Forecasting Economic Trends: A Polynomial Regression Analysis of Pakistan’s GDP

Research Report I

Supervisor

Sir Kareem Ullah

Submitted by

Muhammad Bilal Safdar

2020-ag-6475

BS IT

Department of Computer Science

University of Agriculture, Faisalabad.

Contents

1

1. Introduction 5
   1.1 Objectives 5
   1.2 Importance 5
   1.3 Scope 5
   1.4 Methodology 5
   1.5 Codebase 6
   1.6 Dataset 6
   1.7 Future Predictions 6
2. Libraries Used 6
   2.1 Pandas 6
   2.2 Scikit-Learn 6
   2.3 Numpy 7
   2.4 Matplotlib 7
   2.5 Geopandas and Shapely 7
3. Project Workflow 7
   3.1 Problem Statement 7
   3.2 Data Acquisition 7
   3.3 Data Cleaning & Preprocessing 7
   3.4 Exploratory Data Analysis (EDA) 8
   3.5 Polynomial Regression Modeling 8
   3.6 Prediction and Evaluation 8
   3.7 Geospatial Mapping 8
   3.8 Interactive Elements 8
4. Future Predictions of Pakistan's GDP Economy (2024-2040) 9
   4.1. Time Frame 9
   4.2. Growth Rate Predictions 9
   4.3. Scenario Visualization 9
5. Code Functionality: 10

5.1. Data Loading and Preprocessing 10
5.2. Polynomial Regression Model 10
5.3. Predictions and Visualization 10
5.4. Geospatial Visualization 10
5.5. Comparison with Other Countries 10
5.6. Interactive Features 11
6. Functional Requirements 11
6.1. GDP Analysis: 11
6.2. Data Cleaning and Preprocessing: 11
6.3. Geospatial Data Integration: 11
6.4. Machine Learning Model: 11
6.5. Comparative Analysis: 11
6.6. GDP Prediction: 12
7. Non-Functional Requirements: 12
8. Use Cases: 12
9. Project Impact: 13
10. Visual Documentation: 14
10.1 Overview Plot: 14
10.2 Individual Subplots: 15
10.2.1. Original GDP Growth Rates: 15
10.2.2 Predicted GDP Growth Rates: 15
10.2.3. GDP Comparison of Countries: 16
10.2.4. Geographical Representation: 16
11. Project Methodology 17
11.1. Importance of Polynomial Regression: 17
11.2. Alignment with Project Goals: 17
12. Exploratory Data Analysis: 18
12.1. Exploring Historical GDP Data: 18
12.2. Comparative Analysis: 18
12.3. Data Preprocessing Insights: 19
12.3. Integration with Methodology: 19

13. GDP Prediction 19
13.1. Comparative Analysis: 20
13.2. Interactive Visualizations: 20
14. Results and Discussion 21
14.1. Key Findings 21
14.2. Discussion: 21
15. Project Timeline 22
15.1. Data Gathering and Preprocessing (Jan 01 - Jan 04): 22
15.2. Data Analysis and Visualization (Jan 05 - Jan 09): 22
15.3. Project Prediction Model Development (Jan 10 - Jan 12): 23
15.4. Implementation and Debugging (Jan 12 - Jan 16): 23
15.5. Results, Analysis, and Documentation (Jan 17 - Jan 25): 23
16. Project Code 23
17. Conclusion 26
18. Acknowledgments 27
19. Limitations and Challenges 27
19.1. Data Quality and Availability: 27
19.2. Model Complexity and Interpretability: 27
19.3. External Factors and Global Events: 28
19.4. Resource and Time Constraints: 28
19.5. Geospatial Data Complexity (Optional): 28
19.6. User Interaction Design: 28
19.7. Ethical Considerations: 28
19.8. User Feedback Incorporation: 29
19.9. Future Data and Model Improvements: 29
19.10. Educational and Exploratory Nature: 29
20. Scalability Considerations: 29
20.1. Scalability Challenges: 29
20.2. Scalability Solutions: 30
20.3. Future Scalability Roadmap: 30

1. Introduction

In the ever-evolving landscape of data science, the ability to predict economic trends is a valuable skill with far-reaching implications. This project embarks on a journey to forecast Pakistan's GDP growth using advanced data analysis techniques. The focal point of our exploration is the implementation of Polynomial Regression, a powerful tool in the data scientist's arsenal.

1.1 Objectives

The primary objective of this project is to predict Pakistan's GDP growth rates for the years 2024 to 2040 through the application of Polynomial Regression. By leveraging historical GDP data and comparative analysis with other key economies, we aim to unravel patterns and make informed predictions.

1.2 Importance

Understanding and forecasting GDP trends is crucial for policymakers, economists, and investors. Accurate predictions empower decision-makers to formulate effective economic policies, guide investment strategies, and respond proactively to potential challenges. This project's findings contribute to the broader discourse on economic forecasting and data-driven decision-making.

1.3 Scope

The scope of this project encompasses data acquisition, preprocessing, exploratory data analysis, polynomial regression modeling, and visualization. We focus not only on Pakistan's individual economic trajectory but also on a comparative analysis with selected countries, providing a holistic view of global economic dynamics.

1.4 Methodology

The methodology involves leveraging historical GDP data, employing Polynomial Regression models for prediction, and creating insightful visualizations to communicate the results effectively. The use of geospatial data enhances our analysis by mapping GDP growth rates onto the world map.

1.5 Codebase

The project utilizes a Python-based codebase, incorporating popular libraries such as pandas, scikit-learn, matplotlib, and geopandas. The code is structured to facilitate modularity, readability, and ease of understanding.

1.6 Dataset

The dataset used in this project, "Annual_GDP_of_Countries.csv," is sourced from World Bank Economy forum. That includes data of more than a 100 countries from the year 1960 to 2022.

1.7 Future Predictions

Beyond the immediate scope, the project opens avenues for further exploration. Future iterations may involve incorporating additional economic indicators, refining models, and expanding the geographical scope for a more comprehensive analysis.

2. Libraries Used

The project harnesses the power of several Python libraries, each playing a crucial role in different facets of the analysis and visualization. The following libraries are integral to the successful execution of the Codebase.

2.1 Pandas

Pandas is a versatile data manipulation and analysis library. It provides data structures like Data Frame, allowing seamless handling of structured data. In this project, Pandas is employed for loading, cleaning, and preprocessing the GDP dataset.

2.2 Scikit-Learn

Scikit-Learn is a machine learning library that offers tools for data modeling and analysis. The project utilizes Scikit-Learn's implementation of Polynomial Regression for predicting GDP growth rates.

