

GDP/CAPTIA PREDICTION USING IBM WATSON MACHINE LEARNING

DevelopedBy:M.shravya, S.sanath, B.prasanna, Nizami

SmartBridge–MiniProjectReport

1.INTRODUCTION

1.1 Overview:

Gross Domestic Product (GDP) is the market price of all product and services that area unit made inside the country's national borders in a year. Gross domestic product could be a measure to assess overall economic performance of a country, it includes all product and services made by the economy as well as personal consumption, government purchase, non-public inventories, paid in construction prices and therefore the foreign trade gap. The topic of GDP became of high importance among the indicator of economy variables. Information on Gross Domestic Product is thought to be a crucial indicator for evaluating the national economic development and growth of entire macro economy. GDP aggregates the complete economic motion. It is frequently used as a finest measure to calculate the performance of the economy. GDP is mostly measured in one of a 3 approaches. First, the Expenditure approach, it involves the worth of all domestic expenditures created on final product and services of the year, beside consumption expenditures, investment expenditures, government expenditures, and web exports. Second, the assembly approach, it's involving the summation of all additional activities at each a part of production by all industries inside the country, taxes and product's subsidies of the amount. Third is that the financial gain approach, it's the summation of all aspect of the financial gain created by production inside the economy as remuneration of workers, capital financial gain, and gross in operation surplus of enterprises i.e., profit, taxes on production and imports less grants of the quantity.

1.2 Purpose:

The aim of this study is to predict GDP, using linear regression and random forest for a particular period. Prediction of GDP involves application of applied mathematics and mathematical model to predict future developments within the economy. It permits to review previous economic movements and predict however current economic changes can amend the patterns of previous trend; therefore, a more accurate prediction would provide a significance facilitate to the government in setting up economic development goals, ways and policies. Consequently, a correct Gross Domestic Product prediction presents a number one insight associate an understanding for future economics' trend.

