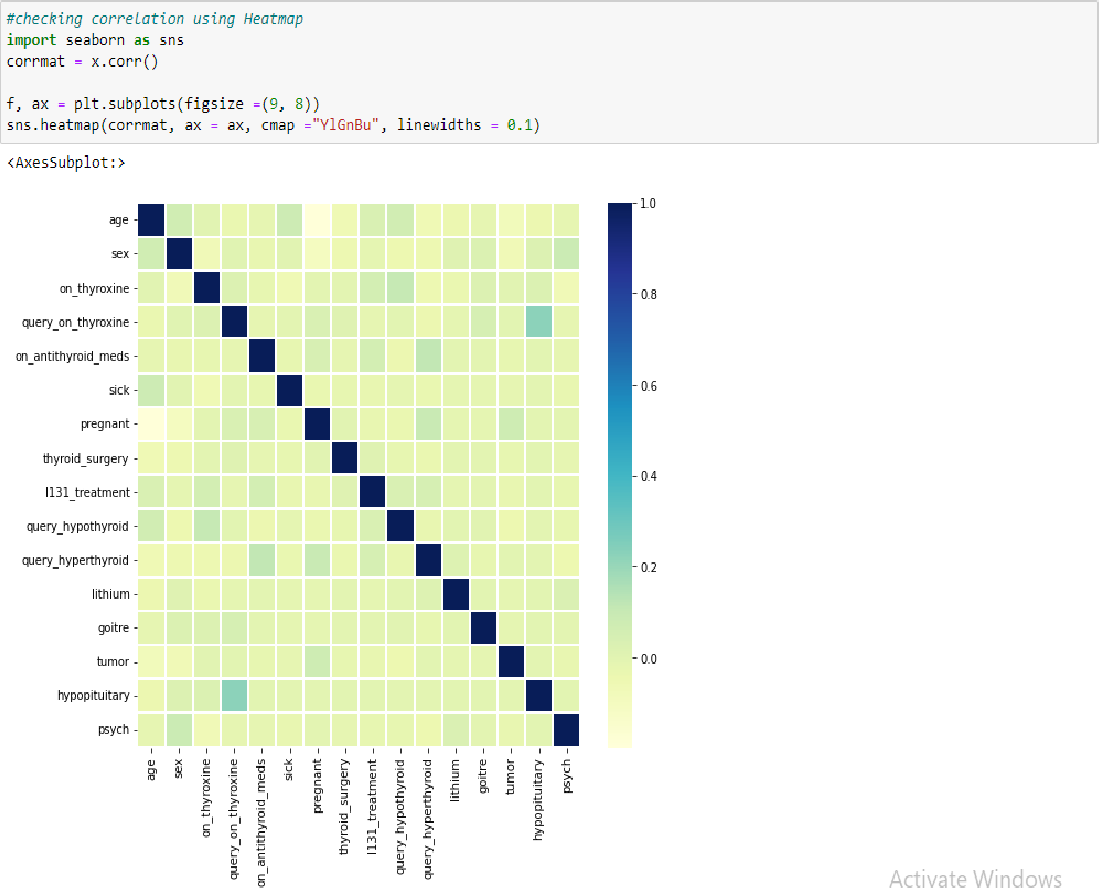
Visual analysis is the process of using visual representations, such as charts, plots, and graphs, to explore and understand data. It is a way to quickly identify patterns, trends, and outliers in the data, which can help to gain insights and make informed decisions.

**Activity 2.1: Checking Correlation.**

Here, I'm finding the correlation using HeatMap. It visualizes the data in 2-D

coloured maps making use of colour variations. It describes the related variablesin the form of colours instead of numbers; it will be plotted on both axes.

Here, there is no correlation between columns.



# Milestone 4: Model Building

## Activity 1: Training the model in multiple algorithms

Now our data is cleaned and it‟s time to build the model. We can train our data ondifferent algorithms. For this project we are applying four classification

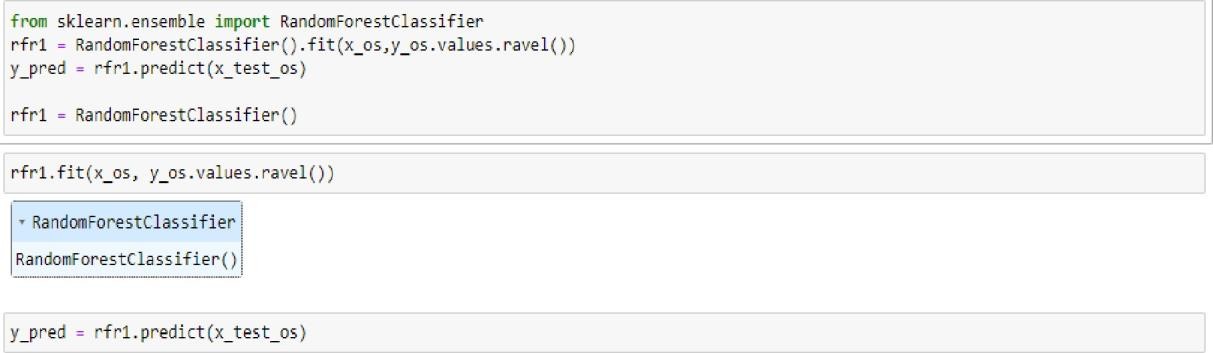
algorithms. The best model is saved based on its performance.

**Activity 1.1: Random Forest Classifier Model**

A function named Random Forest Classifier Model is created and train and test data are passed as the parameters. Inside the function, the Random Forest