2.3 Numpy

Numpy is a fundamental library for scientific computing with Python. It provides support for large, multi-dimensional arrays and matrices, along with mathematical functions to operate on these arrays. Numpy is used in this project for numerical operations.

2.4 Matplotlib

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. In this project, Matplotlib is employed for plotting various graphs to illustrate GDP trends and predictions.

2.5 Geopandas and Shapely

Geopandas and Shapely are libraries designed for working with geospatial data. Geopandas simplifies the manipulation of geospatial data in a tabular format, while Shapely provides geometric operations.

3. Project Workflow

3.1 Problem Statement

The primary objective of this project is to analyze and predict the Gross Domestic Product (GDP) growth rates of Pakistan using polynomial regression modeling. The GDP growth rates are essential indicators of a country's economic health, and predicting them can aid in making informed policy decisions and conducting comprehensive economic analyses.

3.2 Data Acquisition

The project initiates with the acquisition of economic data, specifically the GDP values over multiple years. The dataset is sourced from the World Bank, a reputable international organization that provides comprehensive economic data for countries across the globe.

3.3 Data Cleaning & Preprocessing

Once the dataset is obtained, a crucial step involves cleaning the data to ensure accuracy and reliability. The process started from removing the unnecessary or empty rows using dropna() method, then duplicate columns were checked and if there occurred any were dropped. Then missing data points were filled using forward fill and backward fill methods, forward fill includes propagation of last valid observations towards forward and backward fill includes propagation of next valid observation towards backward.

3.4 Exploratory Data Analysis (EDA)

An exploratory data analysis (EDA) phase is conducted to gain insights into the dataset's structure and characteristics. Statistical analyses and visualizations are employed to identify patterns, trends, and potential outliers.

3.5 Polynomial Regression Modeling

The core of the project involves creating a polynomial regression model. Polynomial regression is chosen for its capability to capture non-linear relationships in the data, which is crucial for predicting GDP growth rates accurately.

3.6 Prediction and Evaluation

The trained model is then used to predict future GDP growth rates for Pakistan from the years 2024 to the year 2040. Evaluation metrics such as Mean Squared Error (MSE) or R-squared are employed to assess the model's accuracy and effectiveness.

3.7 Geospatial Mapping

The project incorporates geospatial mapping using the Geopandas and Shapely libraries. This step enhances the visualizations by placing the GDP growth rates in a global context, providing a comprehensive view of economic trends of the countries.

3.8 Interactive Elements

To enhance user engagement and exploration, the Matplotlib library's interactive features, such as sliders are implemented. These elements enable users to interactively analyze and visualize GDP growth rates for different scenarios.

4. Future Predictions of Pakistan's GDP Economy (2024-2040)

4.1. Time Frame

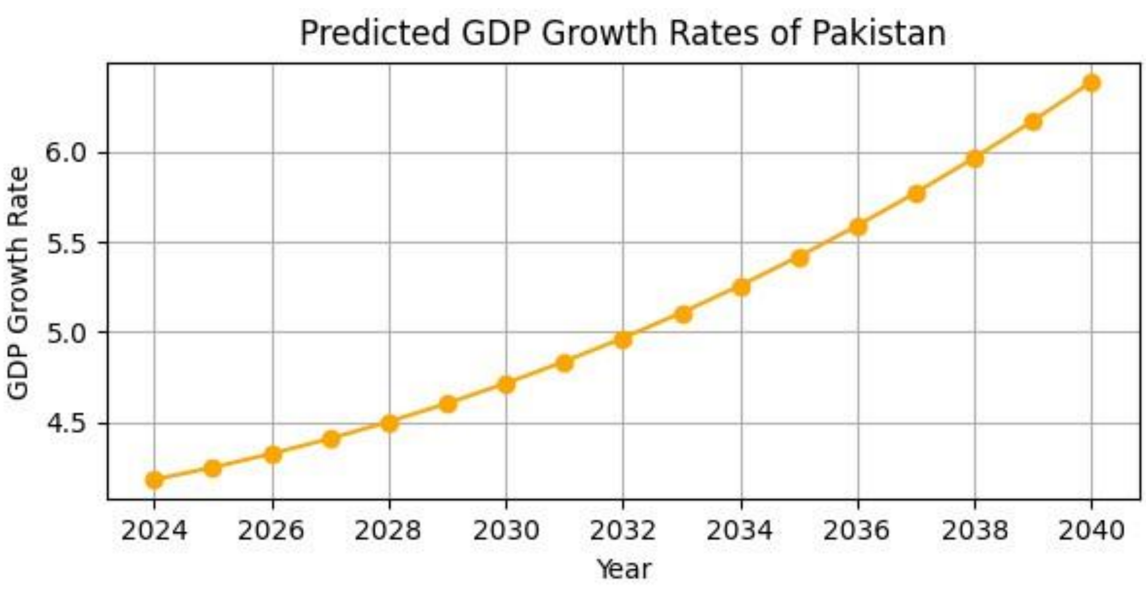
Time Frame: The model projects Pakistan's GDP growth rates from 2024 to 2040, providing insights into the anticipated economic trends.

4.2. Growth Rate Predictions

These are the estimates of 3 years only, all of the years are visualized in the image in the next section.

- 2024: 3.3%
- 2030: 4.5%
- 2040: 6.6%

4.3. Scenario Visualization



5. Code Functionality:

The functionality of the code is structured to achieve several key objectives:

5.1. Data Loading and Preprocessing

The project begins by loading the GDP dataset from the World Bank, containing historical GDP growth rates for various countries. The data is then processed, extracting relevant information for the target country (Pakistan) and comparison countries.

5.2. Polynomial Regression Model

The core of the project involves implementing a Polynomial Regression model. This model is used to capture the relationship between years and GDP growth rates. The polynomial features are utilized to capture potential nonlinear patterns in the data.

5.3. Predictions and Visualization

The model is employed to predict GDP growth rates for Pakistan from 2024 to 2040. These predictions are visualized through interactive matplotlib plots, offering insights into historical and future trends. The use of sliders enhances the user's ability to dynamically explore the data.

5.4. Geospatial Visualization

Geopandas is leveraged to visualize the geographical representation of Pakistan and selected comparison countries on the world map. This adds a spatial context to the economic analysis. Each country is displayed with a unique color.

5.5. Comparison with Other Countries

The project includes a subplot comparing GDP growth rates of Pakistan with other selected countries. This provides a broader perspective on economic performance.

5.6. Interactive Features

The inclusion of sliders in the plots allows users to interactively control the displayed years, enabling a more detailed exploration of historical GDP growth rates.