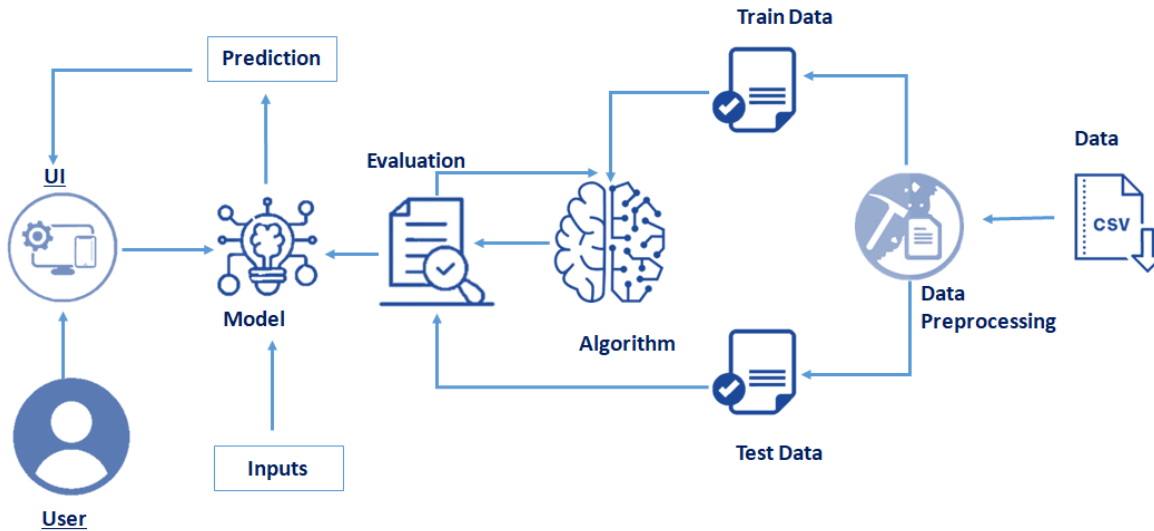
2 LITERATURE SURVEY

Gross Domestic Product's growth rate is treated as a sign of the economic health of the country. A number of studies demonstrates the factors for prediction of GDP using various methodologies. The GDP data ranging from the year 1989 to 2007 of Anhui region in particular was studied by Gang Long [1]. The method depicts the comparative performance of the GA-SVM and RBF neural networks respectively. Jaehyun Yoon [2] explored the Gross Domestic Growth of Japan from the year 2001 to 2018. The data is collected from International Monetary Fund and Bank of Japan. The author worked with gradient boosting and random forest machine learning classifier. MAPE and RMSE method are taken into consideration for the purpose of measuring accuracy of the model. Further, cross validation and hyper parameter tuning are used for the creation of more accurate models. The vector machine was trained with genetic algorithm and henceforth used for GDP forecasting. Relative error method was used to evaluate the model performance. The author concluded that in SVM, optimal solution in short time was acquired by genetic algorithm which worked as a better approach in parameters selection of SVM. For optimizing the support vector machine's parameters, Genetic algorithm was introduced. Various Economic Indicators play a vital role in Gross Domestic Product prediction. Consumption is normally the largest GDP component in the economy. John [3] coined Real Government Consumption Expenditures, Real Personal Consumption Expenditures and Gross Private Domestic Investment as more vital indicators for predicting GDP. Autoregressive approach predicts consistent future growth in terms of factors related to GDP but fails to overcome historic economic recession. Shelly and Wallace [4] studied the relation between M1 money, real GDP and inflation in Mexico. Annual data from the year 1944 to 1991 is studied. This work indicates that a positive effect on real Gross Domestic Product growth is obtained by unpredictable increases in differenced inflation while predictable increases in differenced inflation results in negative impact on real Gross Domestic Product growth.

In order to produce short-term forecasts of real Austrian GDP, Schneider M. and Spitzer M [5] utilized a generalized dynamic factor model. Macro-Economic Variables has a great influence in country's GDP. Amongst factors like service, agricultural and livestock sector, business sector and industrial sector proves to be dominating one as far as contribution to the GDP is concerned [6]. The influence of small medium enterprises was described by author Maciej Woźniak [7] stating small medium enterprises plays key social role as they reduce unemployment. Carlos Encinas Ferrer [8] stated why Foreign Direct Investment does not show as an independent variable since FDI has small proportion within the national investment in the countries like Brazil, China, Peru, Mexico and therefore lead to low multiplier effect on the national economy.

3.THEORITICALANALYSIS

3.1 Block diagram:



3.2 Hardware and software requirements in the project :

The following is the Hardware required to complete this project:

- Internet connection to download and activate
- Administration access to install and run Anaconda Navigator
- Minimum 10GB free disk space
- Windows 8.1 or 10 (64-bit or 32-bit version) OR Cloud: Get started free, *Cloud account required.

Minimum System Requirements To run Office Excel 2013, your computer needs to meet the following minimum hardware requirements:

- 500 megahertz (MHz)
- 256 megabytes (MB) RAM
- 1.5 gigabytes (GB) available space
- 1024x768 or higher resolution monitor

The following are the software required for the project:

- Google Colaboratory Notebook and Jupyter Notebook
- Spyder and Pycharm Community
- Microsoft Excel 2013

4. EXPERIMENTAL INVESTIGATIONS

4.1 Linear Regression:

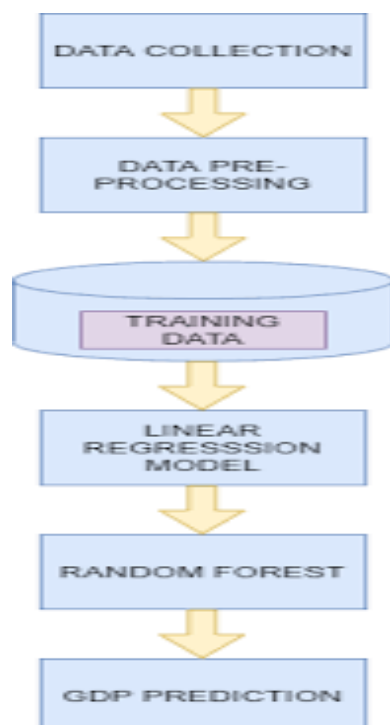
It is supervised machine learning algorithm, the most basic type of regression. Basically, it is the mathematical model that analyses the linear relationship between a dependent variable with given set of independent variables(s). In the project the simple linear regression was used to predict the individual attribute of the dataset. For this 80% of the dataset was the training dataset i.e., used for training the model and remaining 20% was used to test the dataset.

4.2 Random Forest:

Random Forest is one of the well-known machine learning algorithms that belongs to supervised learning technique. Random Forest is used both for regression and classification problems in machine learning. Ensemble Learning is a concept in which multiple classifiers are integrated in order to resolve a complicated drawback and hence it improves the performance of the model. Random Forest relies on the concept of ensemble learning

As the name itself suggests, "Random Forest is basically a classifier that consists of decision trees of the given dataset on varied subsets. Further, the random forest takes the average in order to improve the forecasting accuracy." Predictions from every tree that is formed are taken into consideration instead of just relying on a single decision tree and after that; based on majority votes of prediction, output is predicted. The classification of the classes in the project are done on the basis of the predicted data from linear regression. The results of this implementation and its analysis are mentioned further.

