* **What are the trends in crop yields over the past decade for various countries?**
  + - Calculated the average yield for each area.
    - Got the top 10 areas based on average yield.
    - Filtered the data frame to include only the top 10 areas.
    - Plotted the yield trends for these top 10 areas and obtained the summary statistics.
* Count shows the number of yield records for each year. The count varies slightly year by year, indicating a relatively consistent number of data points collected annually.
* The mean (average) crop yield shows a slight increasing trend over the years, starting from 671,453 in 2016 to 726,737 in 2022. This suggests an overall improvement in crop yields.
* The standard deviation is very high, indicating a wide variation in yield values. This large variation could be due to different crop types, regions, and farming practices.
* The minimum value for crop yield is 0.0 for all years. This could indicate instances where certain crops failed completely or data was not recorded accurately.
* The 25th percentile shows the yield value below which 25% of the data points fall. This value is relatively stable around 2,000 to 2,100, indicating that a quarter of the yields are consistently low.
* The median (50th percentile) values are in the range of approximately 20,000 to 21,335, indicating that half of the data points have yields below this range. This shows a slight upward trend over the years, reflecting gradual improvement in median crop yields.
* The 75th percentile values show yields below which 75% of the data points fall. These values have a gradual increase, suggesting that the top 25% of yields are improving.
* The maximum yield values show substantial fluctuations, with the highest yields reaching up to approximately 439 million in 2022. These extreme values likely reflect outliers or exceptionally high yields in specific regions or crops.
* **Overall Increase in Average Yield**: The mean crop yield has increased from 671,453 in 2016 to 726,737 in 2022, indicating an overall improvement in agricultural productivity over the decade.
* **Wide Variation in Yields**: The high standard deviation and the presence of very high maximum values suggest significant variability in crop yields. This could be due to differences in geographical regions, crop types, farming techniques, and environmental conditions.
* **Consistent Low Yields**: The 25th percentile values are relatively stable around 2,000, indicating that the lower quartile of crop yields has not seen significant improvement. This might point to challenges faced by certain regions or crops that consistently produce lower yields.
* Improvement in Median and Higher Yields: The upward trend in the median and 75th percentile values indicates that the middle and higher yield-producing crops or regions are improving, contributing to the overall increase in average yields.
* **Outliers and Exceptional Yields**: The extremely high maximum values suggest the presence of outliers or exceptional cases where yields are significantly higher than the average. These outliers could be due to advanced agricultural practices, favourable weather conditions, or specific high-yield crop varieties.
* The trends in crop yields over the past decade show an overall improvement in agricultural productivity, with increasing average yields and better performance in the median and higher yield segments. However, the wide variation in yields and the consistent low yields for the lower quartile indicate areas where improvements are needed. Addressing the challenges faced by low-yield regions or crops and reducing the variability in yields could further enhance agricultural productivity.
* **Which countries have the highest average crop yields?**

•This highlights countries with efficient agricultural practices.