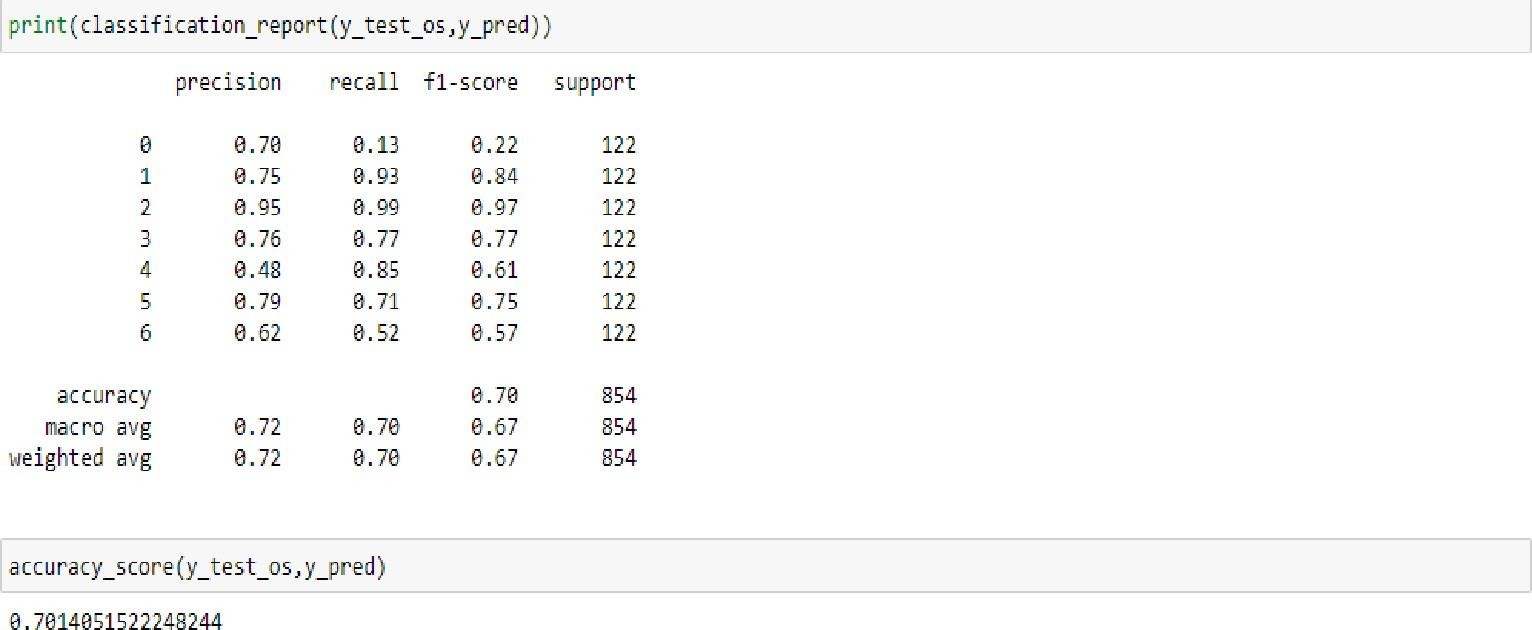
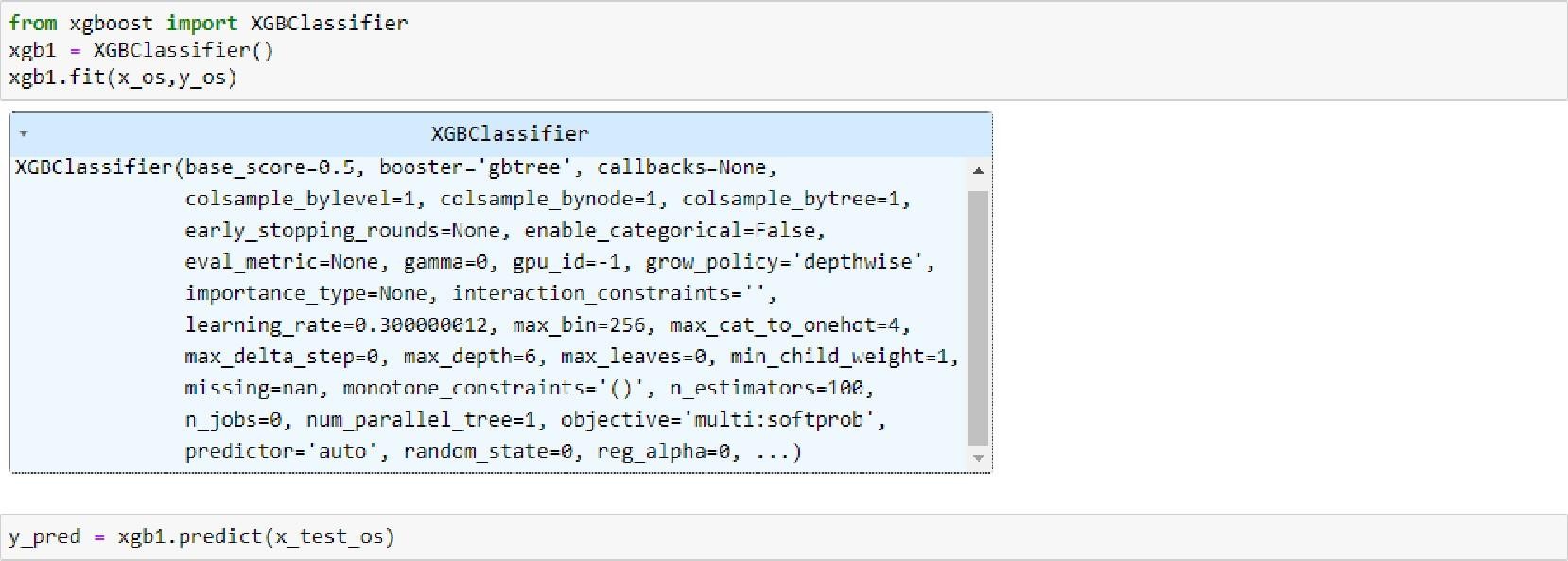
Classifier algorithm is initialized and training data is passed to the model with the

.fit() function. Test data is predicted with the .predict() function and saved in a new variable. For evaluating the model, accuracy\_score and classification report is done.



**Activity 1.2: XGBClassifier model**

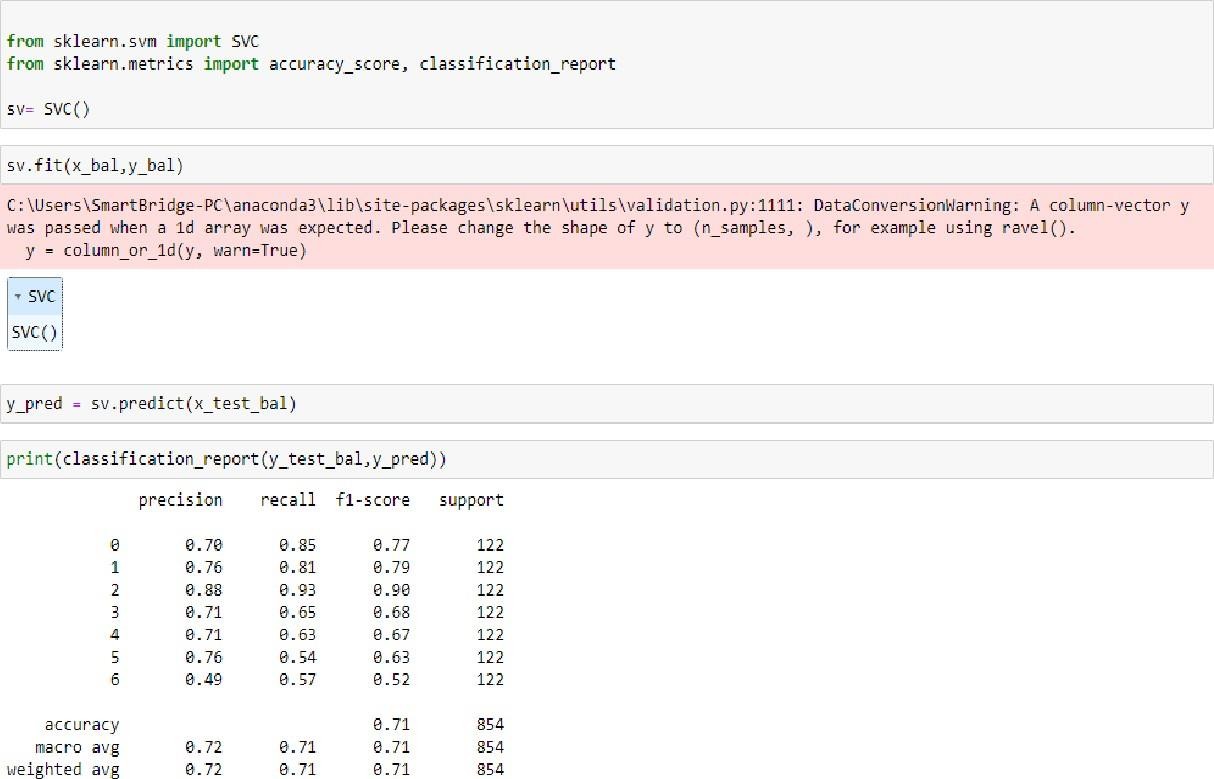
A function named XGBClassifier model is created and train and test data are passed as the parameters. Inside the function, the XGBClassifier algorithm is initialized and training data is passed to the model with the .fit() function. Testdata is predicted with the .predict() function and saved in a new variable. For evaluating the model, the accuracy score and classification report is done.