5.FLOWCHART



6. RESULTS :

Student Dashboard GDP Analysis

127.0.0.1:5000

GDP Analysis

Home Predict

Introduction

GDP stands for "Gross Domestic Product" and represents the total monetary value of all final goods and services produced (and sold on the market) within a country during a period of time (typically 1 year). Purpose: GDP is the most commonly used measure of economic activity.

Type here to search

USD -3.00%

1:00 PM 11/11/2022

Student Dashboard GDP Analysis

127.0.0.1:5000/predict

GDP Analysis

Population: Range 7026 to 1313973713

Area (sq. mi.): Range 2 to 17075200

Population Density (per sq. mi.): Range 0.0 to 16271.5

Coastline (coast/area ratio): Range 0.0 to 870.66

Net migration: Range -20.99 to 23.06

Infant mortality (per 1000 live births): Range 2.29 to 191.19

Literacy (%): Range 17.6 to 100.0

Arable (%): Range 50.0 to 62.11

Crops (%): Range 0.0 to 50.68

Deathrate: Range 2.29 to 29.74

Agriculture: Range 0.0 to 0.769

Industry: Range 0.02 to 0.9059999999999999

Service: Range 0.062 to 0.9540000000000001

Region_label: ASIA

Climate_label: 1.

Type here to search

USD -3.00%

OneDrive - Personal Not signed in

Student Dashboard x GDP Analysis x +

127.0.0.1:5000/predict

GDP Analysis

Population: 9989

Area (sq. mi.): 5

Population Density (per sq. mi.): 55

Coastline (coast/area ratio): 55

Net migration (per 1000): 23

Infant mortality (per 1000 live births): 23

Literacy (%): 23

Arable (%): 57

Crops (%): 26

Deathrate: 27

Agriculture: 0.6

Industry: 0.3

Service: 0.6

Region_label: ASIA

Climate_label: 3

Predict

Student Dashboard x GDP Analysis x +

127.0.0.1:5000/data_predict



7.ADVANTAGES & DISADVANTAGES

ADVANTAGES:

- GDP is important because it gives information about the size of the economy and how an economy is performing.
- Many businesses forecast the economy to make important decisions that may impact their future processes or policies.
- Broad indicator of development
- Easy to measure growth in percentage
- Easy to compare to itself and other countries.

DISADVANTAGES:

- The GDP cannot be fully accurate because it fails to account for transactions that aren't recorded, free labor or changes in actual income, quality changes of goods, and goods produced but not exchanged for money.
- Forecasts are Never Completely Accurate - Forecasts are never 100% and it is almost impossible to predict the future with certainty.
- GDP doesn't count unpaid volunteer work
- Disasters can raise GDP
- GDP doesn't account for quality of goods

8.APPLICATIONS:

- GDP is important because it gives information about the size of the economy and how an economy is performing.
- Economists can use GDP to determine whether an economy is growing or experiencing a recession.

9.CONCLUSION

The resultant study encourages the utilization of machine learning classifiers namely linear regression and Random Forest in macroeconomic data forecasting. On the basis of optimization process, the machine learning algorithm “Random Forest” utilized during this study worked well with the accuracy 86 percentage in order to predict the true GDP per capita. Random Forest Classifier produces more accurate forecasts as compared to the linear regression. Accuracy is measured by MAE and RMSE evaluation metrics. The main focus of traditional economics models is mainly on explanations of relationships whereas machine learning classifiers target predictions. Though it may seem as Machine learning models do not turn out to be good performers while discovering the impact of independent variable on the dependent variable or analyzing a causal relationship. However, as described in this paper and in previous studies as well, machine learning models often tend to convey and exhibit high prediction power. In future this model can be improved by using better machine learning algorithm which may result in even better performance.

10.FUTURESCOPE:-

- GDP was not designed to assess welfare or the well being of citizens. It was designed to measure production capacity and economic growth.
- It is a welfare measure because it captures future consumption opportunities.
- GDP per capita convergence in the future. ... Scope for catch-up in productivity and human capital in many countries.
- Machine learning (ML) in demand forecasting makes it possible to avoid traditional challenges associated with planning such as long delivery lead times, high transport costs, high inventory and waste levels, and incorrect decision making due to inaccurate forecasts.

