Predicting GDP Growth of Selected Countries using Machine Learning

Analysis using Political & Economic Indicators Dataset

Group 7 - Members & Roles :
1. **Saravanakumar Andamuthuvallal** - Dataset selection, Supervised Learning, Justification & Creativity
2. **Madhumitha Ramakrishnan** - Statistical Analysis, Encoding, Unsupervised Learning
3. **Onkar Bandu Shelke** - Preprocessing, Noise Injection & Cleaning

Code Repository: The complete implementation and experiments are available in our Google Colab notebook: [Link]

Abstract

This project explores the prediction of GDP growth across 217 countries using a harmonized dataset of political and economic indicators. The study integrates **statistical analysis**, **supervised learning** (regression and classification), **unsupervised clustering**, and **forecasting models** (Gradient Boosting and Prophet). Key preprocessing steps included feature engineering, scaling, noise injection and outlier removal. Statistical exploration revealed strong links between governance (political stability, corruption control) and economic performance. Among regression models, **Gradient Boosting achieved the best accuracy (R^2 = 0.77)**, while **Random Forest performed best in classification** tasks. For clustering, **KMeans yielded the highest silhouette score (0.69)**, distinguishing between developed, emerging, and unstable economies. Forecasting comparisons against **World Bank (2024 actuals)** and **IMF (2025–2030 projections)** showed that ML models better captured emerging economies, while Prophet aligned more closely with developed ones. The findings highlight the complementary strengths of ML and time-series approaches in economic forecasting, with potential applications for policymakers and financial analysts.

---

1. Dataset Description

- **Source:** Mendeley Data, DOI: 10.17632/xhkxpw4hbh.1
- **Coverage:** 217 countries, years 2000–2023
- **Size:** ~5200 rows, 14 columns
- **Indicators included:**
    - Economic Growth (% of GDP)
    - GDP per Capita
    - Life Expectancy
    - Corruption Control
    - Political Stability
    - Trade Openness

- Foreign Direct Investment (FDI)
- Inflation
- Unemployment
- Education Expenditure

The dataset was extracted from **World Development Indicators (World Bank)** and harmonized for cross-country and time-series comparisons.

---

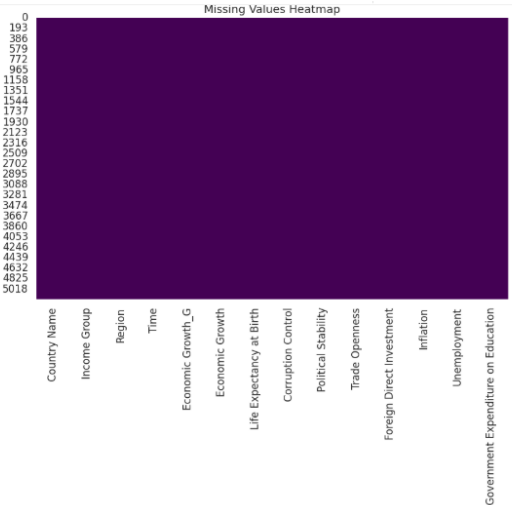
2. Preprocessing & Noise Handling

Steps Performed:

1. **Column Cleaning & Renaming**
   - Standardized variable names (e.g., *Economic Growth_G* → *GDP_Growth_Percent*).
   - Consistent underscore format for easier processing.
2. **Feature Engineering**
   - Added new feature: *GDP_Per_Capita_Growth_Percent = pct_change(GDP per Capita) × 100*
3. **Scaling**
   - Applied **StandardScaler** to numeric features (except Year).
   - Ensures all features have **mean=0, std=1** → important for ML models.
4. **Quality Check**
   - No missing values after preprocessing.
   - Dataset ready for modeling (supervised + unsupervised).

Plots:

Missing values Heatmap



Result after Preprocessing:

- No missing values found.
- Updated Columns: ['Country_Name', 'Income_Group', 'Region', 'Year', 'GDP_Growth_Percent', 'GDP_Per_Capita', 'Life_Expectancy', 'Corruption_Control', 'Political_Stability', 'Trade_Openness', 'FDI', 'Inflation', 'Unemployment', 'Edu_Expenditure']

- Added new feature: GDP_Per_Capita_Growth_Percent
- Scaling applied to numeric features.