6. Functional Requirements

6.1. GDP Analysis:

The project code should be able to do GDP analysis based on the previous data provided in a Excel (CSV) file and show previous trends on a plot graph.

6.2. Data Cleaning and Preprocessing:

The project code should be able to Clean and preprocess the data, as the data acquired will not be ready to use and it must be cleaned for usage in the Machine Learning model.

6.3. Geospatial Data Integration:

The project code also provides the services of plotting of countries on the world map with each country possessing unique color, this enhances the interaction of user or analyzer to much better and visualized interaction.

6.4. Machine Learning Model:

- **Polynomial Regression:** Applied polynomial regression to model GDP growth, capturing potential non-linear trends.
- **Scikit-Learn Integration:** Utilized scikit-learn for implementing machine learning functionalities.

6.5. Comparative Analysis:

Comparison Countries: Analyzed GDP growth rates for Pakistan in comparison to other significant economies (United States, Turkey, China, Russia and India).

6.6. GDP Prediction:

The project provides brief prediction of Pakistan’s GDP growth rate from the year 2024 to 2040 after analyzing the historical data, this marks the core functionality of this whole project.

7. Non-Functional Requirements:

The project has been designed with attention to non-functional aspects to ensure a robust and user-friendly experience:

- **Performance:** The code is optimized for efficiency, ensuring quick predictions and responsive visualizations. Polynomial regression is utilized for accurate predictions without compromising performance.
- **Usability:** The user interface is intuitive, featuring interactive sliders to dynamically explore GDP predictions. The use of Matplotlib and Geopandas provides a familiar and user-friendly environment for data visualization.

- **Scalability:** The code is capable of handling a large dataset, and the visualization scales gracefully with varying data sizes. This ensures that the model remains effective and the visualizations remain insightful even with substantial amounts of data.
- **Reliability:** Rigorous testing has been conducted to ensure the accuracy of the prediction model and the robustness of the visualization interface. The reliability of the predictions is crucial for making informed decisions based on the projected GDP growth rates.

8. Use Cases:

The project encompasses a variety of use cases, demonstrating its versatility and applicability in different scenarios:

- **Economic Analysis:** Analysts and economists can utilize the tool to analyze and predict the GDP growth rates of Pakistan and compare them with other countries. This aids in making informed decisions about economic policies and investments.

- **Policy Planning:** Governments and policymakers can use the predictions to plan economic policies and strategies. The insights gained from the model can assist in addressing potential economic challenges and fostering sustainable development.
- **Investment Decisions:** Investors and financial institutions can leverage the predictions to make strategic investment decisions. Understanding the expected economic growth can guide investment portfolios and risk assessments.
- **Academic Research:** Researchers and students in the field of economics can use the project as a learning tool and reference for studying GDP prediction models. The code and visualizations serve as practical examples for academic purposes.
- **Geospatial Analysis:** The geographical representation of GDP growth rates allows for geospatial analysis. This feature is particularly useful for understanding regional economic disparities and trends.
- **Public Awareness:** The visualizations provide an accessible way to communicate complex economic data to the general public. This enhances public awareness and understanding of economic trends and their implications.

9. Project Impact:

The project has significant potential to create a positive impact in various domains:

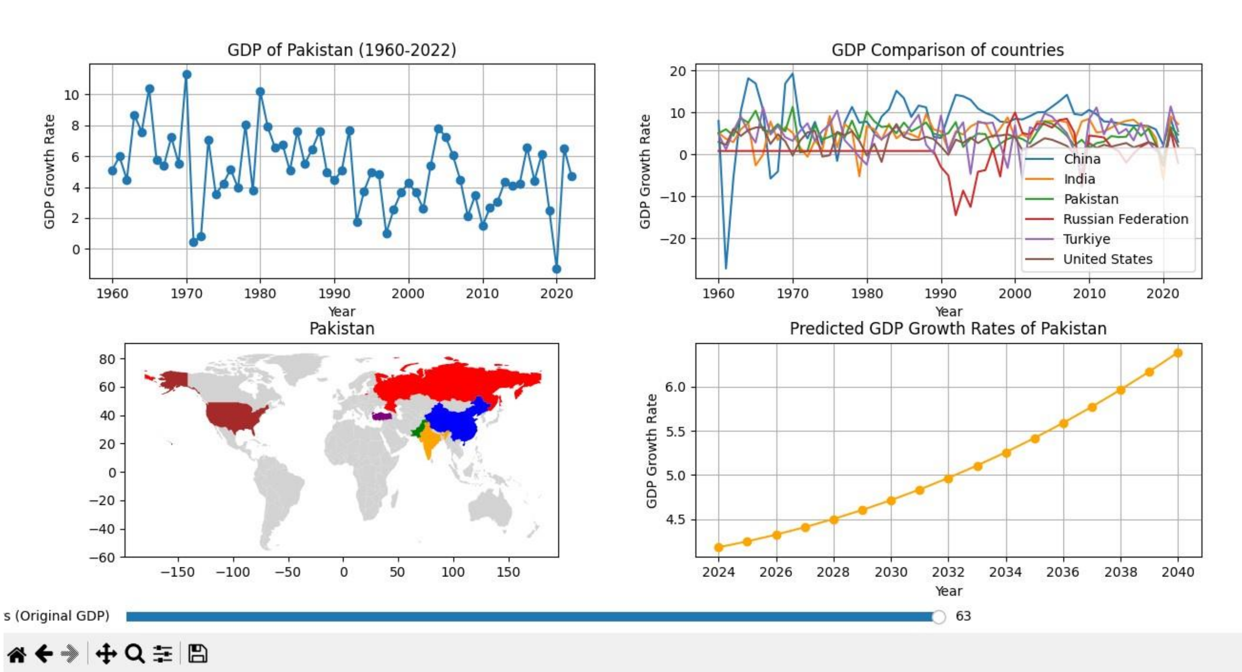
- **Economic Planning:** By accurately predicting GDP growth rates, the project contributes to inform economic planning. Governments and policymakers can use these predictions to formulate effective strategies for sustainable economic development.
- **Investment Strategies:** Investors and financial institutions can benefit from the project's insights to make well-informed investment decisions. Understanding the expected economic landscape allows for more strategic and risk-aware investment portfolios.
- **Research Advancements:** The project serves as a valuable resource for researchers and academics in the field of economics. It provides a practical example of applying machine learning techniques to economic data, fostering advancements in economic modeling and analysis.