11.BIBILOGRAPHY:-

Long Gang, "GDP Prediction by Support Vector Machine Trained with Genetic Algorithm," in 2nd International Conference on Signal Processing Systems (ICSPS), 2010.

[2] Jaehyun Yoon, "Forecasting of Real GDP Growth Using Machine Learning Models: Gradient Boosting and Random Forest Approach," Springer Science+Business Media, LLC, part of Springer Nature, 2020.

[3] John Roush, Keith Siopes, Gongzhu Hu, "Predicting Gross Domestic Product Using Autoregressive Models," in IEEE SERA, London, UK, June 7-9, 2017.

[4] Gary L. Shelley, Frederick H. Wallace, "Inflation, money, and real GDP in Mexico: a causality analysis," Applied Economics Letters, vol. 11, no. 4, p. 223–225, 2004.

[5] Martin Schneider, Martin Spitzer, "Forecasting Austrian GDP using the generalized dynamic factor model," 17 September 2004.

APPENDIX:

. COLAB NOTEBOOK:

<https://drive.google.com/file/d/1cSvqgZdf1gAyLWMT9VXznp5IO-Ts7iSJ/view?usp=sharing>

The screenshot shows a Google Colab notebook titled "GDP_Analysis-checkpoint.ipynb". The left sidebar displays the file explorer with folders "sample_data" and "world.csv". The main code area contains the following code:

```
import numpy as np # linear algebra
import pandas as pd # data processing
import seaborn as sns
from matplotlib import pyplot as plt

from sklearn.preprocessing import LabelEncoder
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_squared_error, mean_squared_log_error
```

Below the code, a "New Section" is created with the following code:

```
[2] data = pd.read_csv('/content/world.csv', decimal=',')
print('number of missing data:')
print(data.isnull().sum())
data.describe(include='all')
```

The notebook interface shows the RAM and Disk usage, and the status bar indicates "0s completed at 6:29 PM".

The screenshot shows the same Google Colab notebook, but the code has been executed. The output of the code is displayed in the cell:

```
number of missing data:
Country          0
Region           0
Population        0
Area (sq. mi.)   0
Pop. Density (per sq. mi.) 0
Coastline (coast/area ratio) 0
Net migration     3
Infant mortality (per 1000 births) 3
GDP ($ per capita) 1
Literacy (%)      18
Phones (per 1000) 4
Arable (%)        2
Crops (%)         2
Other (%)         2
Climate          22
Birthrate         3
Deathrate         4
Agriculture       15
Industry          16
Service           15
dtype: int64
```