**Activity 1.3: SVC model**

A function named SVC model is created and train and test data are passed as the parameters. Inside the function, the SVC algorithm is initialized and training data is passed to the model with .fit() function. Test data is predicted with the .predict() function and saved in a new variable. For evaluating the model, the accuracy

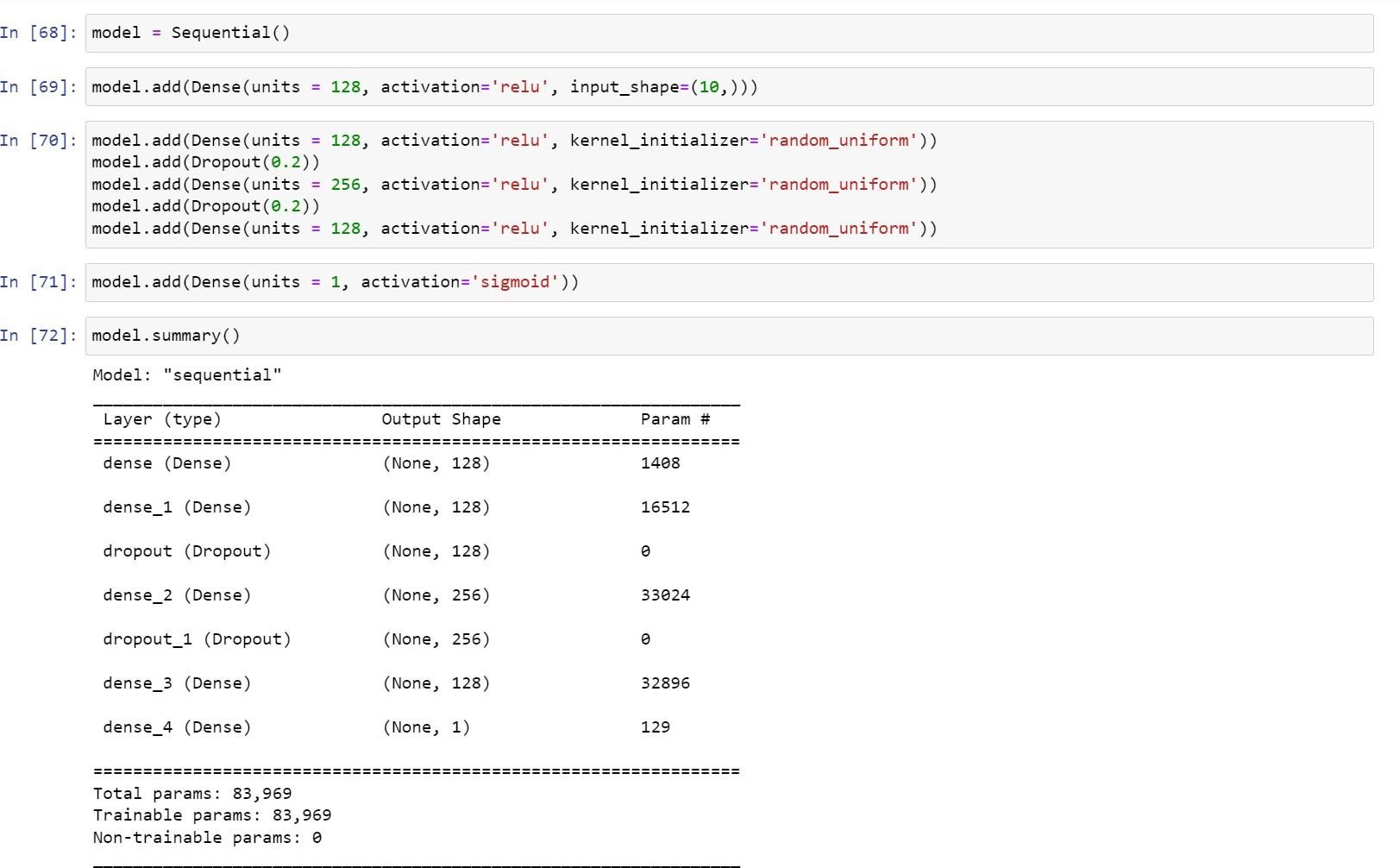
score and classification report is done.

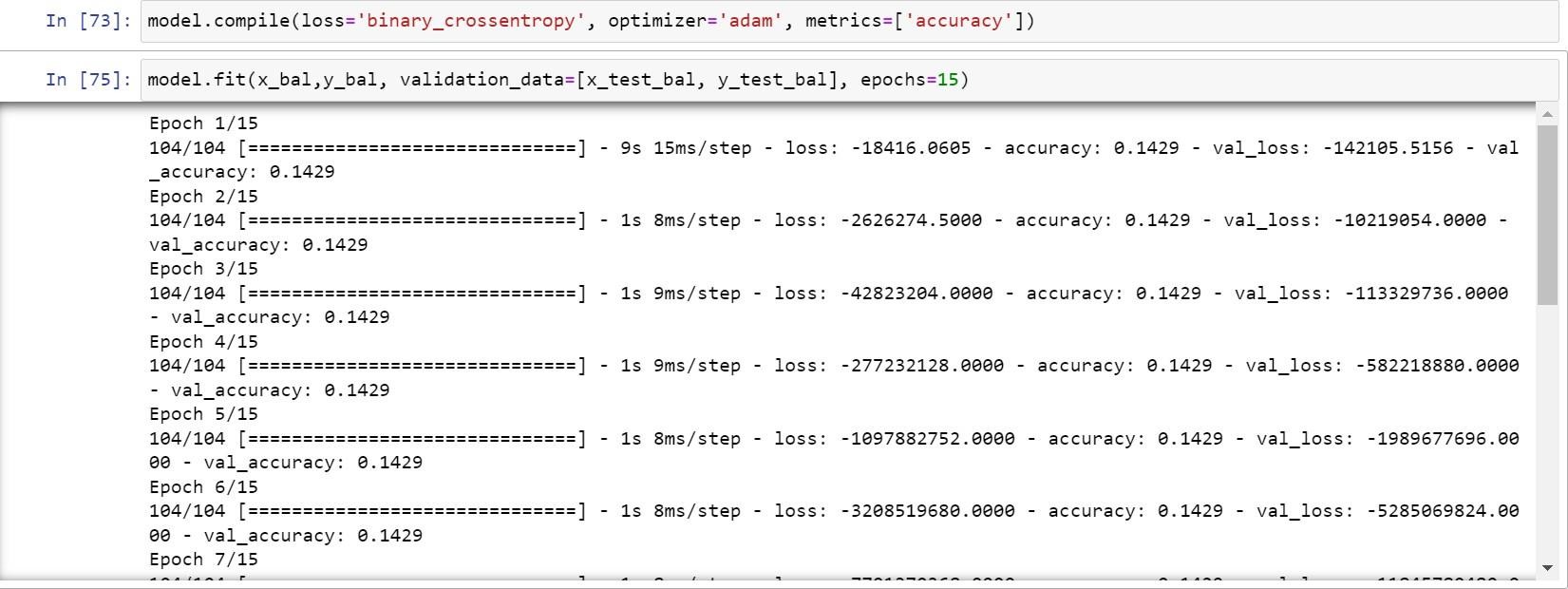




**Activity 1.4 ANN Model**

Artificial Neural Networks (ANN) are multi-layer fully- connected neural nets. They consist of an input layer, multiple hidden layers, and an output layer. Every node in one layer is connected to every other node in the next layer. We make the network deeper by increasing the number of hidden layers





