**Curneu MedTech Innovations Private Limited**

**TASK - SD03Q01**

**Problem statement 1: Social\_Network\_Ads.csv**

**AIM**:

To understand the dataset of Social\_Network\_Ads.csv and find the best suitable ML algorithm.

**PROBLEM ANALYSIS:**

For the given dataset, Random forest algorithm has been used for classification. The dataset consists of five columns. The first two columns ‘User ID’ and ‘Gender’ have been dropped for easier implementation. The two columns, namely the ‘**Age ’**and the ‘**EstimatedSalary ’,** are the independent variables and the last column is the dependent variable ‘**Purchased’**, which is in the binary format denoting whether the individual has bought the product (1) or not (0). Using the dataset, Classifier is built for a product company that will classify whether a person of a particular age and drawing a particular amount of salary will purchase their product that is advertised on a social media platform.

**STEPS:**

The implementation was performed using python code and Google colab. The following are the steps involved in the code:

1. Importing libraries like Numpy, pandas and Matplotlib
2. Importing the dataset
3. Splitting the dataset into the testing and training set
4. Feature scaling
5. Training the random forest classification model on the training set
6. Predicting the test set results
7. Confusion matrix and accuracy
8. Comparing the real values with predicted values
9. Visualizing the results

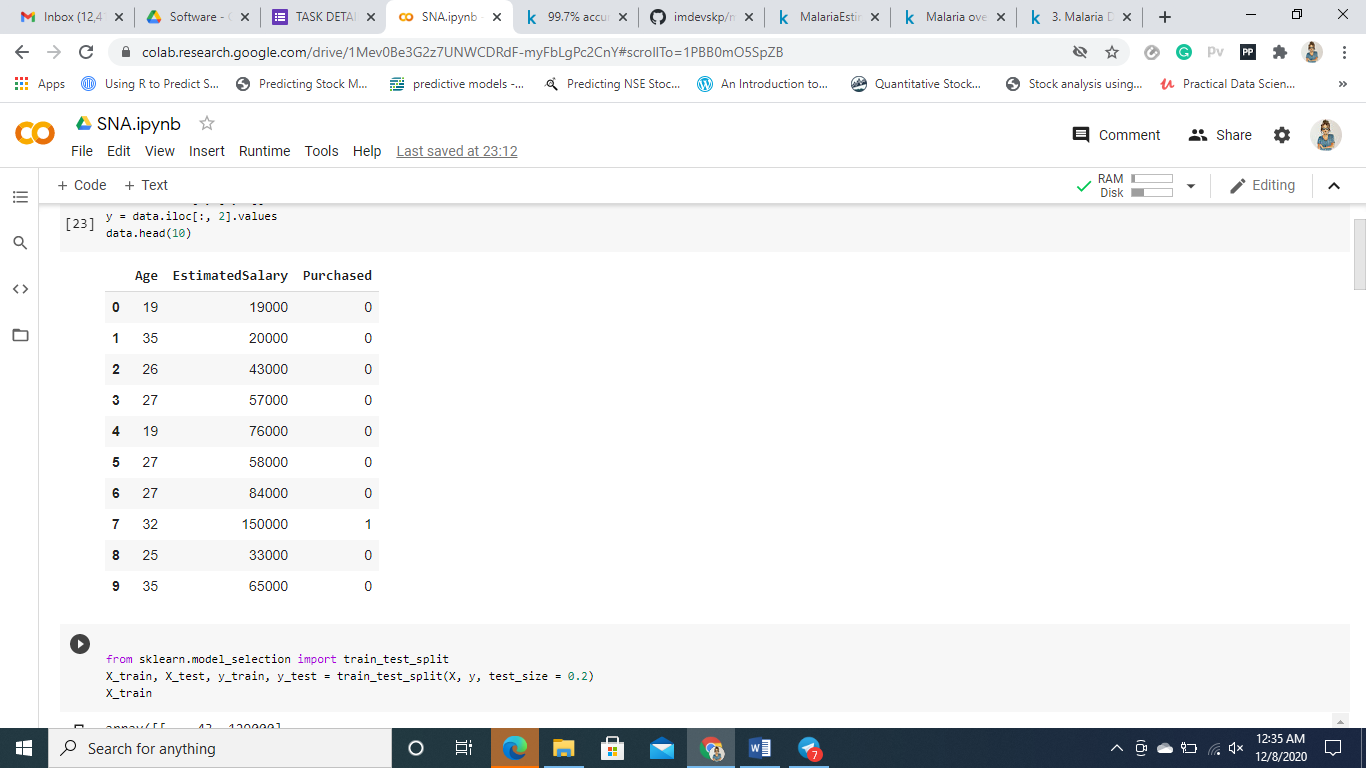