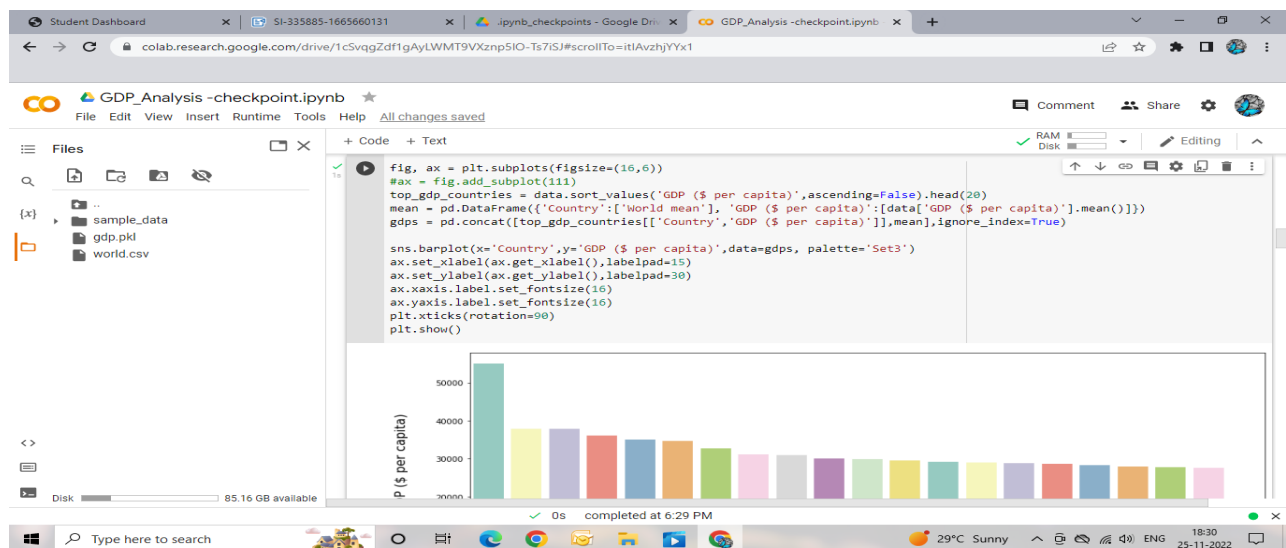
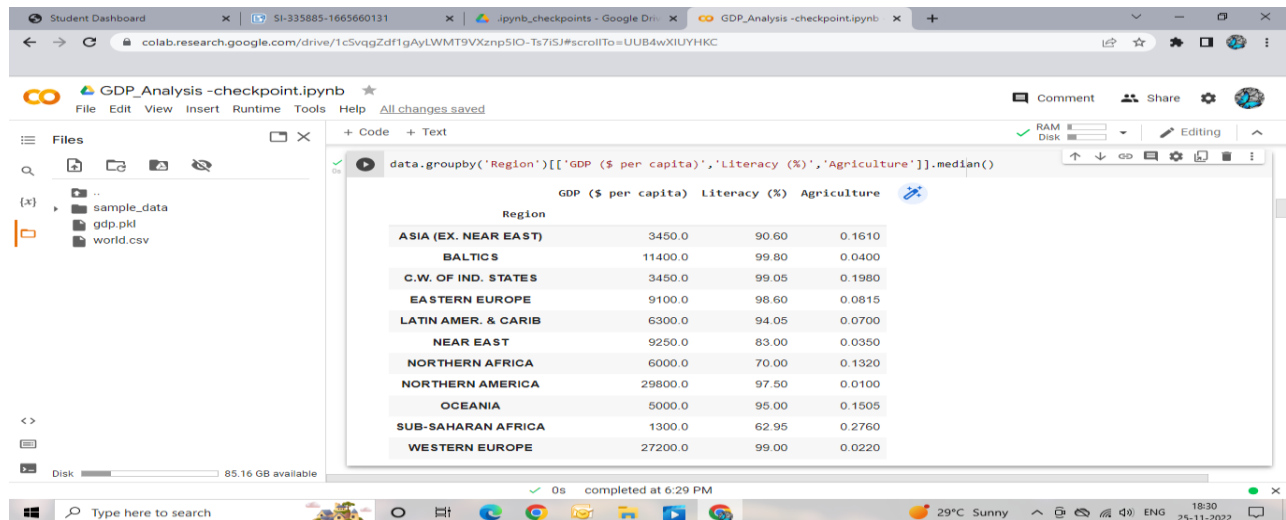
The notebook interface shows the RAM and Disk usage, and the status bar indicates "0s completed at 6:29 PM".