* + Extracted the columns related to area (country) and yield value.
  + Grouped the data by country and calculate the average yield for each country.
  + Sorted the countries based on their average yields in descending order.
  + Printed the top countries with the highest average yields.
  + Converted the result to a Data Frame
  + The output provides the average crop yields for the top 10 countries with the highest yields, highlighting those with the most efficient agricultural practices. Here are some insights:
  + **China, Mainland**:
  + China, Mainland has the highest average crop yield at approximately 5.64 million 100 g/ha. This indicates highly efficient agricultural practices, possibly due to advanced technology, effective farming methods, and significant investment in agriculture.
  + **India**:
  + India follows closely with an average yield of around 5.39 million 100 g/ha. This reflects improvements in agricultural productivity, including better irrigation, use of high-yield crop varieties, and government support for farmers.
  + **Indonesia**:
  + Indonesia's average yield is approximately 2.73 million 100 g/ha, showcasing efficient farming practices, especially in crops like rice and palm oil, which are major agricultural products in the country.
  + **Thailand**:
  + Thailand has an average yield of around 1.08 million 100 g/ha. The country is known for its efficient rice production, benefiting from favourable climatic conditions and effective farming techniques.
  + **Viet Nam**:
  + Viet Nam's average yield is approximately 923,460 100 g/ha. The country has made significant strides in improving agricultural productivity, particularly in rice and coffee production.
  + **Malaysia**:
  + Malaysia has an average yield of about 885,621 100 g/ha. The country's efficient palm oil and rubber production contribute to its high yield.
  + **Pakistan**:
  + Pakistan's average yield is approximately 803,461 100 g/ha. Improvements in irrigation and the adoption of high-yield crop varieties have enhanced agricultural productivity.
  + **Myanmar**:
  + Myanmar's average yield is around 549,950 100 g/ha. The country is focusing on modernizing its agricultural sector to improve efficiency and productivity.
  + **Bangladesh**:
  + Bangladesh has an average yield of about 538,266 100 g/ha. Efficient rice farming practices and government initiatives have contributed to this yield.
  + **Philippines**:
  + The Philippines has an average yield of approximately 529,268 100 g/ha. The country has been working on improving its agricultural practices to increase productivity.
* **Asian Dominance**: All the top 10 countries with the highest average crop yields are from Asia. This region has seen significant agricultural advancements and investments.
* Rice Production: Many of these countries (e.g., China, India, Thailand, Viet Nam, Philippines) are major rice producers. Efficient rice farming techniques and favourable climatic conditions contribute to high yields.
* **Technological and Policy Impact**: Countries with high average yields often benefit from advanced agricultural technology, effective government policies, investment in research and development, and robust infrastructure.
* The countries with the highest average crop yields demonstrate efficient agricultural practices and effective resource utilization. By focusing on technology, infrastructure, and supportive policies, these countries have achieved significant improvements in their agricultural productivity
* **How do agricultural production volumes vary across countries and over time?**

•This provides insight into market sizes and growth trends.

* + Extracted the columns related to area (country), year, and production volume.
  + Grouped the data by country and year, then unstack the year to get a table with countries as rows and years as columns.
  + Used a loop to plot the production volumes for each country over time.
  + Printed the Data Frame to inspect the production volumes across different countries and years.

### Top 5 Countries by Agricultural Production

* China, Mainland
* India
* Indonesia
* Pakistan
* Bangladesh
* China, Mainland:
* Overall Trend: Agricultural production shows a steady increase from 2016 to 2022, reaching a peak in 2022.
* Key Observations: The growth is consistent, reflecting sustained agricultural output despite fluctuations in population and economic factors.
* India:
* Overall Trend: Agricultural production increases gradually over the years, with noticeable peaks in 2018 and 2022.
* Key Observations: India's agriculture remains robust, with periodic spikes possibly influenced by weather conditions and policy changes.
* Indonesia:
* Overall Trend: Agricultural production shows a general upward trend with minor fluctuations, peaking in 2021.
* Key Observations: Similar to India, Indonesia experiences variability likely due to climate impacts and agricultural policies.
* Pakistan:
* Overall Trend: Agricultural production remains relatively stable with slight fluctuations, showing a peak in 2021.
* Key Observations: Pakistan's agriculture sector appears resilient, responding to local and global economic conditions.
* Bangladesh:
* Overall Trend: Agricultural production increases consistently from 2016 to 2022, with significant growth in recent years.
* Key Observations: Bangladesh shows a strong upward trajectory, indicating improvements in agricultural practices and productivity.
* Cross-Country Comparisons:
* Growth Rates: Bangladesh shows the highest growth rate among the top 5, followed closely by China, Mainland.
* Stability: Pakistan exhibits the most stable agricultural production among the top 5 countries, with fewer fluctuations compared to others.
* External Factors: Variability in agricultural production across these countries can be attributed to factors such as climate variability, government policies, technological advancements, and market conditions.
* The top 5 countries in agricultural production demonstrate diverse trends influenced by both internal and external factors. While all countries show growth over the period from 2016 to 2022, the rates and stability vary, reflecting unique challenges and strengths within each agricultural sector.

Understanding these trends can help in strategic planning for agricultural development, resource allocation, and policy formulation to sustain and enhance agricultural productivity.

* **Which countries have shown the most significant growth in agricultural production?**

•Identifying high-growth markets is crucial for expansion decisions.

