Analysing the GDP disparities between North and South American countries and the factors contributing to these differences

Symon Islam
Friedrich-Alexander-Universitat Erlangen-Nurnberg
Erlangen, Germany
symon.islam@fau.de

***Abstract*— this research-based project emphasizes the primary factors of these economic discrepancies in South and North America, this study inspects the notable GDP difference between Brazil and the USA. Analyzing the factorial data for both countries found the advantages especially for the United States—to achieve huge economic output that are highlighted through this project. While Brazil's struggle limits their potential to thrive. Due to the huge collection of data by narrowing the divergence this study uses data from 1989-2023. Finally, this project produces the key factors for this study.**

Keywords—GDP, Preprocessing, Feature Engineering, Data Engineering

I. Introduction

Despite being part of America North and South American countries have huge disparities in GDP. In this project, I took the United States as the clear leader and representative of North America, on the other hand, Brazil is often considered as the dominant country of South America. Then. I aimed to investigate the GDP difference over the year 1989 to 2023 using metadata and data engineering approaches. While this research and analysis includes significant changes in GDP over the decades and also feature engineering for the factors finding which value most for both of the countries.

II. Primary Questions

A. *What are the primary economic, political, and social factors driving the GDP disparities between North and South American countries, and how do these factors impact growth potential across the two regions?*

B. *How does workforce composition, including employment in different economic sectors, contribute to GDP variations between the two regions?*

III. Research Scope

As long as the GDP data and the indicator data are associated from 1989 to 2023, which provided almost two decades of economic flow thus, the scope of research is broad. On the one hand, the global economic leader serves as a benchmark for the sophisticated economies and other hand the strongest economically positioned Brazil offers a challenging and full of opportunities market. By factor analysis, it is evident that a wide range of selections can be found contributing to GDP. That includes sectoral contributions and investment patterns which gives a convex output for both cases.

Research Scope		
Chronological Scope	***Geographic Focus***	***Factor Analysis***
Comparative Analysis, Economic trends and shifts over decades	Highlights two representative nations	Economic Factors, Workforce composition,

Table 1: Research Scope

IV. Data Sources

For datasets, I had to consider various data policies. The following data sets are used for implementing the whole project where most of the data is from the World Bank. However, relevant preprocessing is done by code execution.

GDPData

https://raw.githubusercontent.com/errorsymon/Data/d710147cfb374060422bd86a1889d33e54fa3f2b/API_NY.GDP.MKTP.CD_DS2_en_csv_v2_9865.csv

1. **Source:** This data comes with a wide range of attributes from the official sites of the World Bank. Which is mainly used for the yearly trends representation of the global economy.
2. **Description:** This dataset comes with real-time values from 1960-2023 and onward. Undoubtedly this dataset is valuable for finding the insight of economic research and study.
3. **Licenses:** The data is available under the World Bank's open data policy. The World Bank provides it free of charge for use of educational, research, and non-commercial uses. Though, Proper citation is required for any usage of the data. Open Database License (ODbL), Microdata Research License and the Creative Commons Attribution license (CC-BY 4.0) works for this Data set.

Metadata – Country

https://raw.githubusercontent.com/errorsymon/Data/d710147cfb374060422bd86a1889d33e54fa3f2b/Metadata_Country_API_NY.GDP.MKTP.CD_DS2_en_csv_v2_9865.csv

1. **Source**: This specific metadata also comes from the same database (World Bank) that contains group data for different attributes like income, and region.

2. **Description:** It gives a details classification on the basis of regions where different income levels are nominated by the gross income status. Interestingly I find country codes which gives me more flexibility when working with feature engineering. Additionally, this data set contains the attributes group data.

3. **License:** The data is provided by the World Bank under the same open data policy, meaning it is free for use with proper attribution with respect to educational or research use.

Indicator Data USA

https://raw.githubusercontent.com/errorsymon/Data/b906c2e79b027e4ad1df257277ede4b8f9884ff2/Metadata_Indicator_API_NY.GDP.MKTP.CD_DS2_en_csv_v2_9865.csv

1. **Source**: This indicator data set especially gives valuable insights into the USA gross income and investment in different areas. Also, this data is from the World Bank database.