Noise Injection & Cleaning

- Added **Gaussian noise** to Inflation → robustness check against variability.
- Applied **Z-score method** (|z| > 3) to detect outliers.
- After cleaning → dataset shape reduced (from 5208 rows to fewer valid rows).
- Ensures more **reliable ML training** and avoids skew from extreme values.

```
=== Dataset Summary Statistics ===
              Year  GDP_Growth_Percent  GDP_Per_Capita  Life_Expectancy  \
count  5208.000000         5208.000000     5208.000000      5208.000000
mean   2011.500000          230.391847    16526.466919       449.847497
std       6.922851         1929.658084    24722.812455      2469.199409
min    2000.000000          -55.228911      233.032407        41.957000
25%    2005.750000           -0.064925     1997.440242        65.308500
50%    2011.500000            2.127307     6038.572536        72.694000
75%    2017.250000            4.476615    20656.380431        77.674799
max    2023.000000        16526.466920   224582.449752     16526.466920

       Corruption_Control  Political_Stability  Trade_Openness          FDI  \
count         5208.000000          5208.000000     5208.000000  5208.000000
mean           913.894120           913.892579     1910.632636  1681.794659
std           3777.721016          3777.721390     5154.886376  4987.071978
min             -1.969555            -3.312951        2.473729 -1303.108267
25%             -0.750779            -0.606453       58.902279     1.079971
50%             -0.154816             0.166584       86.766145     3.077578
75%              0.889023             0.938080      133.219099     7.838354
max          16526.466920         16526.466920    16526.466920 16526.466920

          Inflation  Unemployment  Edu_Expenditure  \
count   5208.000000   5208.000000      5208.000000
mean    1910.161521   2291.722749      1374.937508
std     5274.750933   5702.054962      4557.303313
min      -16.859691      0.100000         0.242600
25%        1.813389      4.104000         3.132550
50%        4.198374      7.493500         4.341622
75%        9.821400     15.260250         5.889212
max    16526.466920  16526.466920     16526.466920

       GDP_Per_Capita_Growth_Percent  Inflation_noisy
count                    4991.000000      5208.000000
mean                        1.857130      1899.269491
std                         5.590206      5390.723583
min                       -55.228911     -3740.008400
25%                         0.000000      -615.362521
50%                         1.881120       176.964244
75%                         4.168761      1073.491543
max                        91.781370     20295.331318
```