* + The data is filtered based on Element Code 5510 (sample) to focus on agricultural production
  + The data is pivoted to aggregate production values (Value) by country (Area) and year (Year).
  + Growth rates are calculated by comparing production values between 2016 and 2019.
  + Countries are sorted based on their growth rates in descending order, and the top countries are displayed along with their growth rates.
  + countries that have shown the most significant growth-
* Iraq:
* Growth Rate (%): 67.37%
* Iraq has shown the highest growth rate among the countries listed. From 2016 to 2019, agricultural production more than doubled, indicating substantial growth in this sector. This makes Iraq a high-growth market for agricultural expansion decisions.
* Qatar:
* Growth Rate (%): 49.41%
* Qatar has also experienced significant growth in agricultural production, with a nearly 50% increase from 2016 to 2019. This growth rate suggests a notable expansion in agricultural activities in the country.
* Syrian Arab Republic:
* Growth Rate (%): 47.52%
* Despite challenges, including political and social unrest, Syria has shown strong growth in agricultural production. This growth rate reflects resilience and potential for further development in the agricultural sector.
* Oman:
* Growth Rate (%): 42.13%
* Oman has demonstrated steady growth in agricultural production over the years, indicating ongoing investment and development in this sector. The growth rate suggests favourable conditions for agricultural expansion.
* Kuwait:
* Growth Rate (%): 32.39%
* Kuwait has shown moderate growth in agricultural production. While the growth rate is lower compared to other countries listed, it still signifies progress and potential in the agricultural sector.
* **Expansion Decisions:** Based on these insights, Iraq, Qatar, and Syrian Arab Republic stand out as countries with the highest agricultural production growth rates. These countries could be attractive for expansion decisions due to their robust growth trends.
* **Diverse Growth Rates:** The varying growth rates across these countries also indicate different stages of agricultural development and potential opportunities for investment and partnerships.
* **Consideration of Factors:** When making expansion decisions, factors such as political stability, economic conditions, infrastructure, and market demand should also be considered alongside growth rates to ensure sustainable and profitable investments in the agricultural sector.
* **What are the major crops produced in different countries?**

•Understanding crop specializations helps tailor agribusiness strategies.

* + Grouped the data by 'Area' (country) and 'Item' (crop) to aggregate the production values of each crop in each country.
  + Summed up the production values across different years for each crop in each country.
  + Used idx to identify rows where the production value is the maximum for each country, indicating the major crop.
  + Finally, we printed out the country, major crop, and its production value.
* Diversity of Agricultural Production:
* The dataset includes a variety of crops from different countries, indicating diverse agricultural specializations across regions.
* Regional Crop Specializations:
  + Afghanistan focuses on wheat, reflecting its importance in the country's agricultural output.
  + Armenia prominently produces potatoes, likely due to suitable growing conditions and demand.
  + Azerbaijan also prioritizes wheat, suggesting it is a staple crop in the region.
  + Bangladesh specializes in rice, aligning with its status as a major staple food crop in the country.
  + Bhutan emphasizes bananas, which may be influenced by climate suitability and market demand.
  + Cambodia produces cassava, indicating a significant role in its agricultural sector.
  + China, Mainland focuses on maize (corn), highlighting its importance in China's agricultural landscape.
* Strategic Agribusiness Decisions:
  + Understanding these crop specializations helps tailor agribusiness strategies such as production planning, market targeting, and export/import decisions.
  + It also informs investment decisions in agricultural infrastructure and technology suited to specific crop needs.
* Market Opportunities:
  + Identifying major crops in each country reveals potential market opportunities for agricultural products and technologies tailored to specific crops.
* Impact of Climate and Demand:
  + The choice of crops reflects factors like climate suitability, local demand, and economic considerations in each country.
* **What are the trade policies and regulations affecting agricultural imports and exports in different countries?**
  + To gain insights into trade policies and regulations affecting agricultural imports and exports in different countries -
  + Official websites of agricultural ministries or trade departments often publish information on trade policies and regulations.
  + Understanding regulatory environments helps mitigate risks.
* **Which countries offer the most attractive business climates for agribusiness expansion?**
  + Assessing factors such as ease of doing business, political stability, and infrastructure quality is essential.
  + We can calculate overall score for each country based on the selected indicators and rank countries based on their overall scores to identify top performers.
* **How do agricultural import and export trends vary across countries?**

•Analysed trade data helps understand market demand and supply dynamics.