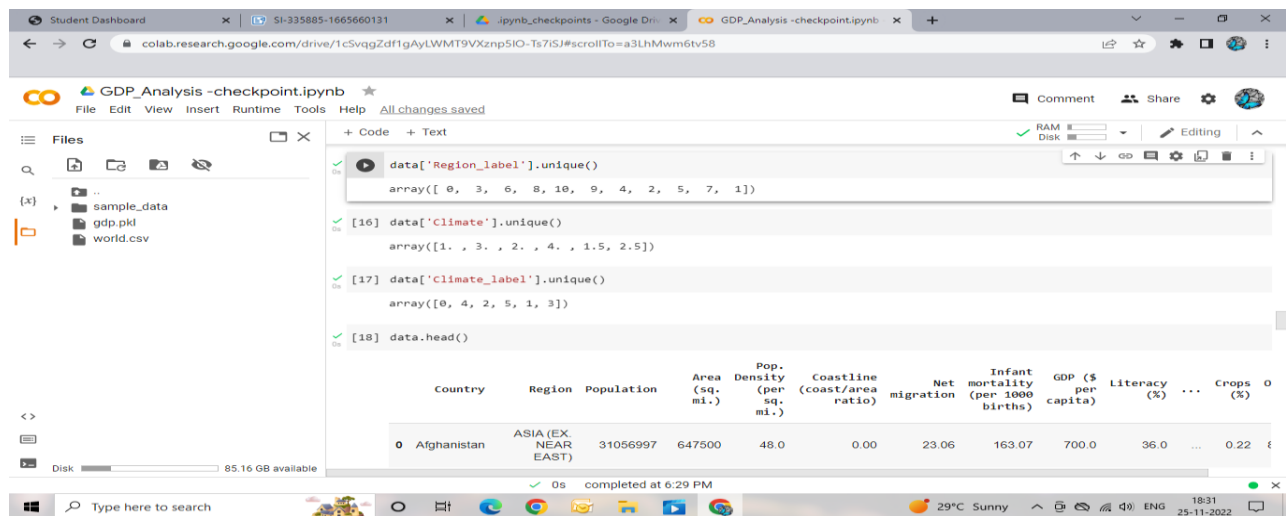
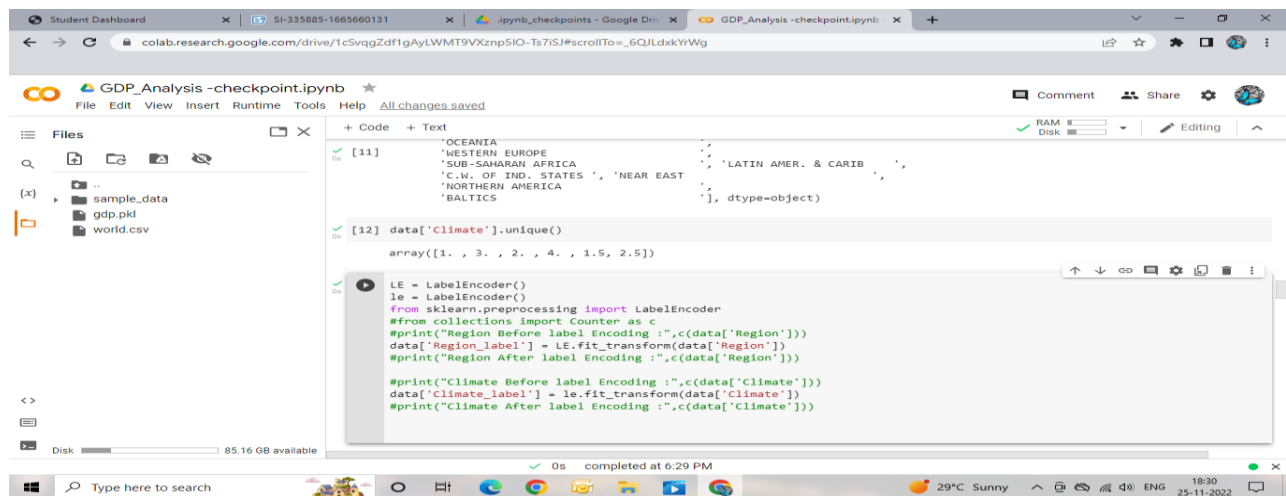
The screenshot shows the same Google Colab notebook, but the code has been executed. The output of the code is displayed in the cell:

```
data.head()
```

	Country	Region	Population	Area (sq. mi.)	Pop. Density (per sq. mi.)	Coastline (coast/area ratio)	Net migration	Infant mortality (per 1000 births)	GDP (\$ per capita)	Literacy (%)	Phones (per 1000)	Arabl (%)
0	Afghanistan	ASIA (EX-NEAR EAST)	31056997	647500	48.0	0.00	23.06	163.07	700.0	36.0	3.2	12.1
1	Albania	EASTERN EUROPE	3581655	28748	124.6	1.26	-4.93	21.52	4500.0	86.5	71.2	21.0
2	Algeria	NORTHERN AFRICA	32930091	2381740	13.8	0.04	-0.39	31.00	6000.0	70.0	78.1	3.2
3	American Samoa	OCEANIA	57794	199	290.4	58.29	-20.71	9.27	8000.0	97.0	259.5	10.0
4	Andorra	WESTERN EUROPE	71201	468	152.1	0.00	6.60	4.05	19000.0	100.0	497.2	2.2

The notebook interface shows the RAM and Disk usage, and the status bar indicates "0s completed at 6:29 PM".





Student Dashboard | SI-33585-166560131 | .ipynb_checkpoints - Google Drive | GDP_Analysis -checkpoint.ipynb

colab.research.google.com/drive/1c5vqgZdf1gAylWMT9VXznPSIO-Ts7SJ#scrollTo=SeCCLUR7awk-

GDP_Analysis -checkpoint.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Files: sample_data, gdp.pkl, world.csv

```
[21] from sklearn.metrics import r2_score
r2_score(test_Y,lr.predict(test_X))
0.5137096069689247

[22] lr.predict([[3581655,28748,124.6,1.26,-4.93,21.52,86.5,71.2,21.09,4.42,74.49,15.11,5.22,0.2320,0.188,0.579,3,4]])
/usr/local/lib/python3.7/dist-packages/sklearn/base.py:451: UserWarning: X does not have valid feature names, but LinearReg
array([[2569.57654235]])

rfr = RandomForestRegressor(n_estimators = 50,
                           max_depth = 6,
                           min_weight_fraction_leaf = 0.05,
                           max_features = 0.8,
                           random_state = 42)

rfr.fit(train_X, train_Y)
train_pred_Y = rfr.predict(train_X)
test_pred_Y = rfr.predict(test_X)
train_pred_Y = pd.Series(train_pred_Y.clip(0, train_pred_Y.max()), index=train_Y.index)
test_pred_Y = pd.Series(test_pred_Y.clip(0, test_pred_Y.max()), index=test_Y.index)

rmse_train = np.sqrt(mean_squared_error(train_pred_Y, train_Y))
```