**METHODOLOGY:**

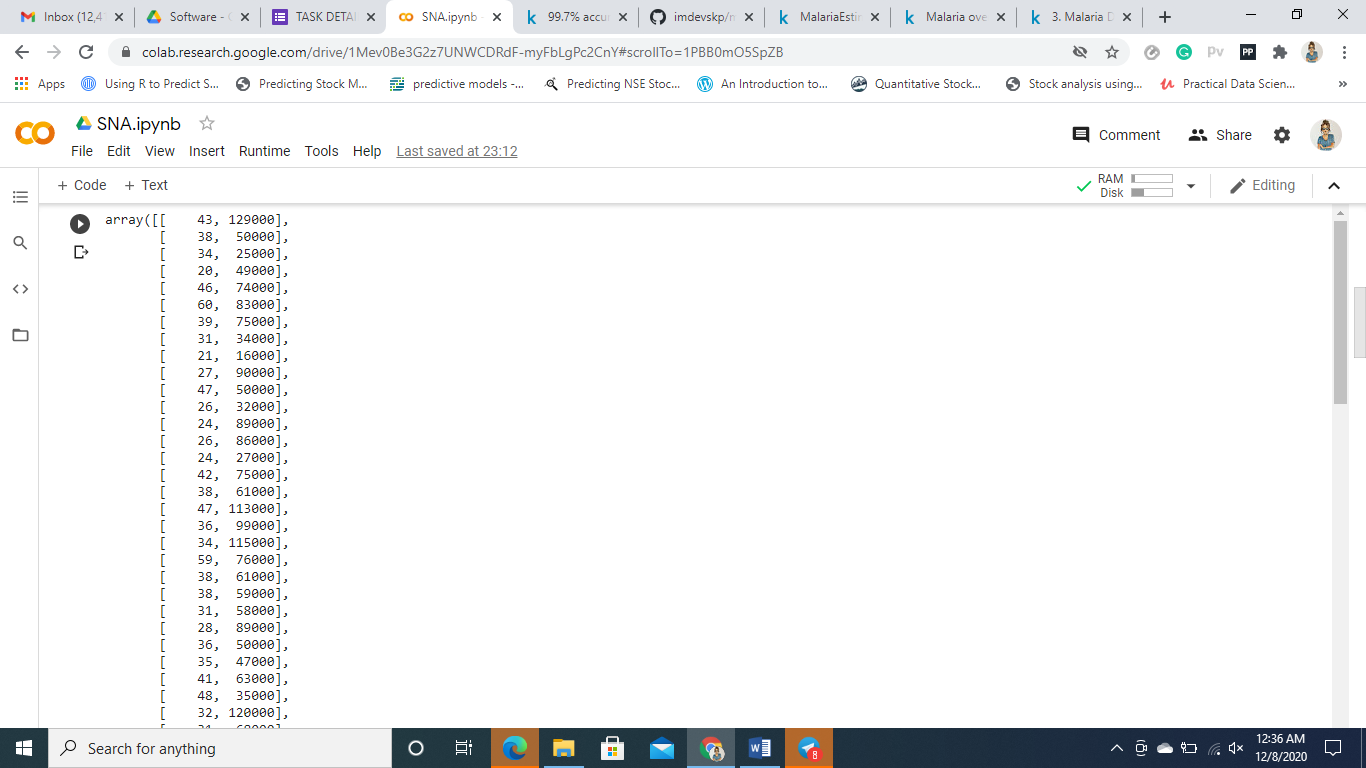
The first step is to import the necessary libraries like NumPy, pandas and Matplotlib. The dataset is then loaded and stored to the variable data. Then the corresponding variables to X and Y are assigned. The first 10 rows of the dataset are displayed. The data is split into training and test set. In this, the test\_size=0.20. It denotes that **20%** of the data will be kept as the **Test set** and the remaining **80%**will be used for training as the **Training set**. As there are 400 rows, about 80 data points will be assigned to the test set and remaining 320 data points will be used for training purposes. This is an additional step that will increase the speed of the program as we scale down the values of X to a smaller range. In this, both the X\_trainand the X\_test  are scaled down to a small range of -2 to +2. For example, the salary 75000 is scaled down to 0.16418997. Once the training test is ready, the RandomForestClassifier Class is imported and fit the training set to the model. The class SVC is assigined to the variable classifier. The criterion used here is “**entropy”** . The other criterion that can also be used is the “**gini**”. The classifier.fit() function is then used to train the model. In this step, the classifier.predict() function is used to predict the values for the Test set and the values are stored to the variable y\_pred.. In this, The next step is calculating the accuracy of the trained model and plot the confusion matrix.