* + Used Element Code (5510 for imports, 5410 for exports) to filter import and export data.
  + Grouped by Area (country) and Year, summing up the Value column to get total import and export quantities.
  + Computed the percentage change in import and export values using pct\_change() within each country group.
* Afghanistan:
* Import Trends: Afghanistan's agricultural imports showed a notable decline in 2017 and 2018, with growth rates of -9.36% and -11.10% respectively. However, imports rebounded strongly in 2019 with a growth rate of 29.24%, indicating increased demand or recovery in economic conditions.
* Export Trends: Export growth rates were relatively stable during these years, indicating consistent performance in agricultural exports.
* Yemen:
* Import Trends: Yemen experienced varying import trends, with a decrease in 2018 (-2.84%) followed by significant increases in 2019 (16.41%) and 2020 (11.07%). This suggests fluctuations possibly influenced by economic conditions or political stability.
* Export Trends: Yemen's agricultural exports showed a steady increase, particularly notable in 2021 (10.06%), indicating potential growth in agricultural production or market access improvements.
  + **Economic Conditions Impact**: Changes in import and export trends often correlate with broader economic conditions, such as GDP growth, inflation rates, and trade policies.
  + **Policy and Market Access**: Stable or growing export trends indicate successful market access strategies or improvements in product competitiveness. Fluctuating import trends may reflect shifts in domestic production capabilities or changes in consumer preferences.
  + **Year-to-Year Variations**: The year-on-year growth rates highlight the volatility and sensitivity of agricultural trade to external factors.
* **Which countries have favourable conditions for sustainable agricultural practices?**

•This ensures long-term viability and compliance with environmental standards.

* + For domain Crops and livestock products, the data is grouped by country (Area) to calculate average values (mean()) of indicators such as area harvested, yield, or production that signify Crops and livestock products.
  + Identified Top Countries
* The dataset shows zero values for both area harvested and production for these crops in Bangladesh across multiple years (2016, 2017, 2018). This suggests that these crops are not cultivated in Bangladesh at a measurable scale.
* Understanding which crops are not cultivated or have negligible production can inform agricultural stakeholders about where to focus resources
* Diversification Opportunities: Identifying crops that are underutilized or have potential for sustainable cultivation can guide agricultural diversification strategies.
* Resource Allocation: Countries can prioritize resources towards crops that align with both economic and environmental sustainability goals.
* **What are the key consumer preferences for agricultural products in different regions?**

•Tailoring products to consumer preferences enhances market success.

* + Found total consumer preferences for each agricultural product (Item) in each country (Area).
  + Top Agricultural Products by Value:
  + The top agricultural products in Afghanistan by value are:
  + Cantaloupes and other melons: 5,422,681.86
  + Apples: 2,498,477.00
  + Grapes: 8,178,737.00
  + Maize (corn): 2,624,415.00
  + Apricots: 1,481,052.92
  + Consumer Preferences and Market Focus:
* Afghanistan shows significant production and likely consumption of fruits such as cantaloupes, apples, grapes, and apricots. These products may indicate consumer preference for fresh fruits, potentially for local consumption and export.
* Maize (corn) also features prominently, suggesting a significant role in both domestic consumption and potentially as a staple crop.
* Specialty items like cotton lint and cottonseed oil show some production but are less dominant compared to fruits and staple grains.
* Market Strategy Considerations:
* Diversification Opportunities: Given the strong production of fruits like grapes and apples, diversifying into value-added products such as juices, preserves, or dried fruits could capitalize on existing strengths.
* Export Potential: Focusing on improving quality and expanding export markets for fruits and maize could enhance revenue streams and economic stability.
* Local Market Demand: Understanding local preferences for fresh fruits versus processed products can guide market segmentation strategies.
  + Monitoring trends in production and consumption can help in adapting agricultural policies and practices to meet evolving consumer preferences and market demands.
  + Exploring sustainable agricultural practices to maintain productivity while minimizing environmental impact could enhance long-term viability.