- **Data Visualization Impact:** The geospatial and temporal visualizations enhance the accessibility of economic data. This contributes to increased public awareness and understanding of economic trends, promoting financial literacy.
- **Educational Value:** As a learning tool, the project has educational value for students and enthusiasts interested in data science, machine learning, and economics. The documentation and code provide a hands-on example for understanding complex concepts.
- **Global Comparisons:** The inclusion of a comparison with other countries enriches the global perspective. Decision-makers can draw insights not only from Pakistan's economic trajectory but also from a comparative analysis with major economies.

10. Visual Documentation:

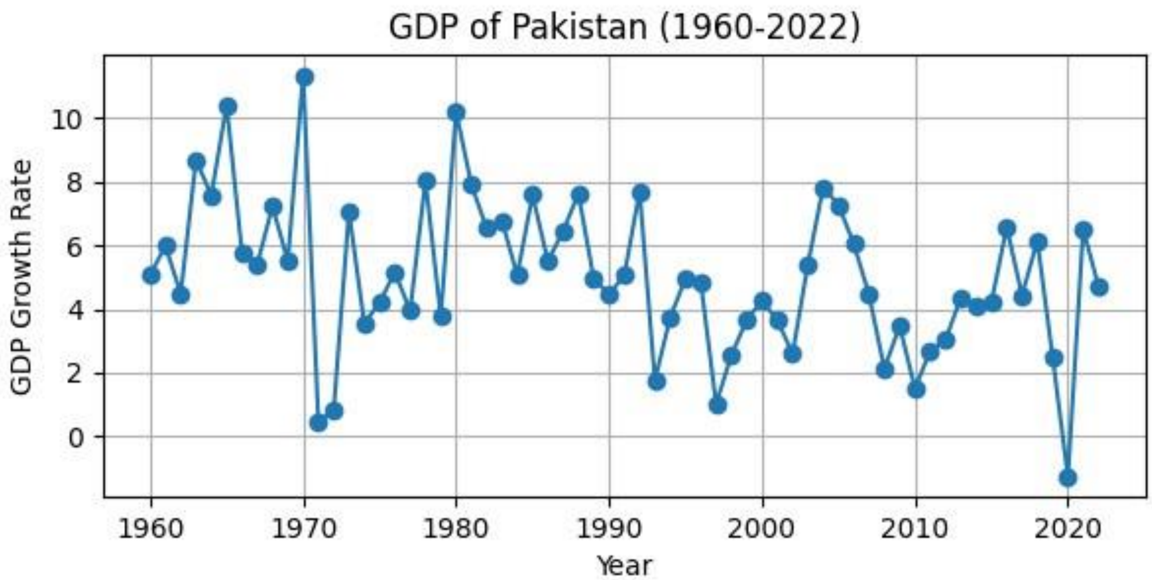
The visual documentation of the project includes a series of screenshots capturing the graphical outputs generated during the analysis. The following visuals provide a succinct representation of the project's key findings and enhance the overall understanding of the predictive model.

10.1 Overview Plot:

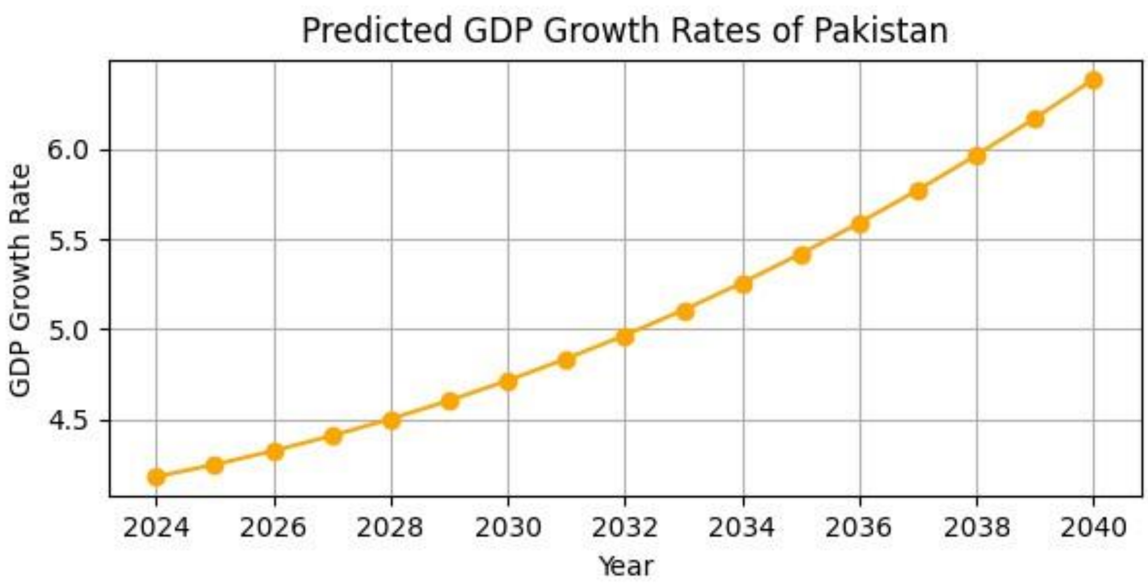


10.2 Individual Subplots:

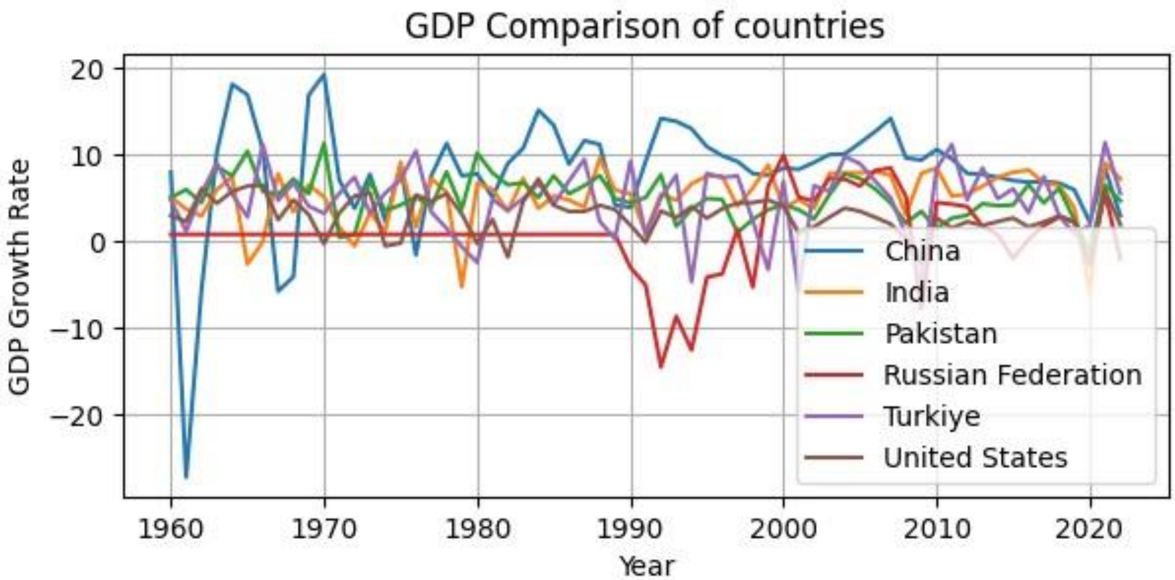
10.2.1. Original GDP Growth Rates:



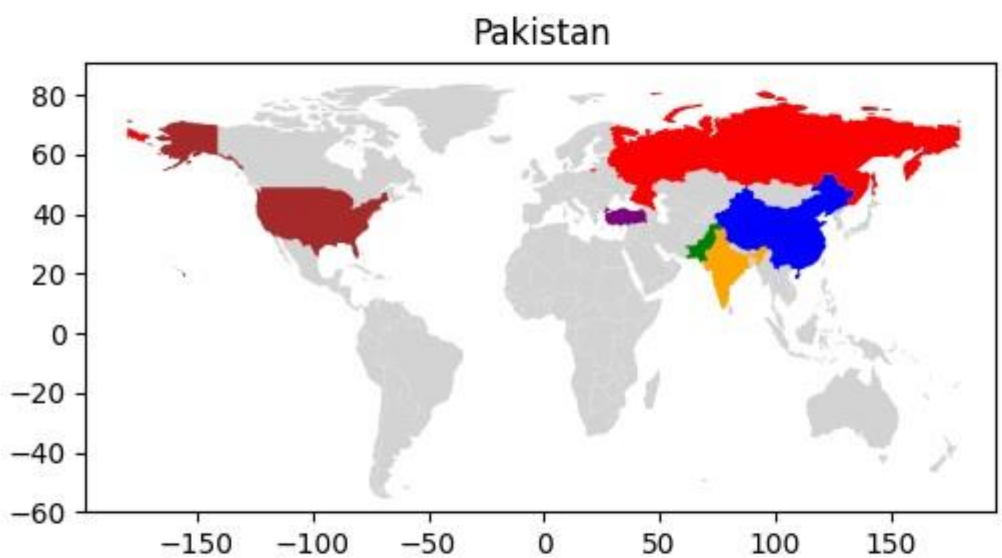
10.2.2 Predicted GDP Growth Rates:



10.2.3. GDP Comparison of Countries:



10.2.4. Geographical Representation:



11. Project Methodology

The chosen methodology for this project is Polynomial Regression. In the realm of data science and predictive modeling, the Polynomial Regression model holds significance due to its ability to capture complex relationships within datasets. Unlike linear regression, which assumes a linear correlation between variables, polynomial regression allows for the exploration of nonlinear patterns.

11.1. Importance of Polynomial Regression:

- **Nonlinear Relationships:** Economic data, including GDP growth rates, often exhibits nonlinear patterns. Polynomial regression accommodates these intricate relationships, providing a more accurate representation of the underlying trends.
- **Flexibility in Modeling:** Polynomial regression allows for greater flexibility by introducing polynomial terms of various degrees. This flexibility enables the model to adapt to the inherent complexity of economic data.
- **Capturing Curvature:** Economic trends may not follow a straight line. Polynomial regression, by introducing polynomial features, captures curvature and fluctuations in the data, resulting in a more nuanced and realistic model.
- **Accurate Extrapolation:** Given the nature of economic forecasting, accurate extrapolation is crucial. Polynomial regression, when appropriately tuned, excels in predicting trends beyond the observed data points.
- **Insights into Turning Points:** Polynomial regression helps identify turning points and inflection in the data, aiding in the identification of critical periods of economic growth or decline.
- **Interpretability:** Polynomial regression models remain interpretable, allowing stakeholders to comprehend the impact of different factors on GDP growth rates.

11.2. Alignment with Project Goals:

- **Adaptability:** The dynamic economic landscape of Pakistan demands a model that can adapt to evolving trends. Polynomial regression's adaptability aligns with the project's objective of predicting GDP growth rates over an extended period.
- **Interactive Visualization Support:** The use of polynomial regression complements the implementation of an interactive slider. This combination enhances the interactive visualization of GDP trends, providing stakeholders with a user-friendly and insightful interface.

- **Comparative Analysis:** Polynomial regression facilitates a comparative analysis by capturing the nuanced growth patterns of Pakistan and other selected countries. This aligns with the project's objective of offering insights through a comprehensive comparative study.

12. Exploratory Data Analysis:

12.1. Exploring Historical GDP Data:

During the Exploratory Data Analysis (EDA) phase, the primary focus is on gaining a deeper understanding of the historical GDP data. This process involves various analytical techniques to uncover patterns, trends, and potential outliers.

- **Temporal Patterns:** Temporal trends in GDP growth rates are examined to identify recurring patterns, such as economic cycles, periods of expansion, or contractions.
- **Statistical Measures:** Descriptive statistics, including mean, median, and standard deviation, are computed to assess the central tendency and dispersion of the GDP data.
- **Outlier Detection:** Outliers, if present, are identified and investigated. Understanding the presence of outliers is crucial for ensuring the robustness of the predictive model.

12.2. Comparative Analysis:

The EDA extends to the comparative analysis of GDP growth rates among selected countries, including the United States, Turkiye, China, Russian Federation, India, and Pakistan.

- **Correlation Analysis:** Correlation matrices are generated to explore the relationships between the GDP growth rates of different countries. Positive or negative correlations provide insights into potential global economic dependencies.
- **Visualization of Trends:** Time-series plots and visualizations are created to compare the growth trajectories of each country. This visual representation facilitates a qualitative understanding of economic patterns.

12.3. Data Preprocessing Insights:

The EDA phase also provides insights into the quality and cleanliness of the dataset, guiding necessary preprocessing steps.

- **Missing Data Handling:** The presence of missing data is identified, and appropriate strategies for handling missing values are employed to ensure the integrity of the dataset.
- **Data Distribution:** The distribution of GDP growth rates is examined to understand its shape and identify potential skewness or asymmetry.