2. **Description**: For feature engineering this data set also was crucial. It gives various indicators for the United States including social, economic and financial factors. Top of that this data set consists of records of key developments over two decades. However, an issue was found with the alignment. Further study was resolved through a sophisticated algorithm.

3. **Licenses**: Creative Commons Attribution license (CC-BY 4.0) which defines that anyone can use and modify that data for study and research purposes. However commercial use is forbidden.

Indicator Data Brazil

https://raw.githubusercontent.com/errorsymon/Data/refs/heads/main/brazil.csv

1. **Source**: I took Brazil indicator data from the similar data sources to avoid the data formatting hassle. Indeed both of the data sources gives me positiveness to work with different columns and rows as they are from similar sources.

2. **Description*:*** Just like the other three data sets the file type for this data set is csv. While it's closely similar with the USA indicator data set. However, the attributes and the unique code makes it different.

3. **Licenses:** The World Bank Group licenses datasets under the Creative Commons Attribution 4.0 International license (CC-BY 4.0) and makes data publicly accessible in accordance with open data guidelines. Numerous datasets are accessible under other licenses. Users who access them consent to abide by all of the terms of the corresponding licenses, which are described below, and they are appropriately labeled.

A. DATA PIPELINE

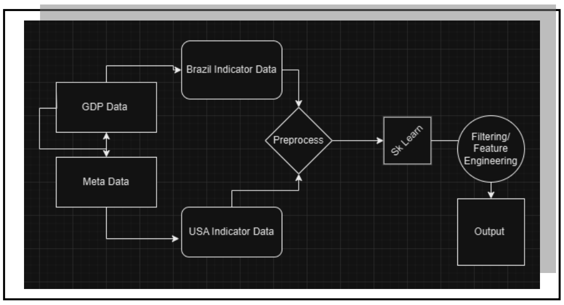


Fig. 1. Pipeline Architecture

For this specific project work, I built a relevant, robust and fully automated pipeline by following the instructions given by the course teacher. Therefore, it allows the three valuable terms of standardization called ETL that can allow the user to download all the datasets and give an overview of the output data.

To make it possible legible format different data transformations were done over the whole project to address the potential expected value. When any redundant data is found to eradicate it appears namely deleting inserting and interpolation strategies used. As you can see above the pipeline consists of different stages namely., Preprocess, Sklearn, feature engineering.

B. Preprocessing

In this study, multiple preprocessing steps are employed to clean and prepare the data with the required format. The `load_datasets ()` function loads the GDP data and metadata from CSV files, while `preprocess_gdp_data()` filters relevant columns and reshapes the data as long as we need column-based. The `preprocess_usa_brazil_data()` function cleaned the data for the USA and Brazil by removing unnecessary rows. `preprocess_country_metadata ()` extracts key metadata columns such as "Region" and "IncomeGroup." The `merge_data()` function merges the cleaned GDP data with country metadata. Missing values are imputed using `SimpleImputer` for numeric and categorical data. Label encoding is applied to categorical columns using `LabelEncoder`. The `calculate_gdp_growth_rate()` function calculates the GDP growth rate using percentage change. Finally, the data is filtered for the countries of interest (USA and Brazil) and the years 1989 to 2023 using string matching and the `between()` method.

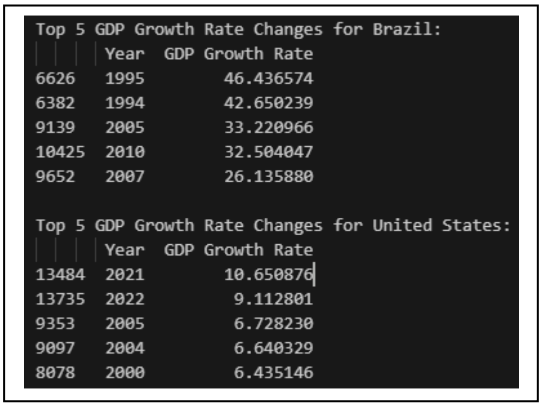
C. Feature Engineering

The code's feature engineering tasks include utilizing LabelEncoder to transform categorical variables into numeric ones for columns such as "Region" and "IncomeGroup." SimpleImputer handles missing values by employing the most common approach for categorical data and the mean for numerical data. The calculate_gdp_growth_rate() function, which determines the percentage change in GDP over time, is also used to calculate the GDP growth rate. Lastly, to focus the analysis, the data is filtered by years (1989–2023) and countries (USA, Brazil).