0s completed at 6:29 PM

Student Dashboard | SI-33585-166560131 | .ipynb_checkpoints - Google Drive | GDP_Analysis -checkpoint.ipynb

colab.research.google.com/drive/1c5vqgZdf1gAylWMT9VXznPSIO-Ts7SJ#scrollTo=SeCCLUR7awk-

GDP_Analysis -checkpoint.ipynb

File Edit View Insert Runtime Tools Help All changes saved

Files: sample_data, gdp.pkl, world.csv

```
[24] r2_score(test_Y,rfr.predict(test_X))
0.7188195631832104

[25] rfr.predict([[3581655,28748,124.6,1.26,-4.93,21.52,86.5,71.2,21.09,4.42,74.49,15.11,5.22,0.2320,0.188,0.579,3,4]])
/usr/local/lib/python3.7/dist-packages/sklearn/base.py:451: UserWarning: X does not have valid feature names, but RandomFor
array([[3592.58180805]])

[26] import pickle
pickle.dump(rfr,open('gdp.pkl','wb'))

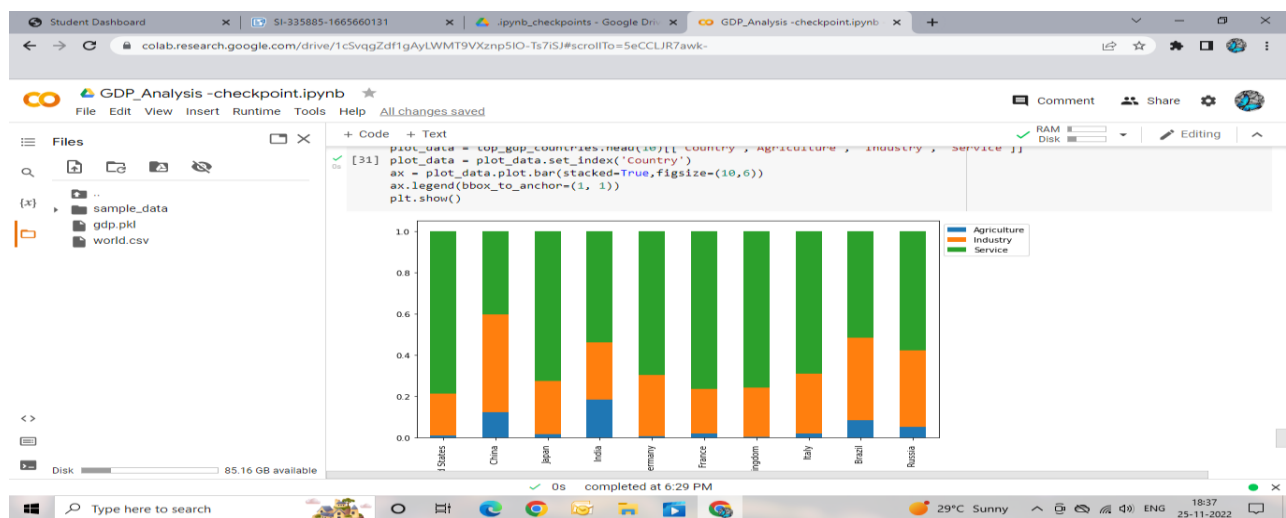
[27] plt.figure(figsize=(18,12))

train_test_Y = train_Y.append(test_Y)
train_test_pred_Y = train_pred_Y.append(test_pred_Y)

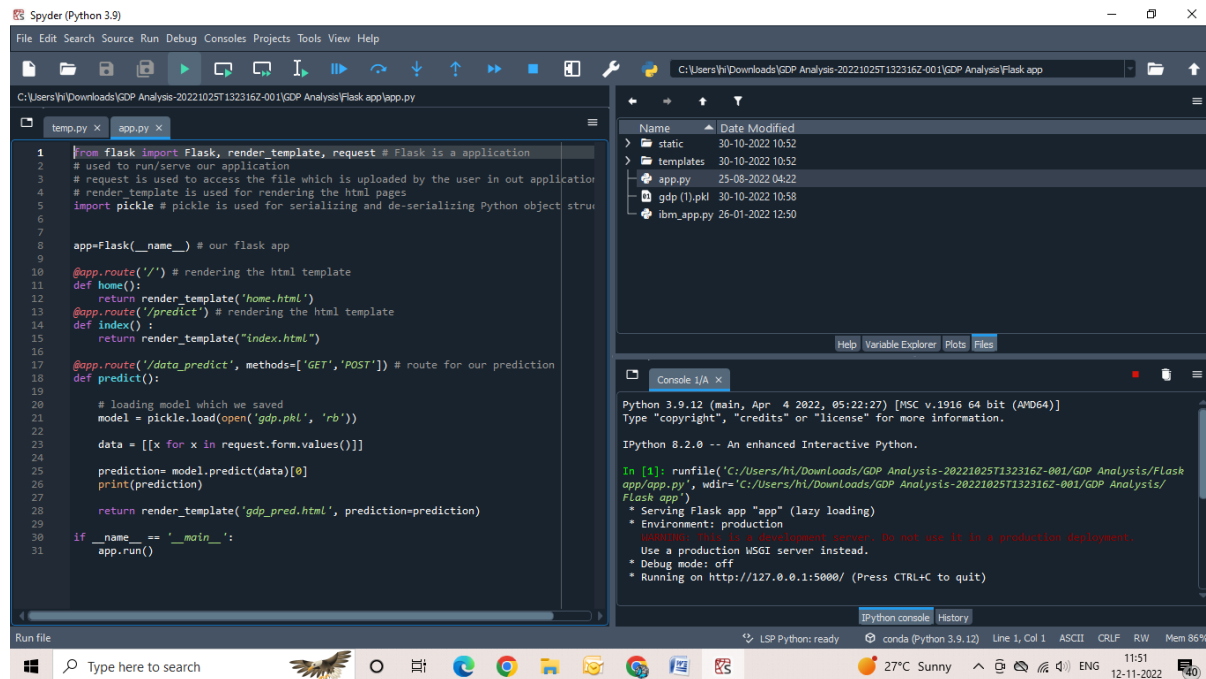
data_shuffled = data.loc[train_test_Y.index]
label = data_shuffled['Country']

colors = {'ASIA (EX. NEAR EAST)':'red',
          'EASTERN EUROPE':'orange'}
```