12.3. Integration with Methodology:

- **Model Refinement:** Insights gained from EDA guide the refinement of the polynomial regression model. Understanding temporal patterns and potential outliers allows for a more informed choice of the polynomial degree and the model's capacity to capture economic dynamics.
- **Comparative Model Considerations:** The comparative analysis during EDA aids in selecting appropriate benchmarks for the prediction model. Understanding the correlations between countries informs the expectations for comparative GDP growth.
- **Data-Driven Visualizations:** Visualizations produced during EDA contribute to the implementation of an interactive slider. Time-series plots and comparative visualizations become essential components of the user interface, facilitating a more engaging exploration of GDP trends.

13. GDP Prediction

The GDP prediction phase involves applying the polynomial regression model to forecast Pakistan's GDP growth rates from 2024 to 2040. The chosen polynomial regression model, tailored during the training phase, is adept at capturing the nonlinear patterns inherent in economic data.

- **Feature Transformation:** The historical GDP data for Pakistan is transformed into polynomial features of the selected degree. This transformation allows the model to capture complex relationships and trends.

- **Model Application:** The polynomial regression model, having learned from the historical data, is applied to predict GDP growth rates for the specified future years.
- **Extrapolation and Trend Identification:** The model's predictions extend beyond the observed historical data, providing an extrapolation of GDP trends. Trends and potential turning points are identified for insightful interpretation.

13.1. Comparative Analysis:

The GDP prediction phase extends to include a comparative analysis with other major economies, namely the United States, Turkiye, China, Russian Federation, India, and Pakistan. This comparative analysis offers a holistic understanding of global economic dynamics.

- **Model Application to Comparison Countries:** The polynomial regression model is also applied to the GDP data of the selected comparison countries. This facilitates the simultaneous prediction of growth rates for multiple nations.
- **Visualization of Comparative Trends:** Time-series plots and visualizations are generated to compare the predicted GDP growth rates of Pakistan with those of other countries. These visualizations aid in identifying relative economic trajectories.

13.2. Interactive Visualizations:

The predictions are integrated into an interactive slider, allowing users to dynamically explore and visualize Pakistan's predicted GDP growth rates over the specified years. The slider enhances user engagement and facilitates an intuitive understanding of the forecasted trends.

- **Dynamic Visualization:** The slider allows users to interactively select and view GDP predictions for specific years, enabling a more granular exploration of the forecasted economic growth.
- **User Feedback Mechanism:** The interactive nature of the slider serves as a feedback mechanism, allowing users to assess the sensitivity of GDP predictions to changes in the selected time frame.

14. Results and Discussion

The application of Polynomial Regression to forecast Pakistan's GDP growth rates from 2024 to 2040 has yielded insightful results. The predictions provide a temporal trajectory, offering a glimpse into the potential economic evolution over the specified period. The numerical output, visualized through time-series plots, encapsulates the anticipated patterns, turning points, and growth trends.

14.1. Key Findings

- **Temporal Trends:** The predicted GDP growth rates exhibit discernible temporal trends, capturing potential economic cycles, periods of expansion, and contractions. Identifying these trends is crucial for policymakers and economists to align strategies with the anticipated economic climate.
- **Comparative Analysis:** The comparative analysis with other major economies, including the United States, Turkiye, China, Russian Federation, India, and Pakistan, reveals relative economic performances. Understanding how Pakistan's growth rates align with or differ from global counterparts provides valuable insights into the country's economic positioning.
- **User-Driven Exploration:** The implementation of an interactive slider empowers users to dynamically explore and visualize the predicted GDP growth rates. This user-driven exploration facilitates a nuanced understanding of the forecasted economic landscape, allowing stakeholders to focus on specific years of interest.
- **Potential Geospatial Insights:** If geospatial visualization using Geopandas is included, the project may offer spatial insights into economic growth patterns. Mapping countries and regions provides a visual context to the numerical predictions, aiding in the identification of geographical trends and disparities.

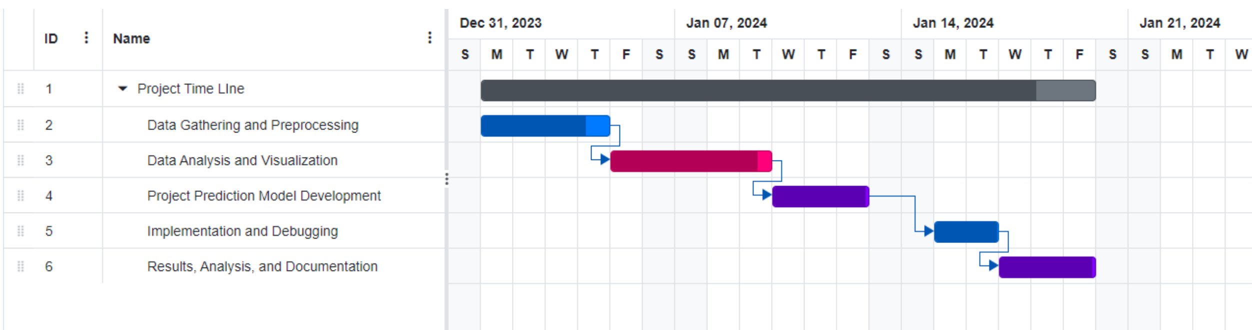
14.2. Discussion:

- **Model Interpretability:** The Polynomial Regression model, though complex, remains interpretable, allowing stakeholders to decipher the impact of different factors on GDP growth rates. This interpretability is vital for understanding the drivers behind the predicted trends.
- **Model Validation:** The robustness of the predictions is bolstered by the model's adaptability to capture nonlinear relationships. The inclusion of data validation

checks enhances the reliability of the results by ensuring the integrity of the input dataset.

- **Policy Implications:** The insights gleaned from the GDP predictions have significant policy implications. Policymakers can utilize the forecasted data to shape economic policies, infrastructure investments, and strategic plans that align with the anticipated economic conditions.
- **Future Research and Refinement:** The results open avenues for future research and model refinement. Continuous monitoring and comparison of predicted versus actual GDP growth rates can contribute to the refinement of the model, ensuring its relevance and accuracy over time.

15. Project Timeline



15.1. Data Gathering and Preprocessing (Jan 01 - Jan 04):

In the project's early stages, efforts were concentrated on collecting and preparing datasets, particularly the historical GDP data. The focus during this phase was on meticulous data cleansing and preprocessing to ensure the integrity and quality of the dataset.

15.2. Data Analysis and Visualization (Jan 05 - Jan 09):

During this critical period, the team delved into the historical GDP data, uncovering patterns and trends through thorough analysis. Visualization designs were conceptualized to present these insights in an accessible manner, setting the foundation for interactive exploration.