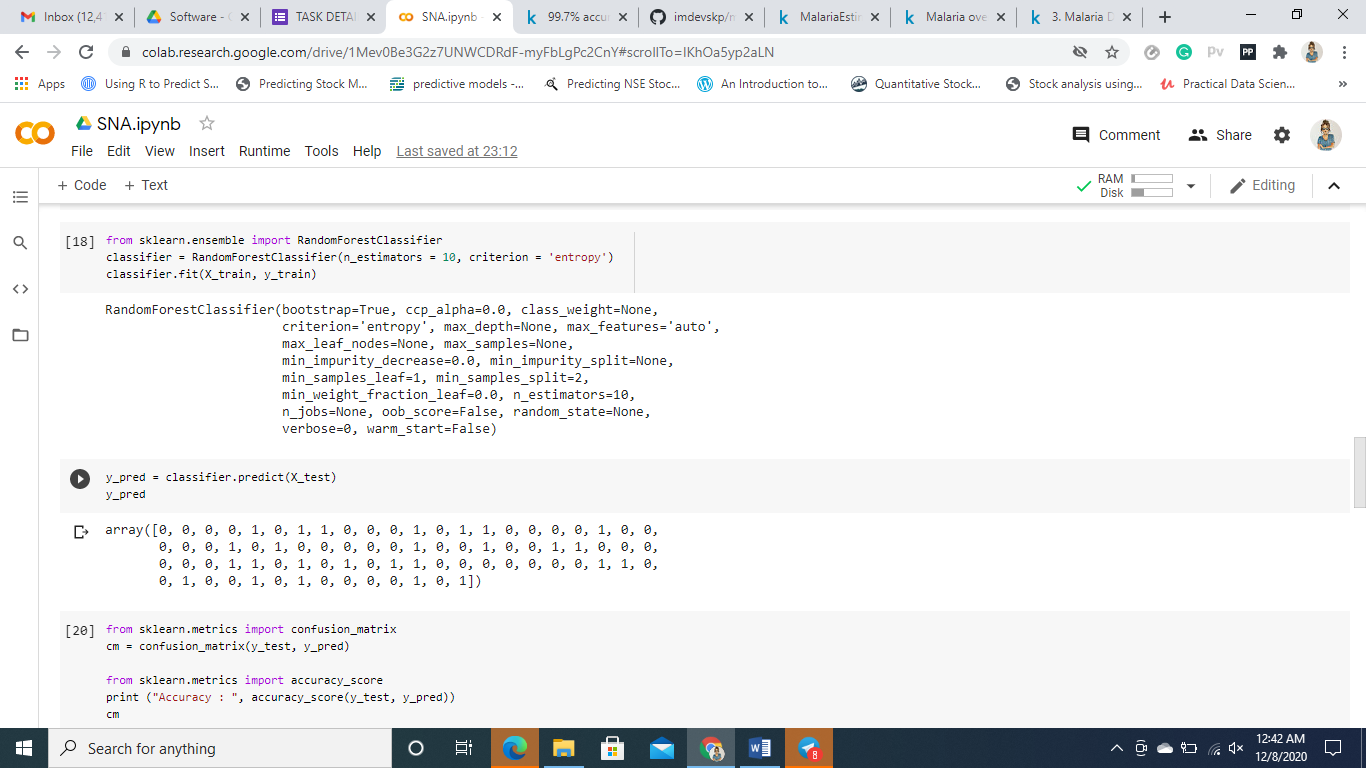
The confusion matrix is a table that is used to show the number of correct and incorrect predictions on a classification problem when the real values of the Test Set are known. The True values are the number of correct predictions made.

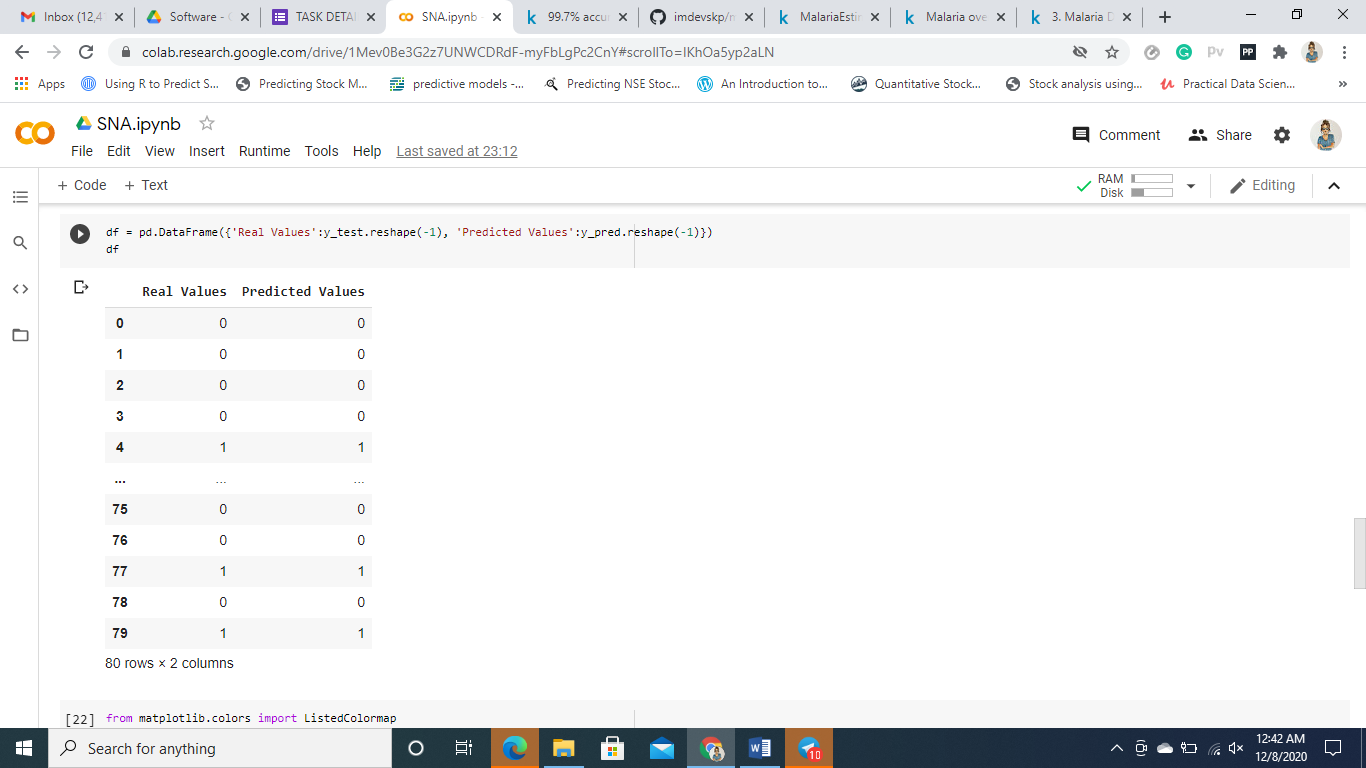
The confusion matrix of the model built shows that, out of 80 test set data, 77 were correctly classified and only 3 were incorrectly classified, leaving us with an accuracy of 96.25%. In the next step, a Pandas DataFrame is created to compare the classified values of both the original Test set (**y\_test**) and the predicted results (**y\_pred**). This step is an additional step which is not much informative as the Confusion matrix and is mainly used in regression to check the accuracy of the predicted value. The last step is to visualize the results of the Random Forest Classification model on a graph that is plotted along with the two regions. In the plot obtained, there are two regions. The **Red**region denotes **0**, which consists of people who have not bought the product and the **Green**region denotes **1**, which consists of the people who have bought the product.