**Activity 2: Testing the model**



# Milestone 5: Performance Testing & Hyperparameter Tuning

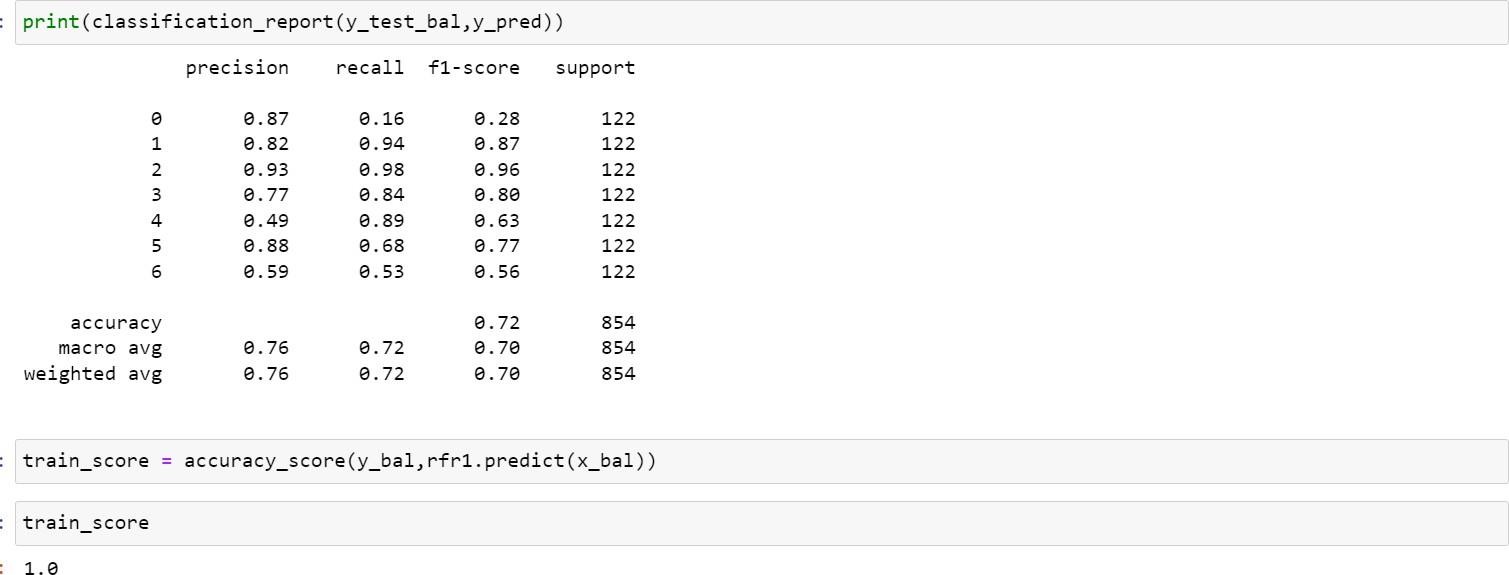
## Activity 1: Testing model with multiple evaluation metrics

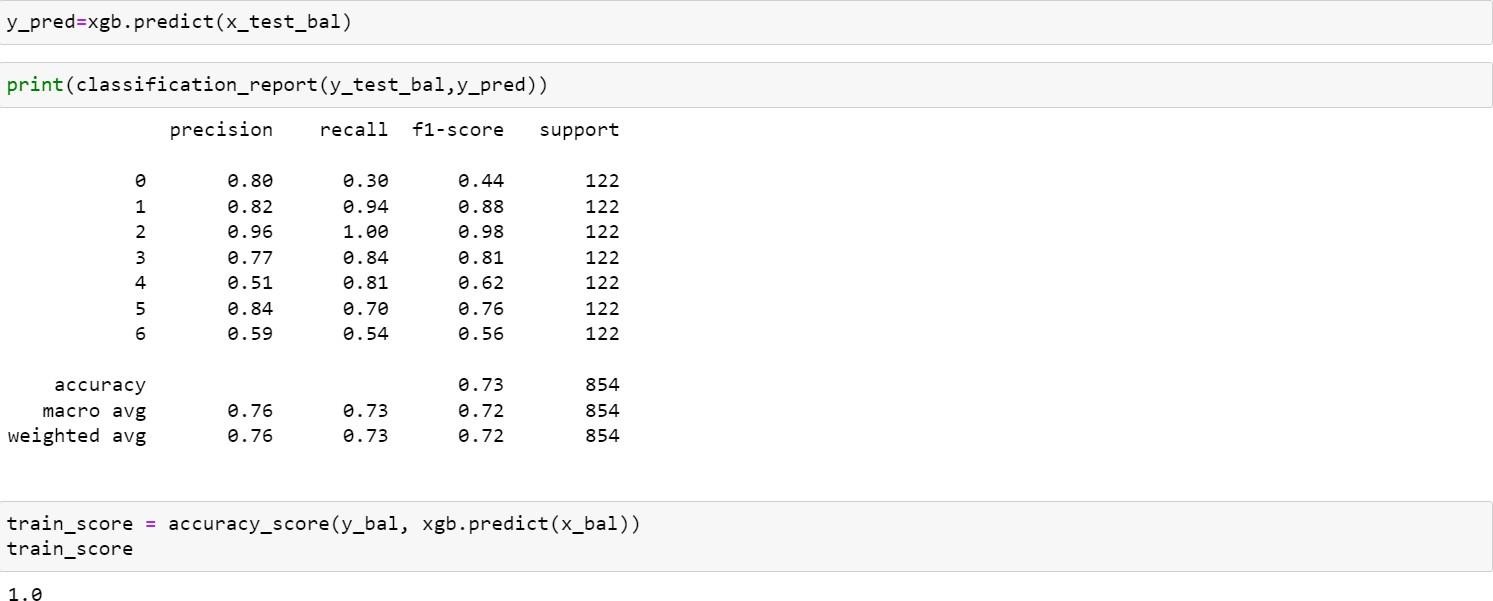
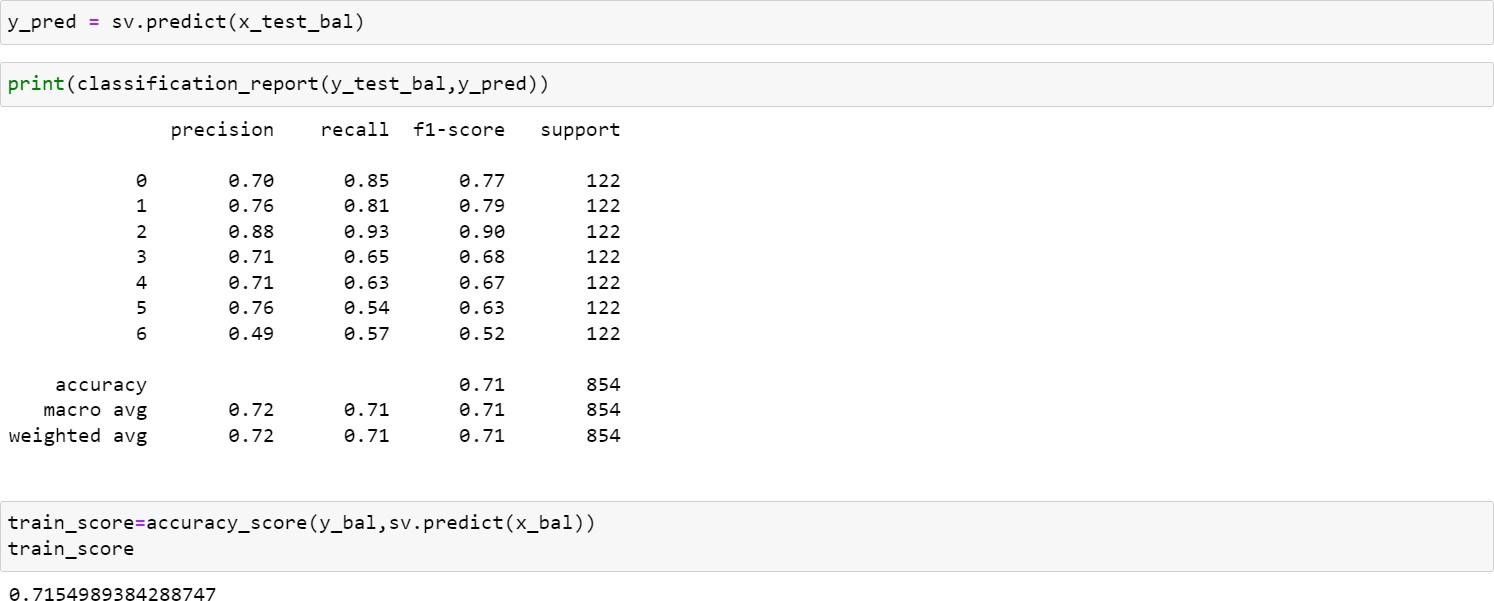
Multiple evaluation metrics means evaluating the model's performance on a test