V. Result and Limitations

A. *Major Chnages of GDP value over the year 1989-2023*

To ensure that only pertinent data is left, the code first filters the GDP data to extract entries pertaining to Brazil. It then determines whether the filtered data includes accurate GDP figures for Brazil. The code then applies a percentage change to the GDP variables to determine the GDP growth rate. The top five years with the biggest changes in Brazil's GDP growth rate are then chosen. The results, which display the top five years with the biggest rises in GDP growth rate, are printed out at the end.



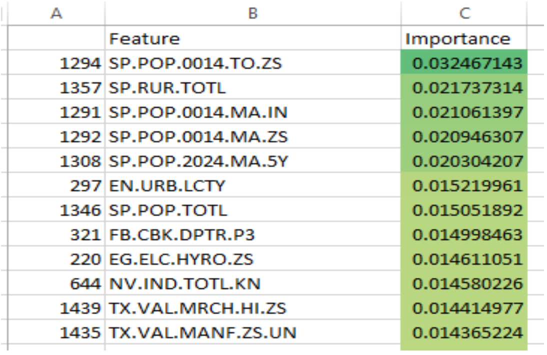
```
Top 5 GDP Growth Rate Changes for Brazil:
       Year  GDP Growth Rate
6626   1995        46.436574
6382   1994        42.650239
9139   2005        33.220966
10425  2010        32.504047
9652   2007        26.135880

Top 5 GDP Growth Rate Changes for United States:
       Year  GDP Growth Rate
13484  2021        10.650876
13735  2022         9.112801
9353   2005         6.728230
9097   2004         6.640329
8078   2000         6.435146
```

Fig. 2. Major changes in GDP over year

a) Limitation to data selection: Due to missing data on 3rd and 4th data set I had to eradicate and work with the data from 1989-2023 where in the meta data and the gdp data data can be found from earlier years. For r

B. *Feature detection for USA GDP contributing most*

Uses machine learning (Random Forest Regressor) to identify key factors influencing GDP growth. We found important features as namely Population ages 0-14 (% of total population), Rural population, Population in the largest city (%) of urban population. Those are the major factor of USA GDP changes.



A	B	C
	Feature	Importance
1294	SP.POP.0014.TO.ZS	0.032467143
1357	SP.RUR.TOTL	0.021737314
1291	SP.POP.0014.MA.IN	0.021061397
1292	SP.POP.0014.MA.ZS	0.020946307
1308	SP.POP.2024.MA.5Y	0.020304207
297	EN.URB.LCTY	0.015219961
1346	SP.POP.TOTL	0.015051892
321	FB.CBK.DPTR.P3	0.014998463
220	EG.ELC.HYRO.ZS	0.014611051
644	NV.IND.TOTL.KN	0.014580226
1439	TX.VAL.MRCH.HI.ZS	0.014414977
1435	TX.VAL.MANF.ZS.UN	0.014365224

Fig. 3. Important Feature

C. *Further study*

Build interactive dashboards for better insights using tools like Power BI or Tableau useful for policy makers.

Challenges

In this venture, a few challenges developed, especially around information quality and accessibility. Guaranteeing the datasets were stacked accurately and were clean, with lost or twisted values suitably dealt with, was a major jump. The blending of different dataset GDP information, nation metadata, and marker data was complex, as nation names and codes did not continuously adjust impeccably. Taking care of lost values in both numeric and categorical columns required successful ascription procedures. Changing the information (e.g., dissolving and rotating) whereas guaranteeing appropriate organizing, as well as encoding categorical highlights for machine learning, included to the trouble. At long last, sending out the information to designs like SQLite and CSV and approving the incorporation of both Brazil and the USA required cautious investigating and consideration to information judgment.

Additionally, I used try function for error handling. When working with external data sources, such as CSV files from URLs, various errors can occur—e.g., the file might not exist, the URL could be invalid, or the file format might not match expectations.