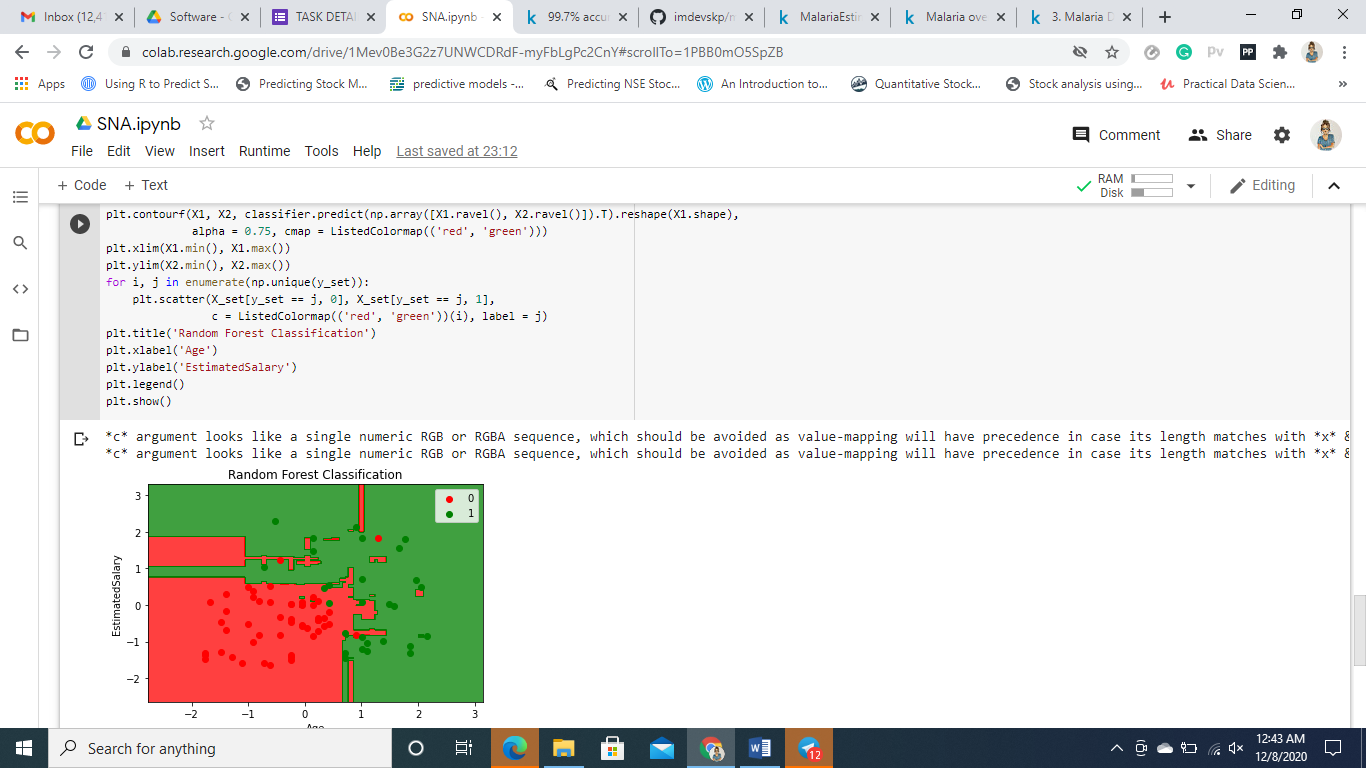
**OUTPUT**:











**CONCLUSION:**

Thus in the task, a **Random Forest Classification**Model that is able to predict if a person is going to buy a product based on his age and salary, built successfully.