set using different performance measures. This can provide a more comprehensive understanding of the model's strengths and weaknesses. We are using evaluation metrics for classification tasks including accuracy, precision, recall, support and F1-score.

**Activity 1.1: Compare the model**

For comparing the above four models, the compareModel function is defined.



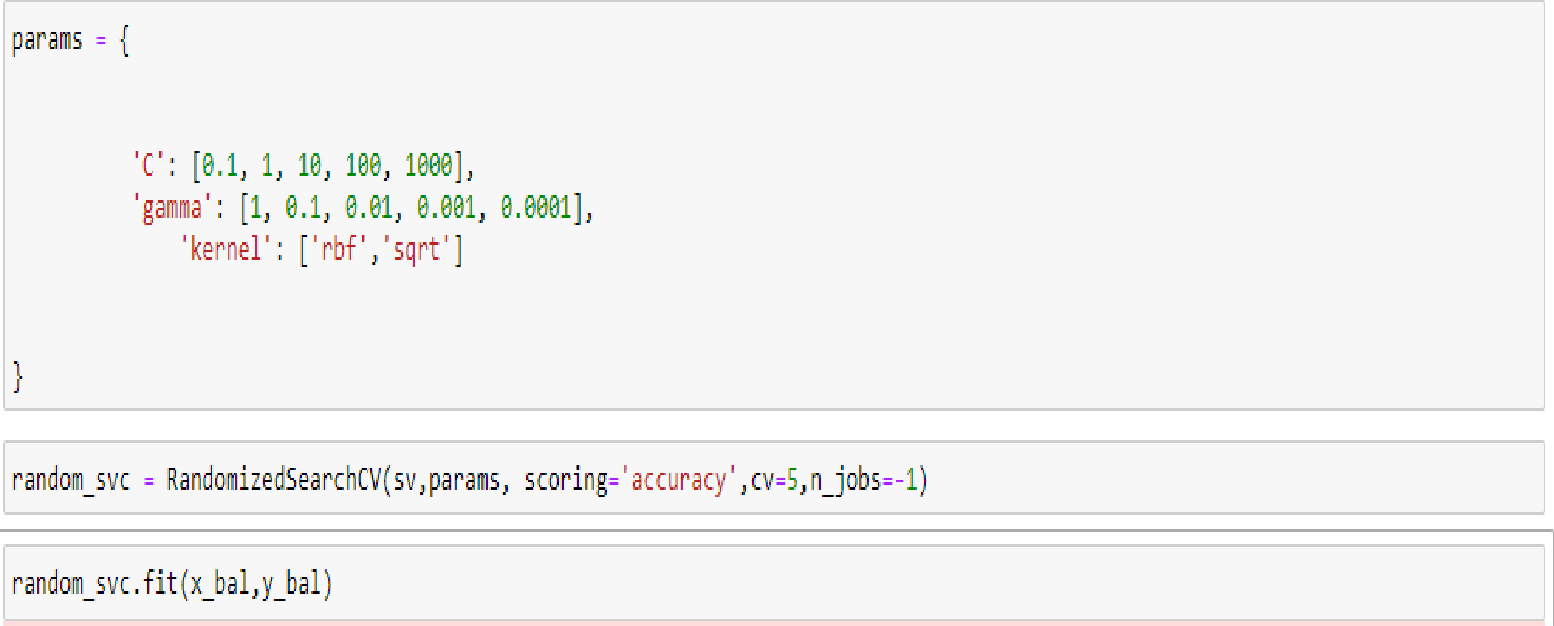




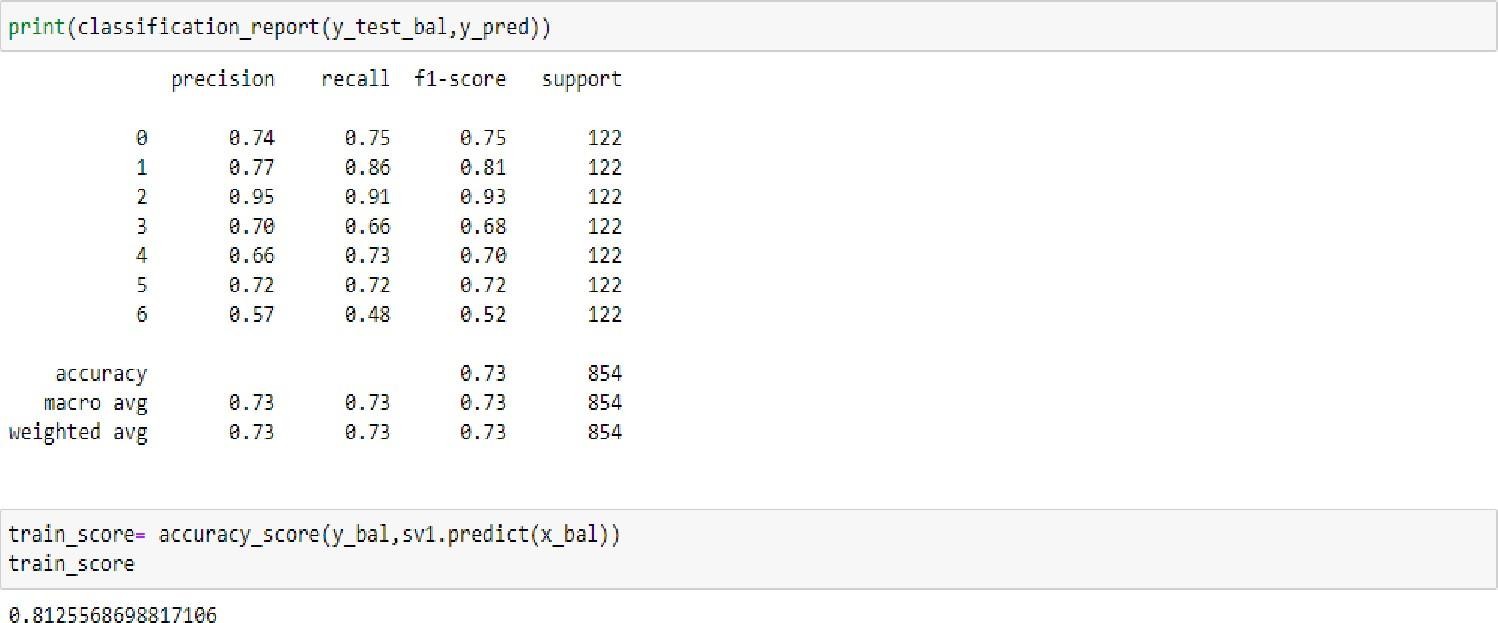
## Activity 2:Comparing model accuracy before & after applying hyperparameter tuning

From sklearn, accuracy is used to evaluate the score of the model. On the parameters, wehave given xgb1 (model name), x, y, cv (as 3 folds). Our model is performing well. So, we are saving the model by pickle.dump().