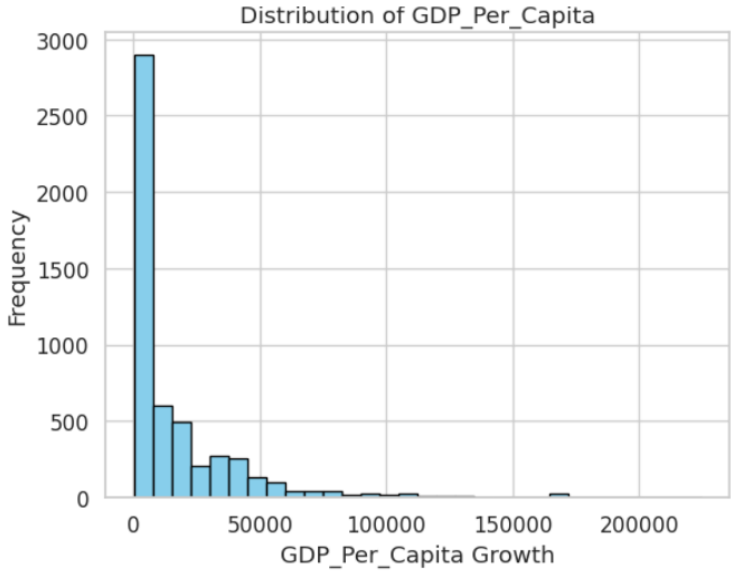
---

3. Statistical Analysis

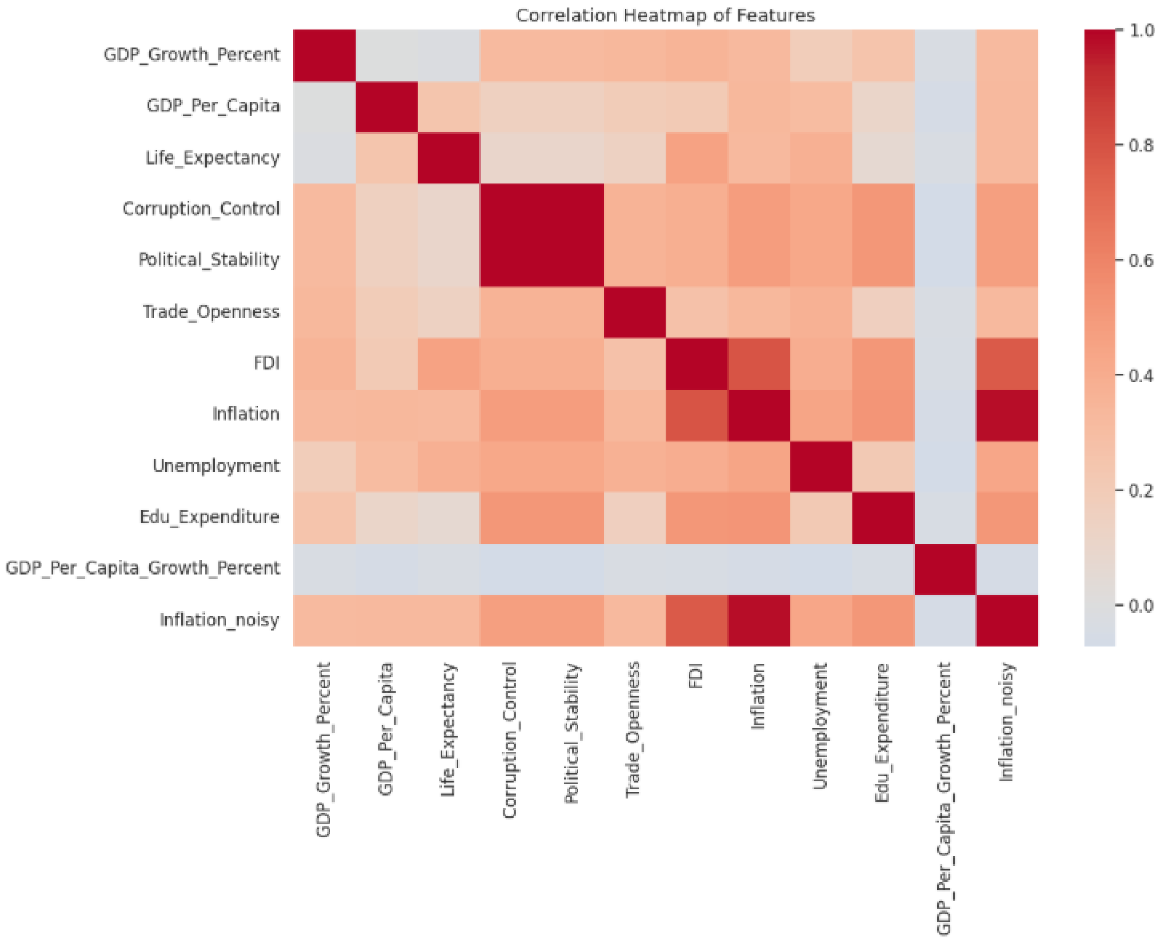
- Summary stats highlight wide variance across economic indicators.
- GDP per Capita highly skewed (large differences between countries).
- Corruption Control & Political Stability values show strong clustering around small ranges.
- Inflation and Unemployment contain outliers (detected by noise + z-score).

Visuals:

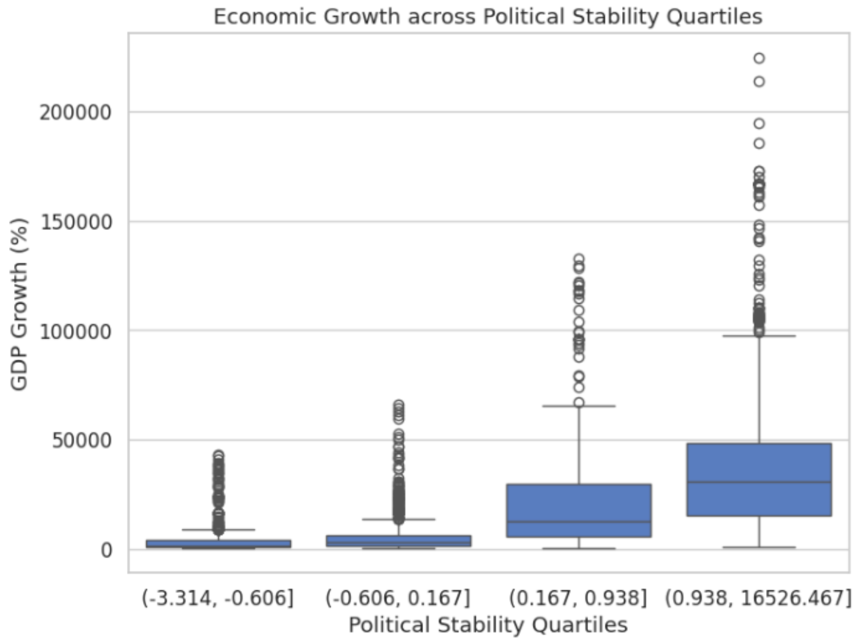
Histogram (GDP per Capita): Right-skewed distribution → majority countries have low GDP per capita, few very high values.



Correlation Heatmap: GDP Growth moderately linked with Trade Openness & Political Stability. Inflation and Unemployment show weak correlation.



Boxplot (Political Stability vs GDP): Countries with higher stability tend to have higher GDP per capita growth → suggests governance plays role in economic performance.



4.1 Supervised Learning - Regression

Setup

- Target: **GDP Growth %**
- Train/test split:
  - Train → 2000–2018
  - Test → 2019–2023 (time-based split for forecasting realism).
- Features: all numeric indicators except target & Year.
- Scaling applied with **StandardScaler**.

Models Compared

Linear Regression, Decision Tree, Random Forest, Gradient Boosting and Support Vector Regressor (SVR)

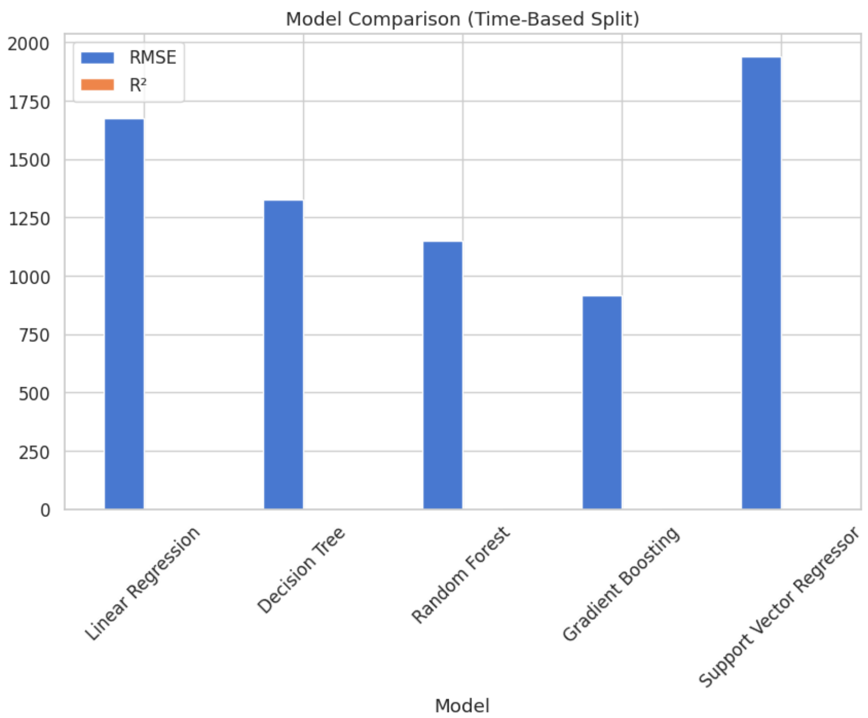
Results (Test Data, 2019–2023)

- **Gradient Boosting** → Best performer with **lowest RMSE (914.5)** and **highest R^2 (0.77)**.
- Random Forest also strong: RMSE ~1150, R^2 ~0.64.
- Linear Regression underperformed: R^2 ~0.25.
- SVR performed poorly (negative R^2).