**Problem statement 2: Malaria**

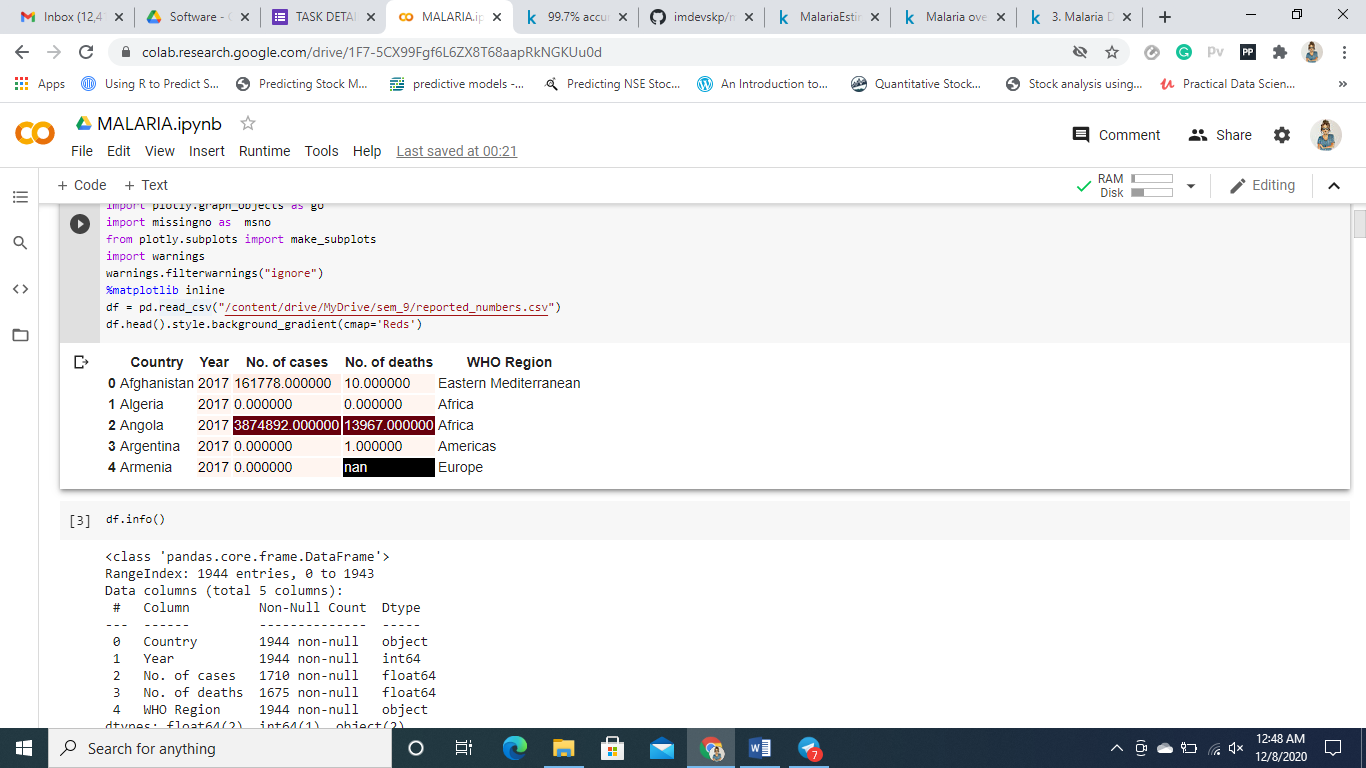
**AIM:**

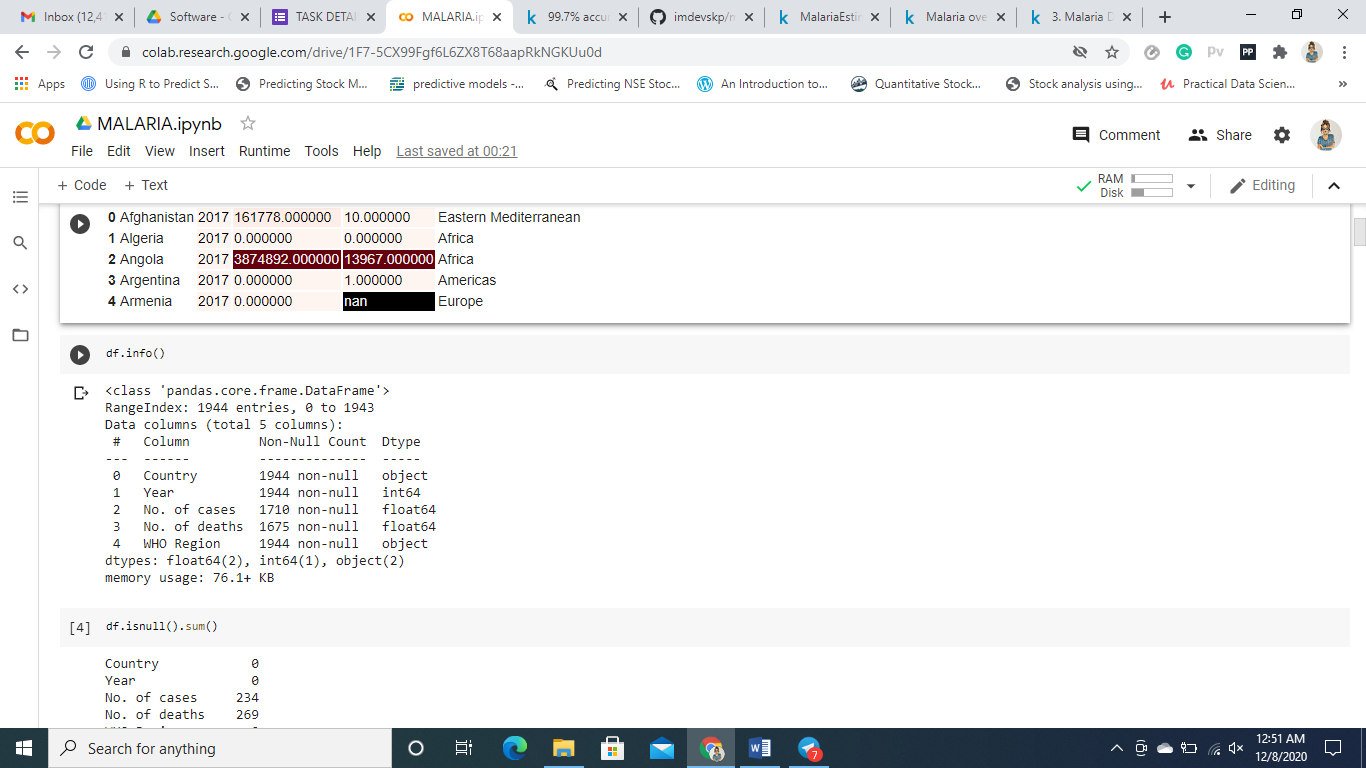
To explore whether the no. of cases, deaths, CFR of malaria increases every year.