Note: To understand cross validation, refer to this link. [https://towardsdatascience.com/cross-validation-explained-](https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85) [evaluating-estimator-performa nce-e51e5430ff85](https://towardsdatascience.com/cross-validation-explained-evaluating-estimator-performance-e51e5430ff85).



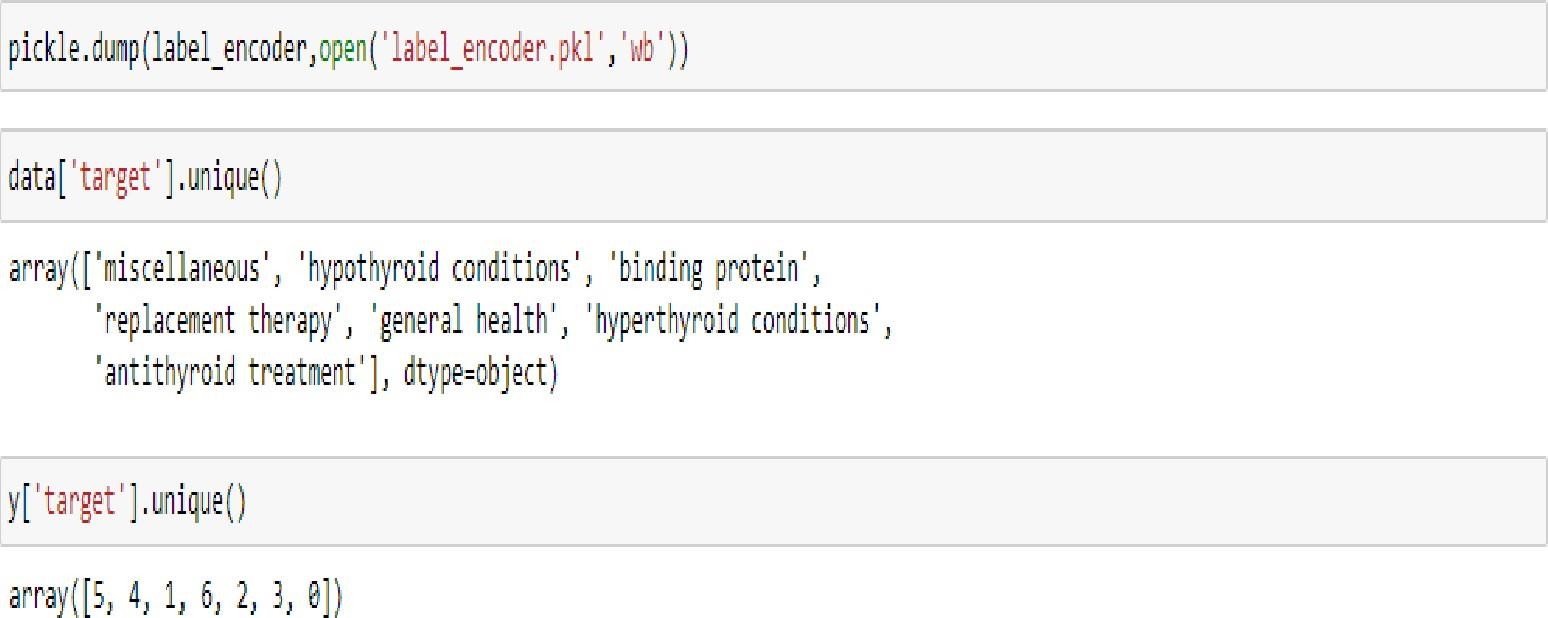




Saving the model as thyroid1\_model.pkl



Here, we are saving label\_encoding also as label\_encoder.pkl



# Milestone 6: Model Deployment

## Activity 1:Save the best model

Saving the best model after comparing its performance using different evaluation metrics means selecting the model with the highest performance and saving its

weights and configuration. This can be useful in avoiding the need to retrain themodel every time it is needed and also to be able to use it in the future.



## Activity 2: Integrate with Web Framework

In this section, we will be building a web application that is

integrated to the model we built. A UI is provided for the uses where he has to enter the values forpredictions. The enter values are given to the saved model and prediction is

showcased on the UI.