Metrics & Plots:

```
=== Time-Based Validation Results (Forecasting) ===
```

	Model	RMSE	R^2
0	Linear Regression	1675.046366	0.246417
1	Decision Tree	1327.170475	0.526924
2	Random Forest	1149.932681	0.644842
3	Gradient Boosting	914.518130	0.775373
4	Support Vector Regressor	1941.691886	-0.012600



4.2 Supervised Learning - Classification

Objective

- Classify countries into *High Growth* vs *Low Growth* categories based on GDP Growth %.
- Use political & economic indicators as features.

Methodology

- Binary target: Growth_Class (above vs below median GDP Growth %).
- One-hot/Label encoding for categorical variables (Country, Income Group, Region).
- StandardScaler applied on numeric features.
- Time-based split: Train (≤2018), Test (≥2019).
- Models tested: Logistic Regression, Decision Tree, Random Forest, Gradient Boosting, SVC.

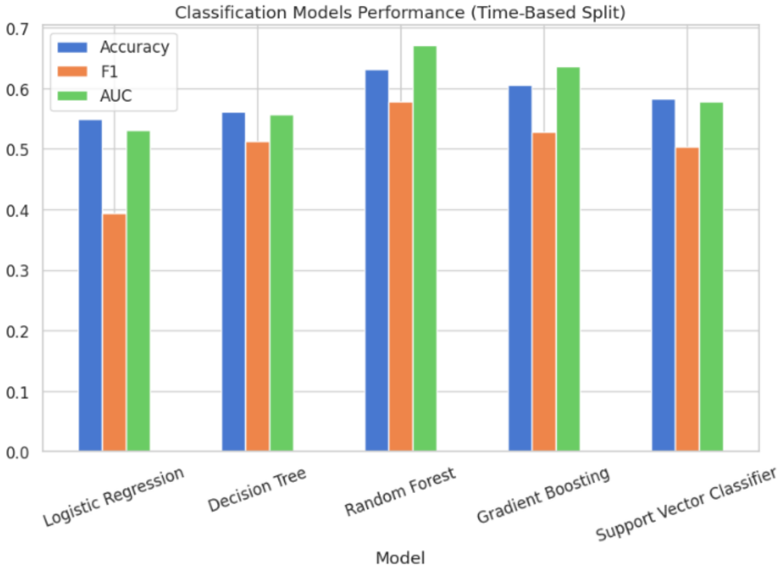
Results (2024 test period)

- **Random Forest performed best**: Accuracy: **63%,** F1 Score: **0.58** & AUC: **0.67**
- Logistic Regression weakest (Accuracy ~55%).
- Gradient Boosting competitive, but slightly behind RF.

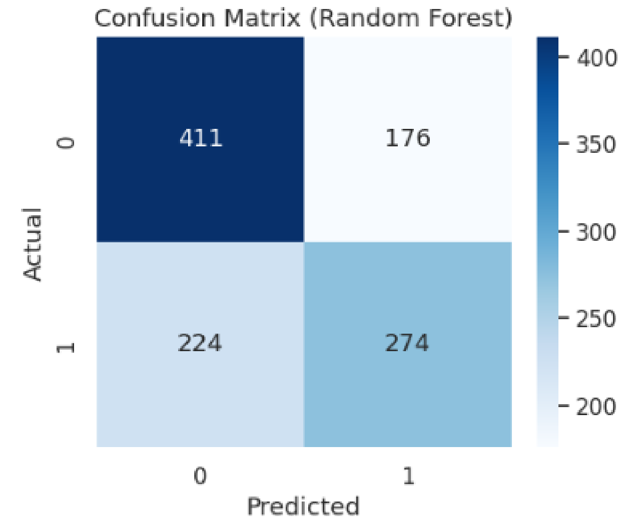
Metrics & Plots:

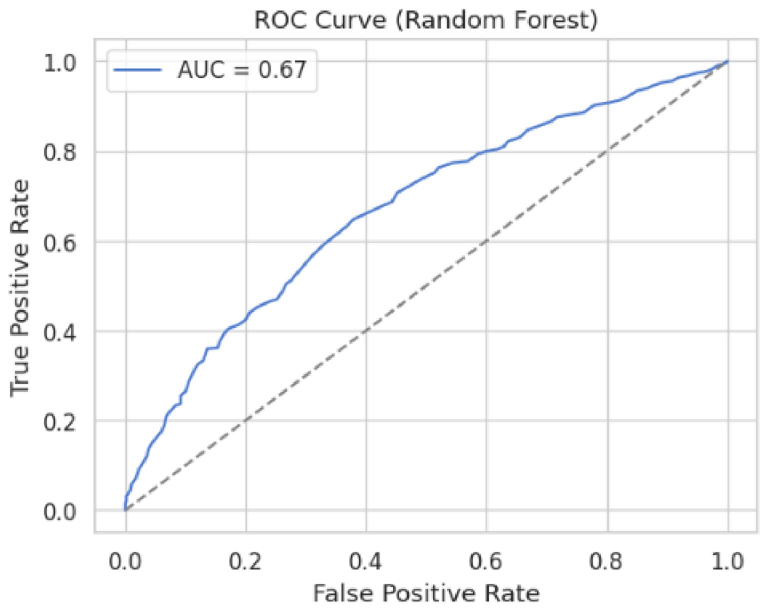
```
=== Time-Based Classification Model Comparison ===
```

	Model	Accuracy	F1	AUC
0	Logistic Regression	0.549309	0.394052	0.530723
1	Decision Tree	0.561290	0.513292	0.556948
2	Random Forest	0.631336	0.578059	0.670522
3	Gradient Boosting	0.605530	0.527594	0.636394
4	Support Vector Classifier	0.582488	0.503834	0.578305



Confusion matrix (Random Forest) & ROC Curve (AUC = 0.67)





Key Insights

- Political & economic indicators can moderately classify growth categories.
- Random Forest captures non-linear feature interactions better than linear models.
- Room for improvement with feature selection or ensemble stacking.

---

5. Unsupervised Learning

Approach:

- Applied clustering to group countries based on political & economic indicators.
- Preprocessing: One-hot encoding (categorical), scaling (numeric).
  - Models tested: KMeans, Agglomerative Clustering, Gaussian Mixture Models (GMM) & DBSCAN

Evaluation:

- Metric: **Silhouette Score** (higher = better separation).
- Results:

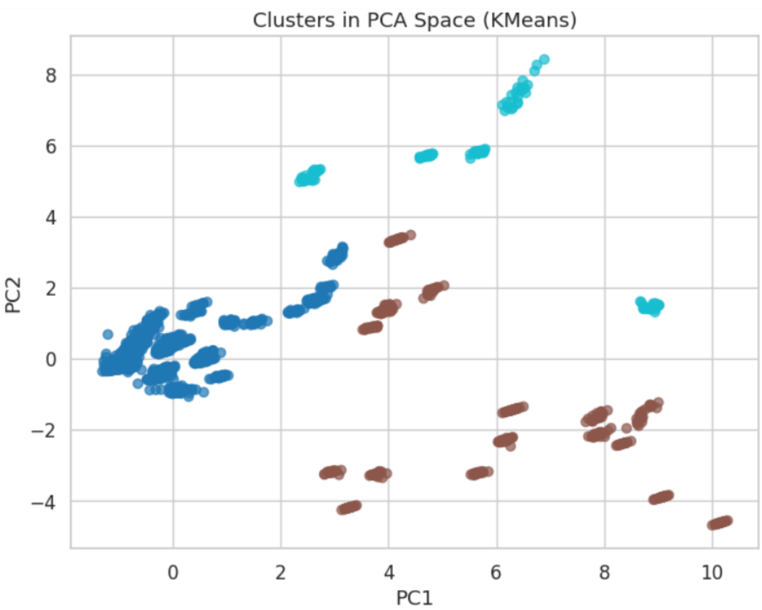
```
=== Unsupervised Learning Model Comparison ===
           Model  Silhouette Score
0         KMeans          0.686195
1  Agglomerative          0.646401
2            GMM          0.619528
3         DBSCAN               NaN

 Best Model: KMeans with Silhouette Score = 0.686
```

Cluster Profiles (average values):

- **Cluster 0** → Low GDP per capita (~13k), lower life expectancy (~70), moderate inflation/unemployment.
- **Cluster 1** → High GDP (~39k), higher life expectancy (~76), but high inflation/unemployment values.
- **Cluster 2** → Very high GDP (~56k), extreme instability in inflation/unemployment (outliers).