0s completed at 6:29 PM



FLASK CODE :



The screenshot shows the Spyder Python IDE interface. The main editor displays a Python script for a Flask application. The script imports Flask, render_template, request, and pickle. It defines a Flask app, routes for home and index, and a prediction route. The prediction route loads a model from a file named 'gdp.pkl' and makes a prediction based on request data. The console output shows the successful execution of the script and the serving of the Flask app.

```
1 from flask import Flask, render_template, request # Flask is a application
2 # used to run/serve our application
3 # request is used to access the file which is uploaded by the user in our application
4 # render_template is used for rendering the html pages
5 import pickle # pickle is used for serializing and de-serializing Python object structures
6
7
8 app=Flask(__name__) # our flask app
9
10 @app.route('/') # rendering the html template
11 def home():
12     return render_template('home.html')
13 @app.route('/predict') # rendering the html template
14 def index():
15     return render_template("index.html")
16
17 @app.route('/data_predict', methods=['GET','POST']) # route for our prediction
18 def predict():
19
20     # loading model which we saved
21     model = pickle.load(open('gdp.pkl', 'rb'))
22
23     data = [[x for x in request.form.values()]]
24
25     prediction= model.predict(data)[0]
26     print(prediction)
27
28     return render_template('gdp_pred.html', prediction=prediction)
29
30 if __name__ == '__main__':
31     app.run()
```

Python 3.9.12 (main, Apr 4 2022, 06:22:27) [MSC v.1916 64 bit (AMD64)]
Type "copyright", "credits" or "license()" for more information.

IPython 8.2.0 -- An enhanced Interactive Python.

In [1]: runfile('C:/Users/h1/Downloads/GDP Analysis-20221025T132316Z-001/GDP Analysis/Flask app/app.py', wdir='C:/Users/h1/Downloads/GDP Analysis-20221025T132316Z-001/GDP Analysis/Flask app')
* Serving Flask app "app" (lazy loading)
* Environment: production
 WARNING: This is a development server. Do not use it in a production deployment.
 Use a production WSGI server instead.
* Debug mode: off
* Running on http://127.0.0.1:5000/ (Press CTRL+C to quit)