This section has the following tasks

* + Building HTML Pages
  + Building server side script

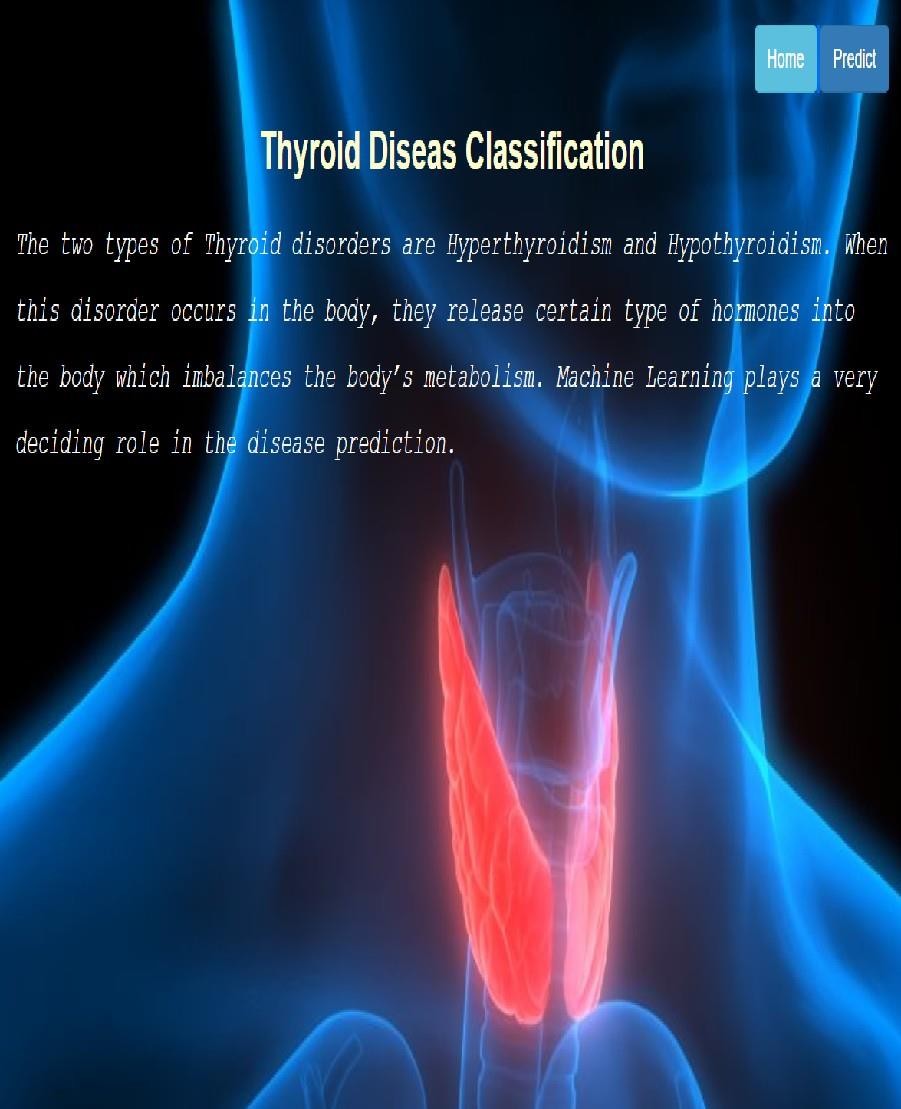
**Activity 2.1: Building Html Pages:**

For this project create three HTML files namely

* + home.html
  + predict.html
  + submit.html

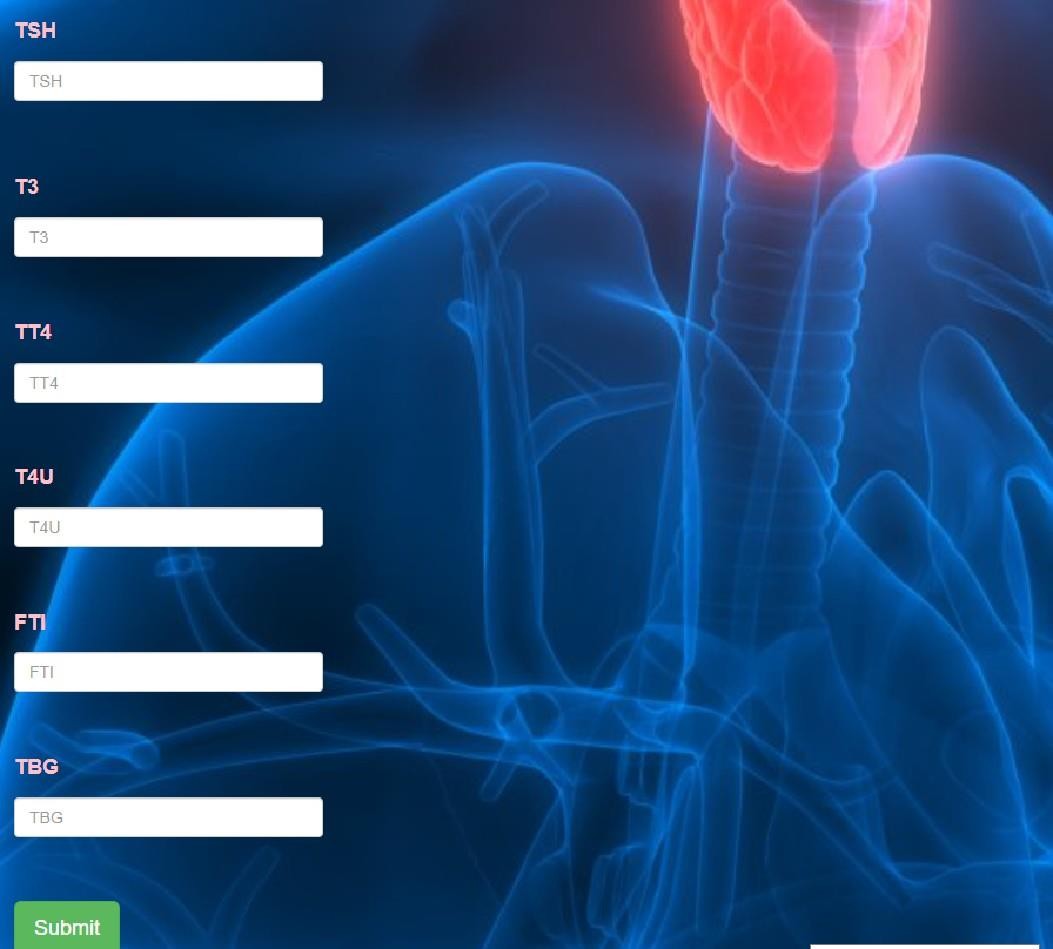
and save them in the templates folder.

Let‟s see how our home.html page looks like:



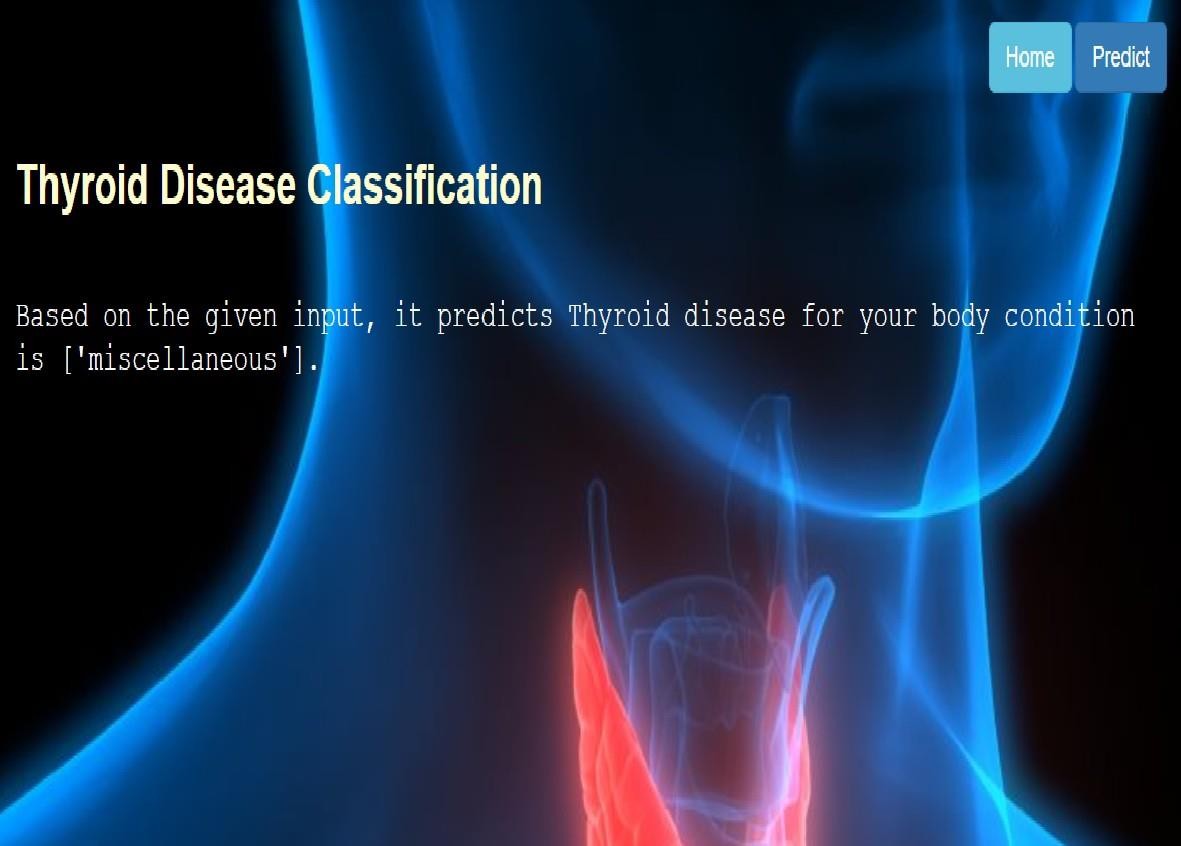
Now when you click on predict button from top right corner you will get redirected to predict.html

Let's look how our predict.html file looks like:



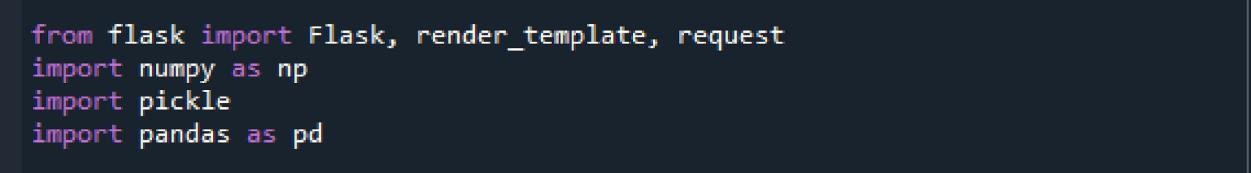
Now when you click on submit button from left bottom corner you will get redirected to submit.html

Let's look how our submit.html file looks like: it is ['miscellaneous'].