**Readme qs**

**An executive summary:**

**- What is your goal?**

The goal of this analysis is to identify trends and opportunities within the agricultural sector across various countries and crops using the FAOSTAT dataset.

By analysing key agricultural metrics such as area harvested, yield, and production volumes, we aim to uncover potential markets for agribusiness expansion.

This analysis seeks to guide strategic decisions for international expansion by pinpointing regions with favourable agricultural trends and robust production capabilities.

**Where did you get your data?**

The primary data source for this analysis is the FAOSTAT dataset, which includes comprehensive agricultural data on various crops and livestock products from numerous countries.

**What are your metrics?**

Area harvested

Yield

Production volumes

**What were your findings?**

Significant growth trends in agricultural production and yields were identified across various countries and products.

Certain crops showed consistent improvement in yield and production, highlighting regions with expanding agricultural productivity.

Potential markets for agribusiness expansion were identified based on these trends.

**What risks/limitations/assumptions affect these findings?**

Data Limitations: The FAOSTAT dataset may have gaps or inconsistencies in certain regions or time periods.

Political, economic, and environmental factors affecting agriculture in different countries were not accounted for.

It is assumed that the trends observed will continue into the future, though this may not always hold true.

**Summarize your statistical analysis, including:**

**- Implementation**

Data Acquisition: The FAOSTAT dataset was imported and pre-processed for analysis.

Data Transformation: Key metrics (area harvested, yield, and production volumes) were extracted and cleaned for analysis.

Statistical Methods: Descriptive statistics and trend analysis were performed to identify significant patterns and opportunities.

**- Evaluation**

Descriptive Statistics: Summarized the data to understand the central tendencies and variability of key metrics.

Trend Analysis: Identified growth trends and significant changes in agricultural productivity over time.

**- Inference**

Growth Trends: Certain countries and crops exhibit consistent growth, indicating potential markets for expansion.

Productivity Analysis: Regions with improving yield and production volumes suggest areas with robust agricultural capabilities.

**How did you acquire your data?**

Data was acquired from the FAOSTAT database, focusing on relevant agricultural metrics.

**How did you transform or engineer your data? Why?**

Exploratory Data Analysis: Visualized and explored key metrics.

Missing values were handled by imputation or removal.

Data types were converted as necessary for analysis.

Analysed trends over time.

Descriptive statistics were used to summarize the data.

Summarized findings and proposed next steps.

**How did you select your model? How did you optimize hyperparameters?**

**Results and Findings from models-**

Linear Regression Models

Baseline Model:

Mean Absolute Error: 3.394417681078034e-09

Mean Squared Error: 1.0107790084728872e-16

R-squared: 1.0

Linear Regression with hyperparameter tuning

**Tuned Model Metrics:**

Mean Absolute Error: 3.394417681078034e-09

Mean Squared Error: 1.0107790084728872e-16

R-squared: 1.0

Insight: Both the baseline linear regression model and the hyperparameter-tuned model perform exceptionally well with an R-squared of 1.0, indicating a perfect fit.

This is highly unusual and suggests that the data might be extremely linear.

Ridge Regression model (with hyperparameter tuning)

Ridge Regression MSE: 365.43432534456224

Insight: The Ridge Regression model significantly deviates from the perfect scores of the linear models. While it regularizes the coefficients to prevent overfitting, its performance suggests that the

linear model was already optimal, and that Ridge regression might not be the best choice for this dataset.

Lasso Regression model (with hyperparameter tuning)

Lasso Regression MSE: 8141336.52291233

Insight: The Lasso Regression model performs much worse than Ridge Regression. This indicates that the features in the dataset might not be sparse, or that Lasso’s ability to set some coefficients to zero is not beneficial for this problem.

Random Forest Regression model (with hyperparameter tuning)

Random Forest Regression MSE: 36166277273.16929

Insight: The Random Forest model has a very high MSE, indicating poor performance compared to linear models. This might suggest that the data has a strong linear relationship that tree-based models fail to capture effectively.