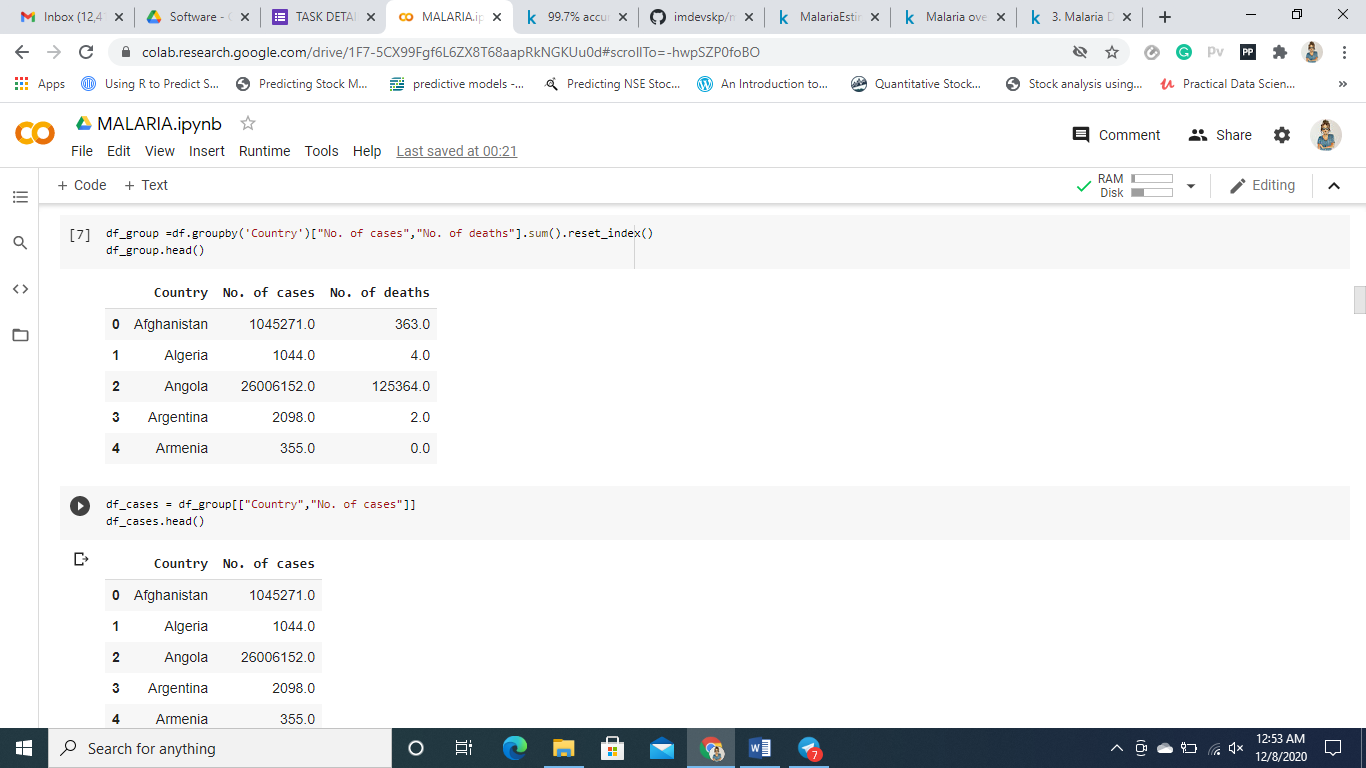
**STEPS:**

The first steps is importing the necessary libraries like numpy, pandas, matplotlib, seabor, plotly etc. The dataset is then imported and stored in the variable *df.* The dataset is checked for null values. If any null values are present, they are dropped. The next step is data visualization. The results are compared with the results of the respective countries, group the data with respect to country. Top 20 countries with highest number of malaria cases and number of deaths are visualized.