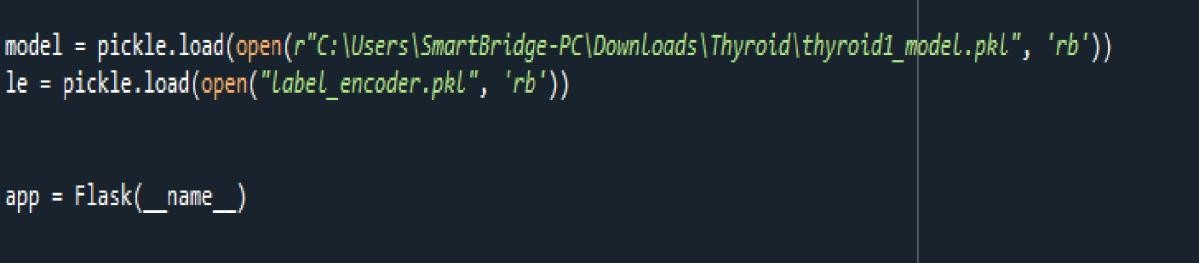


**Activity 2.2: Build Python code:**

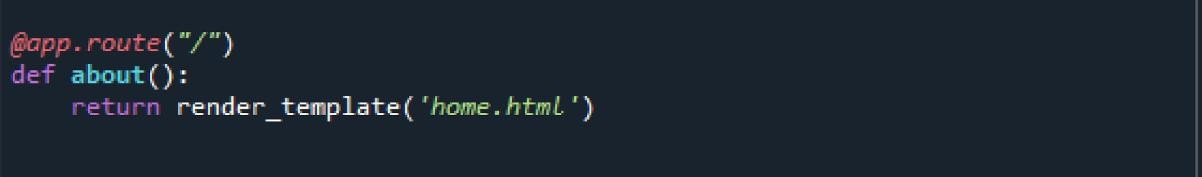
Import the libraries



Load the saved model. Importing the flask module in the project is mandatory. An object of Flask class is our WSGI application. Flask constructor takes the name of the current module ( name ) as argument.



Render HTML page:

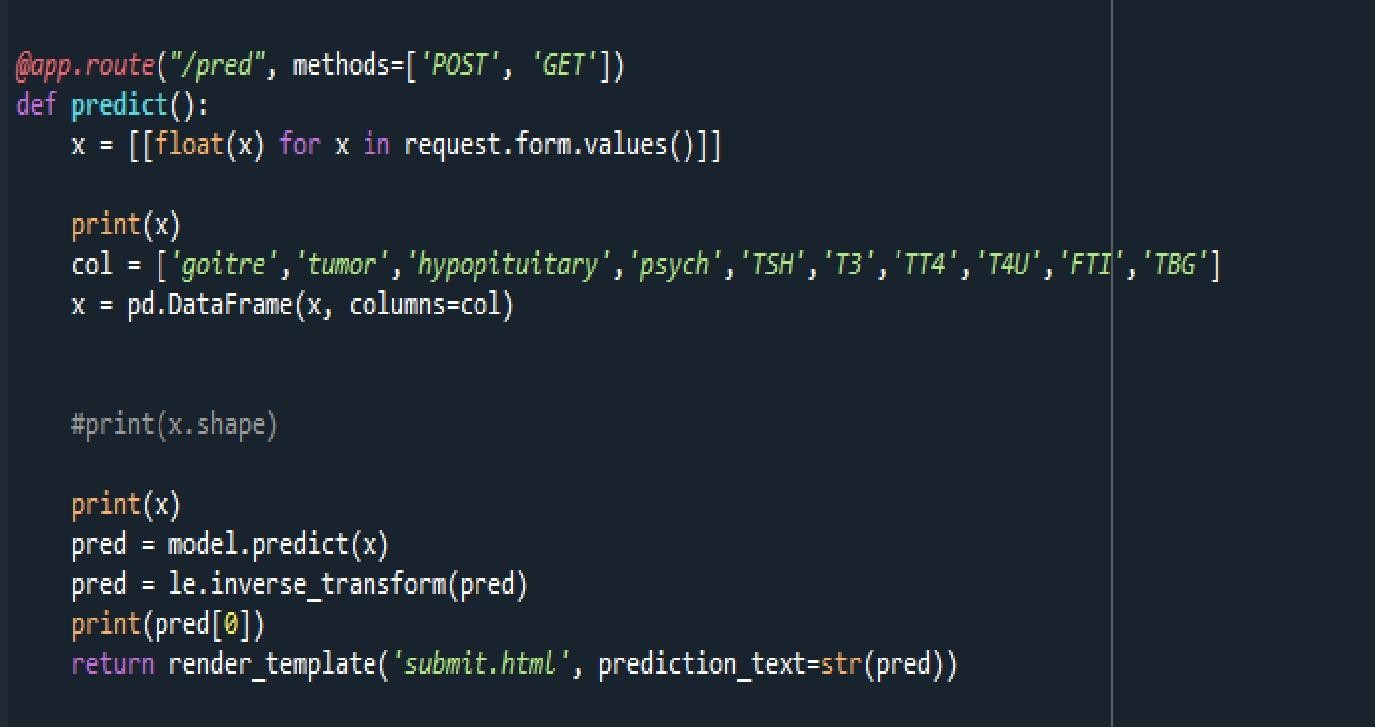


Here we will be using a declared constructor to route to the HTML page which we have created earlier.

In the above example, „/‟ URL is bound with the home.html function. Hence, when the home page of the web server is opened in the browser, the html page will be rendered.

Whenever you enter the values from the html page the values can be retrieved using POST Method.

Retrieves the value from UI:



Here we are routing our app to predict() function. This function retrieves all thevalues from the HTML page using Post request. That is stored in an array. This array is passed to the

model.predict() function. This function returns the

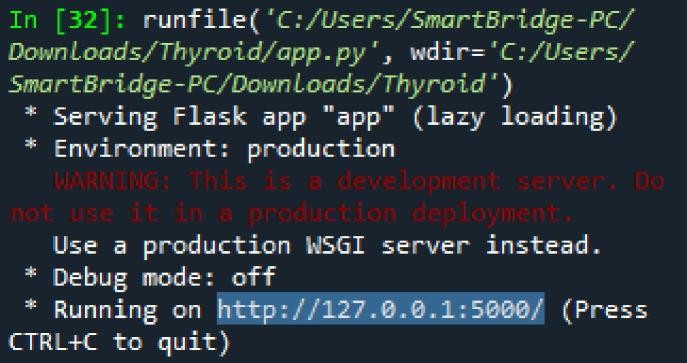
prediction. And this prediction value will be rendered to the text that we havementioned in the submit.html page earlier.

Main Function:



**Activity 2.3: Run the application**

* Open anaconda prompt from the start menu
* Navigate to the folder where your python script is.
* Now type “python app.py” command
* Navigate to the localhost where you can view your web page.
* Click on the predict button from the top right corner, enter the inputs,click on the submit button, and see the result/prediction on the web.



# Milestone 7: Project Demonstration & Documentation

Below mentioned deliverables to be submitted along with other deliverables

## Activity 1:- Record explanation Video for the project end to end solution

**Activity 2:- Project Documentation-Step by step project developmentprocedure**

Create document as per the template provided