Gradient Boosting Regressor with hyperparameter tuning

Gradient Boosting Regression MSE: 57660245667.82349

Insight: Gradient Boosting also shows high MSE, further indicating that boosting methods might not be suitable for this dataset if the underlying relationships are largely linear.

AdaBoost Regressor with hyperparameter tuning

AdaBoost Regression MSE: 1096522517664.4102

Insight: AdaBoost has an even higher MSE than Gradient Boosting and Random Forest, suggesting that boosting weak learners is not effective for this particular problem.

Decision Tree Regressor with hyperparameter tuning

Decision Tree Regression MSE: 58915043854.47922

Insight: The Decision Tree model also shows a very high MSE, indicating overfitting or that it fails to capture the linear relationships in the data.

Linear Relationships:

The linear regression models (both baseline and tuned) have near-perfect performance, suggesting the data has a very strong linear relationship. This also hints that the dataset may be small or overly simplistic.

Regularization:

The Ridge regression, though not as perfect as the linear regression, still performs better than Lasso, suggesting that regularization might be necessary but the data does not benefit from sparsity.

Non-Linear Models:

Tree-based and boosting models (Random Forest, Gradient Boosting, AdaBoost, and Decision Trees) perform significantly worse, indicating that non-linear relationships or complex interactions are not prominent in the data.

**Model Selection:**

**For this dataset, simpler models like Linear Regression (with or without regularization) are much more effective than complex ensemble methods.**

**Create a medium post of at least 1,000 words summarizing your approach in a tutorial format**

Medium Post: Uncovering Agricultural Trends and Opportunities with FAOSTAT Data

Introduction

Goal of the Analysis:

The primary objective of this analysis is to identify trends and opportunities within the agricultural sector across various countries and crops using the FAOSTAT dataset.

By examining key agricultural metrics such as area harvested, yield, and production volumes, we aim to uncover potential markets for agribusiness expansion, providing strategic insights for international growth.

FAOSTAT Dataset:

FAOSTAT, maintained by the Food and Agriculture Organization (FAO) of the United Nations, offers comprehensive agricultural data from numerous countries.

This dataset includes a wide array of metrics for various crops and livestock products, making it an invaluable resource for analysing agricultural trends globally.

Data Acquisition and Preparation

Data Collection:

The data was sourced directly from the FAOSTAT database, focusing on relevant agricultural metrics. This dataset encompasses information from different countries and years, offering a broad view of agricultural productivity and trends.

Data Cleaning and Preparation:

The dataset required several preprocessing steps to ensure its quality and usability.

These steps included handling missing values, filtering relevant columns, and transforming data types. Key metrics such as area harvested, yield, and production volumes were extracted and cleaned for a more focused analysis.

Exploratory Data Analysis

Key Metrics analysed:

The analysis centered on three primary metrics:

Area Harvested: The total area used for cultivating a particular crop.

Yield: The amount of crop produced per unit area.

Production Volumes: The total quantity of crop produced.

Initial Visualizations and Descriptive Statistics:

To understand the data better, we conducted an exploratory data analysis (EDA), which included generating summary statistics and visualizing the distribution of key metrics.

For instance, we examined the average yield of wheat across different countries to identify significant variations.

Trend Analysis

Methods Used to Analyse Trends:

Trend analysis was performed to identify growth patterns over time. We used statistical methods to analyse changes in yield, area harvested, and production volumes across different countries and crops.

Key Findings:

The analysis revealed several significant trends:

Consistent Growth in Crop Production: Certain countries showed a steady increase in production volumes for key crops, indicating robust agricultural practices and favourable conditions.

Improving Yields: Countries like Afghanistan demonstrated significant improvements in crop yields, particularly for almonds and wheat.

Expansion Opportunities: Countries with expanding agricultural productivity and consistent growth trends were identified as potential markets for agribusiness expansion.

Visualizations:

We created various plots to illustrate these trends. For example, a line chart showing the yield of almonds in Afghanistan over the past decade highlighted a positive growth trend, suggesting a promising market for expansion.

Afghanistan: Significant improvement in almond yields.

Brazil and China: Steady growth in wheat and rice production.