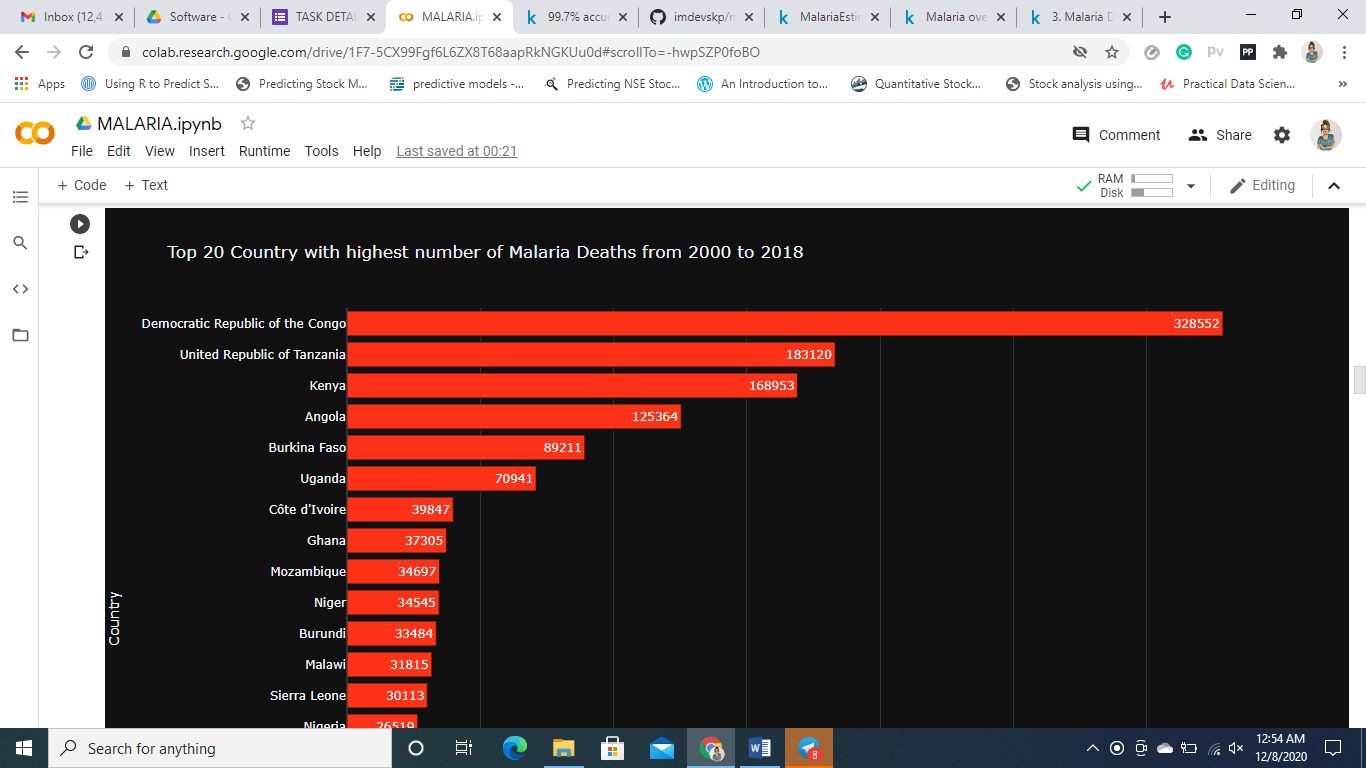
**OUTPUT:**

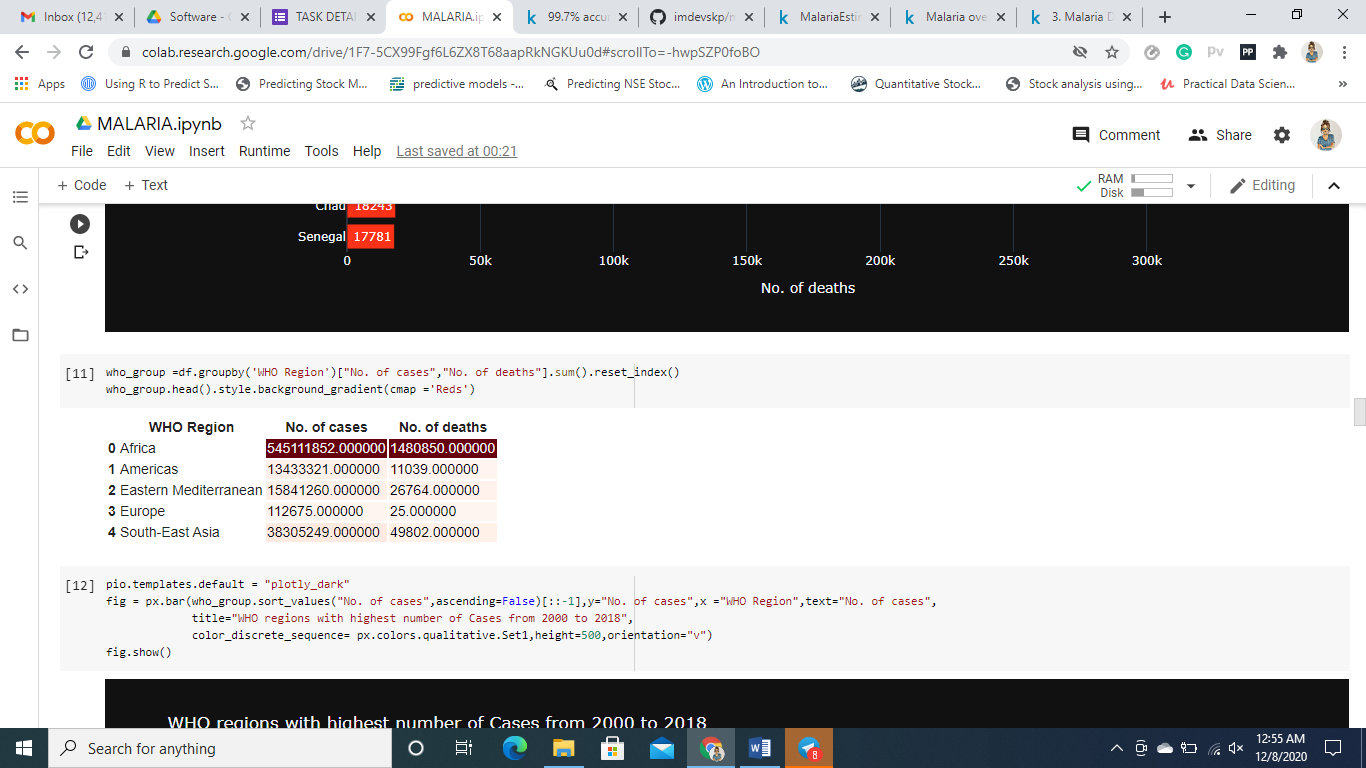


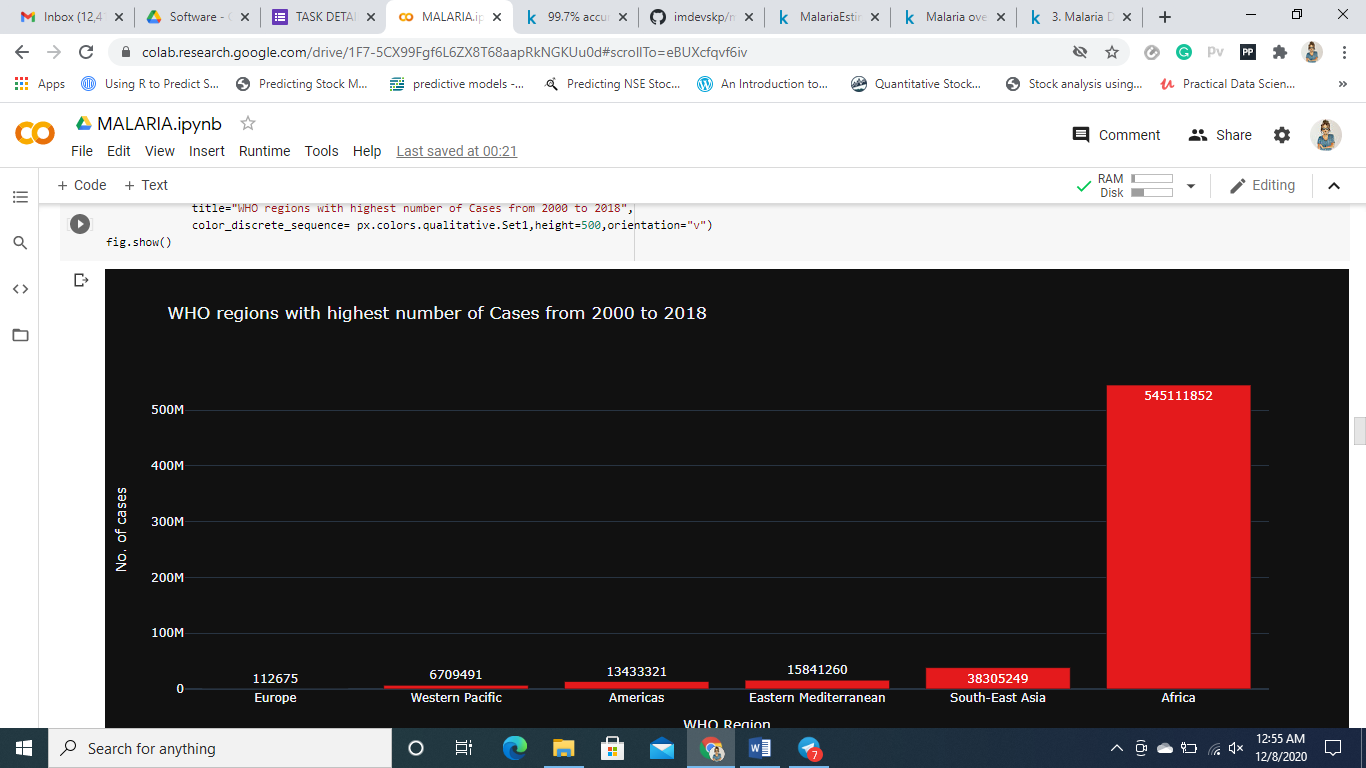














**CONCLUSION:**

From the above analysis, it can be concluded that there has been a steady decrease in the number of malaria cases over the years in all the countries and the number of deaths have also decreased steeply.