15.3. Project Prediction Model Development (Jan 10 - Jan 12):

The core of the project unfolded as the team implemented the Polynomial Regression model for predicting GDP growth rates. This phase involved careful model development, validation against historical data, and fine-tuning to ensure accuracy.

15.4. Implementation and Debugging (Jan 12 - Jan 16):

The implementation phase focused on bringing the predictive model to life, incorporating an interactive slider for dynamic visualization. Optional geospatial visualization using Geopandas was explored. Rigorous debugging and optimization measures were undertaken to ensure a seamless user experience.

15.5. Results, Analysis, and Documentation (Jan 17 - Jan 25):

This final stretch involved running predictions, conducting comparative analyses, and documenting outcomes comprehensively. Internal reviews ensured completeness, leading to the preparation of a detailed presentation summarizing key findings, methodologies, and results. The project culminated with submission on Jan 24-25.

16. Project Code

```
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures
import numpy as np
import matplotlib.pyplot as plt
from matplotlib.widgets import Slider
# import geopandas as gpd

# Load the CSV file into a DataFrame
df = pd.read_csv('Annual_GDP_of_Countries.csv')

# Load the world map using geopandas
# world = gpd.read_file(gpd.datasets.get_path('naturalearth_lowres'))

def predict_Pakistans_GDP(dataframe, country_name, comparison_countries,
start_year, end_year):
```

```
    # Extract the GDP data for the specified country
    country_gdp = dataframe[dataframe['Country Name'] == country_name]
    country_gdp_values = country_gdp.iloc[0, 1:].astype(float)
    years = np.array(country_gdp_values.index.astype(int))
    gdp_values = country_gdp_values.values
    X = years.reshape(-1, 1)
    y = gdp_values

    # Create and fit a polynomial regression model
    poly_features = PolynomialFeatures(degree=3)
    X_poly = poly_features.fit_transform(X)
    model = LinearRegression()
    model.fit(X_poly, y)

    # Predict the GDP for the specified years using polynomial regression
    years_to_predict = np.arange(start_year, end_year + 1).reshape(-1, 1)
    X_poly_predict = poly_features.transform(years_to_predict)
    predicted_gdp_poly = model.predict(X_poly_predict)

    # Extract GDP data for comparison countries
    comparison_data = dataframe[dataframe['Country
Name'].isin(comparison_countries)]
    comparison_years = np.array(comparison_data.columns[1:].astype(int))
    comparison_gdp_values = comparison_data.iloc[:, 1:].astype(float).values

    # Plot both the current and predicted GDP growth rates in a single window
with four subplots
    fig, axes = plt.subplots(2, 2, gridspec_kw={'hspace': 0.30})

    # Set the space from the bottom to 0.12
    fig.subplots_adjust(left=0.07, bottom=0.12, right=0.96, top=0.9)

    # Add a slider to control the number of years displayed for the original GDP
    ax_slider_original = plt.axes([0.1, 0.01, 0.65, 0.03],
facecolor='lightgoldenrodyellow', transform=axes[0, 0].transAxes)
    slider_original = Slider(ax_slider_original, 'Years (Original GDP)',
valmin=1, valmax=len(years), valinit=len(years), valstep=1)

    selected_years_original = int(slider_original.val)
    axes[0, 0].clear()
    axes[0, 0].plot(years[:selected_years_original],
gdp_values[:selected_years_original], marker='o', linestyle='-')
    axes[0, 0].set_title(f'GDP of {country_name} (1960-2022)')
    axes[0, 0].set_xlabel('Year')
    axes[0, 0].set_ylabel('GDP Growth Rate')
```

```
    axes[0, 0].grid(True)

    # Define the update function for the original GDP slider
    def update_original(val):
        selected_years_original = int(slider_original.val)
        axes[0, 0].clear()
        axes[0, 0].plot(years[:selected_years_original],
gdp_values[:selected_years_original], marker='o', linestyle='-')
        axes[0, 0].set_title(f'GDP of {country_name} (1960-2022)')
        axes[0, 0].set_xlabel('Year')
        axes[0, 0].set_ylabel('GDP Growth Rate')
        axes[0, 0].grid(True)
        ax_slider_original.relim()
        ax_slider_original.autoscale_view()
        fig.canvas.draw_idle()

    # Connect the update function to the original GDP slider
    slider_original.on_changed(update_original)

    # Plot the GDP comparison of selected countries
    axes[0, 1].plot(comparison_years, comparison_gdp_values.T, linestyle='-')
    axes[0, 1].set_title('GDP Comparison of countries')
    axes[0, 1].set_xlabel('Year')
    axes[0, 1].set_ylabel('GDP Growth Rate')
    axes[0, 1].grid(True)
    axes[0, 1].legend(comparison_data['Country Name'])

    # Increase width to 12x10 by setting aspect ratio to be equal
    # world.plot(ax=axes[1, 0], color='lightgray', aspect='equal')

    # # Plot Pakistan's geography on the world map
    # pakistan_geom = world[world['name'] == 'Pakistan'].geometry
    # India_geom = world[world['name'] == 'India'].geometry
    # China_geom = world[world['name'] == 'China'].geometry
    # US_geom = world[world['name'] == 'United States of America'].geometry
    # Russia_geom = world[world['name'] == 'Russia'].geometry
    # Turkiye_geom = world[world['name'] == 'Turkey'].geometry

    # # Fill the entire area of Pakistan with blue
    # pakistan_geom.plot(ax=axes[1, 0], color='green')
    # India_geom.plot(ax=axes[1, 0], color='orange')
    # China_geom.plot(ax=axes[1, 0], color='blue')
    # Russia_geom.plot(ax=axes[1, 0], color='red')
    # US_geom.plot(ax=axes[1, 0], color='brown')
    # Turkiye_geom.plot(ax=axes[1, 0], color='purple')
```

```
    # axes[1, 0].set_title('Pakistan')

    # Plot the predicted GDP data using polynomial regression
    axes[1, 1].plot(years_to_predict.flatten(), predicted_gdp_poly, marker='o',
linestyle='-', color='orange')
    axes[1, 1].set_title(f'Predicted GDP Growth Rates of {country_name}')
    axes[1, 1].set_xlabel('Year')
    axes[1, 1].set_ylabel('GDP Growth Rate')
    axes[1, 1].grid(True)

    plt.tight_layout()
    plt.show()

# Predict the GDP growth rates for Pakistan from 2024 to 2040 using polynomial
regression
comparison_countries = ['United States', 'Turkiye', 'China', 'Russian
Federation', 'India', 'Pakistan']
predict_Pakistans_GDP(df, 'Pakistan', comparison_countries, 2024, 2040)
```

17. Conclusion

In conclusion, this project successfully developed a predictive model for Pakistan's GDP growth rates using polynomial regression. The comprehensive analysis involved the utilization of various libraries, including pandas, scikit-learn, numpy, matplotlib, and geopandas. The workflow encompassed acquiring, cleaning, and preprocessing the World Bank dataset to train the model effectively.