Recommendations for Agribusiness Expansion:

To focus on expanding into Afghanistan for almonds, and explore opportunities in Brazil and China for wheat and rice.

To establish local partnerships, leverage existing infrastructure, and invest in technology to enhance productivity.

Data Limitations:

The FAOSTAT dataset, while comprehensive, may have gaps or inconsistencies in certain regions or time periods. These limitations could affect the analysis's accuracy.

External Factors:

Political, economic, and environmental factors impacting agriculture were not considered in this analysis, which could influence the trends observed.

Assumptions:

The analysis assumes that observed trends will continue in the future, which may not always be the case due to unforeseen changes in agricultural practices or environmental conditions.

Based on the performance metrics provided for various regression models, we can draw the following insights:

Linear Regression Models

Baseline Model:

Mean Absolute Error: 3.394417681078034e-09

Mean Squared Error: 1.0107790084728872e-16

R-squared: 1.0

Linear Regression with hyperparameter tuning

Tuned Model Metrics:

Mean Absolute Error: 3.394417681078034e-09

Mean Squared Error: 1.0107790084728872e-16

R-squared: 1.0

Insight: Both the baseline linear regression model and the hyperparameter-tuned model perform exceptionally well with an R-squared of 1.0, indicating a perfect fit.

This is highly unusual and suggests that the data might be extremely linear.

Ridge Regression model (with hyperparameter tuning)

Ridge Regression MSE: 365.4343253445622

Insight: The Ridge Regression model significantly deviates from the perfect scores of the linear models. While it regularizes the coefficients to prevent overfitting, its performance suggests that the linear model was already optimal, and that Ridge regression might not be the best choice for this dataset.

Lasso Regression model (with hyperparameter tuning)

Lasso Regression MSE: 8141336.522912335

Insight: The Lasso Regression model performs much worse than Ridge Regression. This indicates that the features in the dataset might not be sparse, or that Lasso’s ability to set some coefficients to zero is not beneficial for this problem.

Random Forest Regression model (with hyperparameter tuning)

Random Forest Regression MSE: 36166277273.16929

Insight: The Random Forest model has a very high MSE, indicating poor performance compared to linear models. This might suggest that the data has a strong linear relationship that tree-based models fail to capture effectively.

Gradient Boosting Regressor with hyperparameter tuning

Gradient Boosting Regression MSE: 57660245667.82349

Insight: Gradient Boosting also shows high MSE, further indicating that boosting methods might not be suitable for this dataset if the underlying relationships are largely linear.

AdaBoost Regressor with hyperparameter tuning

AdaBoost Regression MSE: 1096522517664.4102

Insight: AdaBoost has an even higher MSE than Gradient Boosting and Random Forest, suggesting that boosting weak learners is not effective for this particular problem.

Decision Tree Regressor with hyperparameter tuning

Decision Tree Regression MSE: 58915043854.47922

Insight: The Decision Tree model also shows a very high MSE, indicating overfitting or that it fails to capture the linear relationships in the data.

Overall Insights

Linear Relationships:

The linear regression models (both baseline and tuned) have near-perfect performance, suggesting the data has a very strong linear relationship. This also hints that the dataset may be small or overly simplistic.

Regularization:

The Ridge regression, though not as perfect as the linear regression, still performs better than Lasso, suggesting that regularization might be necessary but the data does not benefit from sparsity.

Non-Linear Models:

Tree-based and boosting models (Random Forest, Gradient Boosting, AdaBoost, and Decision Trees) perform significantly worse, indicating that non-linear relationships or complex interactions are not prominent in the data.

Model Selection:

For this dataset, simpler models like Linear Regression (with or without regularization) are much more effective than complex ensemble methods.

**How would you explain and walk through your capstone project to your earlier self?**

Step 1: Understanding the Goal

Step 2: Gathering Data

Step 3: Setting up the Environment and libraries

Step 4: Data Preprocessing and data cleaning

Step 5: Calculating Composite Scores including data normalisation

Step 6: Ranking Countries

Step 7: Visualizing Results

Step 8: Model Building on processed data and hyperparameter tuning

Step 9: Conclusion for the best model without overfitting or underfitting

Step 10: Medium Post summarisation with steps