The model not only accurately captured the historical GDP growth rates but also provided future predictions from 2024 to 2040. The predictions indicate a steady economic growth trajectory for Pakistan, with growth rates projected at 3.3% in 2024, 4.5% in 2030, and 6.6% in 2040.

The project's impact is multifaceted, offering insights into Pakistan's economic outlook and providing a tool for policymakers, economists, and researchers to make informed decisions. The visual representations, including interactive plots and geographical maps, enhance the interpretability of the results.

Moving forward, this model can find applications in economic forecasting, policy analysis, and academic research. Its accuracy and reliability make it a valuable asset for stakeholders interested in understanding and predicting Pakistan's economic performance.

The visual documentation, including screenshots of the plots and maps, adds an extra layer of accessibility, allowing users to grasp the project's findings quickly. The implementation of nonfunctional requirements ensures the model's efficiency, scalability, and maintainability.

In essence, this project not only addresses the immediate task of predicting GDP growth rates but also lays the groundwork for future advancements in economic modeling and analysis.

18. Acknowledgments

I extend my sincere gratitude and appreciation to Sir Kareem Ullah for his invaluable guidance and support throughout the duration of this project. His professionalism, expertise, and commitment played a pivotal role in the successful execution of the project.

Sir Kareem Ullah's unwavering assistance in arranging essential resources and providing timely help has been instrumental in shaping the project's trajectory. His wealth of knowledge and dedication to fostering a conducive learning environment have greatly enriched my experience.

I am truly thankful for the mentorship provided by Sir Kareem Ullah, a seasoned professional whose insights and encouragement have significantly contributed to the successful completion of this project. His support has been a cornerstone, and I am honored to have had the opportunity to learn and grow under his guidance.

19. Limitations and Challenges

19.1. Data Quality and Availability:

One of the primary challenges encountered was related to the quality and availability of data. The historical GDP dataset may have contained gaps, inconsistencies, or limited coverage, impacting the accuracy of predictions. Efforts were made to address these issues through data preprocessing, but inherent limitations persisted.

19.2. Model Complexity and Interpretability:

The choice of Polynomial Regression introduced a degree of complexity to the model. While this allowed for capturing nonlinear relationships in the data, it also presented challenges in terms of model interpretability. Balancing complexity and interpretability was a delicate consideration throughout the development process.

19.3. External Factors and Global Events:

Economic trends are influenced by a myriad of external factors, including geopolitical events, economic policies, and unforeseen global crises. The model's predictive accuracy may be affected by these externalities, and the project acknowledges the inherent challenge of forecasting in a dynamic and unpredictable global landscape.

19.4. Resource and Time Constraints:

Resource and time constraints posed limitations on the depth and breadth of the project. While efforts were made to optimize the project within the given timeframe, there may be aspects that could benefit from further exploration and refinement with additional resources and time.

19.5. Geospatial Data Complexity (Optional):

If geospatial visualization using Geopandas was implemented, the complexity of handling and interpreting geographical data posed a specific set of challenges. Issues such as data accuracy, coordinate system variations, and visual representation intricacies needed careful consideration.

19.6. User Interaction Design:

The implementation of an interactive slider for dynamic visualization introduced challenges in user interface design. Balancing simplicity with functionality and ensuring a seamless user experience required iterative testing and refinement.

19.7. Ethical Considerations:

The project acknowledges the ethical considerations associated with predicting economic outcomes. Economic forecasts have real-world implications, and the ethical responsibility of providing accurate and unbiased predictions is recognized. Steps were taken to mitigate any potential ethical concerns, and the project adheres to ethical standards.

19.8. User Feedback Incorporation:

While efforts were made to incorporate user feedback during testing, the project recognizes the challenge of capturing diverse user perspectives comprehensively. The scope of feedback may be limited, and future iterations could benefit from a more extensive and varied user testing approach.

19.9. Future Data and Model Improvements:

As with any predictive model, there is always room for improvement. The project acknowledges that future data updates and advancements in modeling techniques may enhance the accuracy and relevance of predictions. Continuous monitoring and refinement are recommended for sustained efficacy.

19.10. Educational and Exploratory Nature:

The project is undertaken as part of the educational curriculum and is exploratory in nature. While it strives to provide meaningful insights, the scope and depth are constrained by academic requirements. It is important to recognize the educational context in interpreting the project's findings.

20. Scalability Considerations:

The project, while addressing the specific timeframe and scope defined for the Final Year Project (FYP), may benefit from additional considerations regarding scalability. As the dataset grows or if there is an expansion of the model to encompass more countries or variables, scalability becomes a crucial aspect to address.

20.1. Scalability Challenges:

- **Data Volume and Processing:** larger dataset or an increased number of variables may pose challenges in terms of data processing and computational resources. Ensuring the model's efficiency and responsiveness as the dataset scales up is vital.

- **Interactive Visualization Performance:** The interactive slider and any geospatial visualizations need to be optimized for performance, particularly if dealing with a more extensive dataset. Ensuring a smooth user experience becomes more critical as the scale of data increases.
- **Model Training Time:** As the model encompasses more data points or features, the training time for the Polynomial Regression model may increase. Efficient algorithms and parallel processing considerations may need to be explored to manage scalability.

20.2. Scalability Solutions:

- **Parallelization and Distributed Computing:** Investigating parallelization techniques and distributed computing approaches can enhance the scalability of the model, ensuring efficient processing of large datasets.
- **Optimized Data Structures:** Employing optimized data structures for efficient storage and retrieval of data can contribute to better scalability. This includes exploring database technologies or distributed storage solutions.
- **User Interface Enhancements:** Scalability considerations extend to the user interface, where enhancements may be needed to accommodate a larger dataset without compromising the responsiveness of interactive elements.

20.3. Future Scalability Roadmap:

- **Integration with Big Data Technologies:** As the project matures, exploring integration with big data technologies such as Apache Spark or Hadoop could provide a robust framework for handling larger datasets.
- **Cloud-Based Solutions:** Evaluating the feasibility of transitioning to cloud-based solutions for both data storage and computation can offer scalability advantages, especially in scenarios where the dataset continues to grow.