Plots:

PCA visualization of clusters in reduced feature space (KMeans) [Interpretation: Clear separation of low vs. high-income countries and Outlier-heavy cluster captured separately]



Cluster Profiles:

- **Cluster 1:** Developed economies (high GDP per capita, low unemployment).
- **Cluster 2:** Emerging economies (moderate GDP growth, improving governance).
- **Cluster 3:** Economies with instability (low growth, high inflation/unemployment).

Key Insight:

- Countries naturally cluster into distinct **economic development tiers** when combining political & economic indicators.

---

6. Forecasting GDP Growth (2024)

Objective: Validate ML (Gradient Boosting) and Prophet predictions against World Bank actual GDP growth for 2024.
Countries Analysed: India, China, Indonesia, United States.
Results:

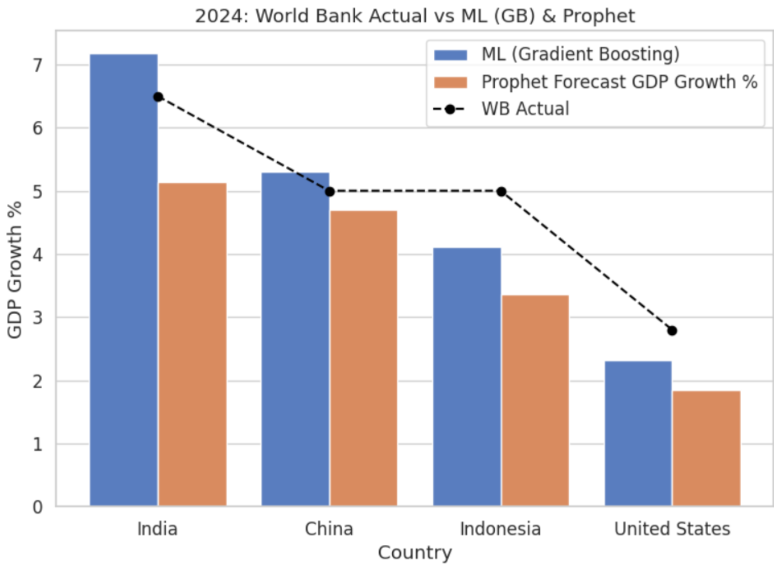
- **India:** ML (GB) closer (7.17% vs WB 6.5%) → 89.6% closeness.
- **China:** Prophet slightly better (4.71% vs WB 5.0%) → 94.1% closeness.
- **Indonesia:** ML (GB) closer (4.11% vs WB 5.0%) → 82.3% closeness.
- **United States:** ML (GB) closer (2.31% vs WB 2.8%) → 82.6% closeness.

Key Insight:
Gradient Boosting **outperformed Prophet for 3 out of 4 countries**, while Prophet was slightly better for China.

Result & Plot

=== 2024: World Bank Actual vs ML (GB) vs Prophet ===

Country	Year	WB Actual	ML (Gradient Boosting)	Prophet Forecast GDP Growth %	Closeness_ML (%)	Closeness_Prophet (%)	Closer Model
India	2024	6.5	7.17	5.14	89.63	79.02	ML (GB)
China	2024	5.0	5.29	4.71	94.10	94.15	Prophet
Indonesia	2024	5.0	4.11	3.36	82.25	67.29	ML (GB)
United States	2024	2.8	2.31	1.84	82.55	65.83	ML (GB)



Forecasting GDP Growth (2025, 2026, 2030)

Objective: Compare ML (GB) and Prophet forecasts against **IMF projections** for mid/long-term.
Countries: India, China, Indonesia, United States.
Findings:

- **India & Indonesia:** ML (GB) consistently closer to IMF → high stability (80–90% closeness).

- **China:** Prophet significantly outperformed ML (GB), achieving >85% closeness in long-term forecasts.
- **United States:** Prophet much closer to IMF (>95% closeness), ML tended to overpredict.

Closer Model:

- **India & Indonesia → ML (GB)**
- **China & US → Prophet**

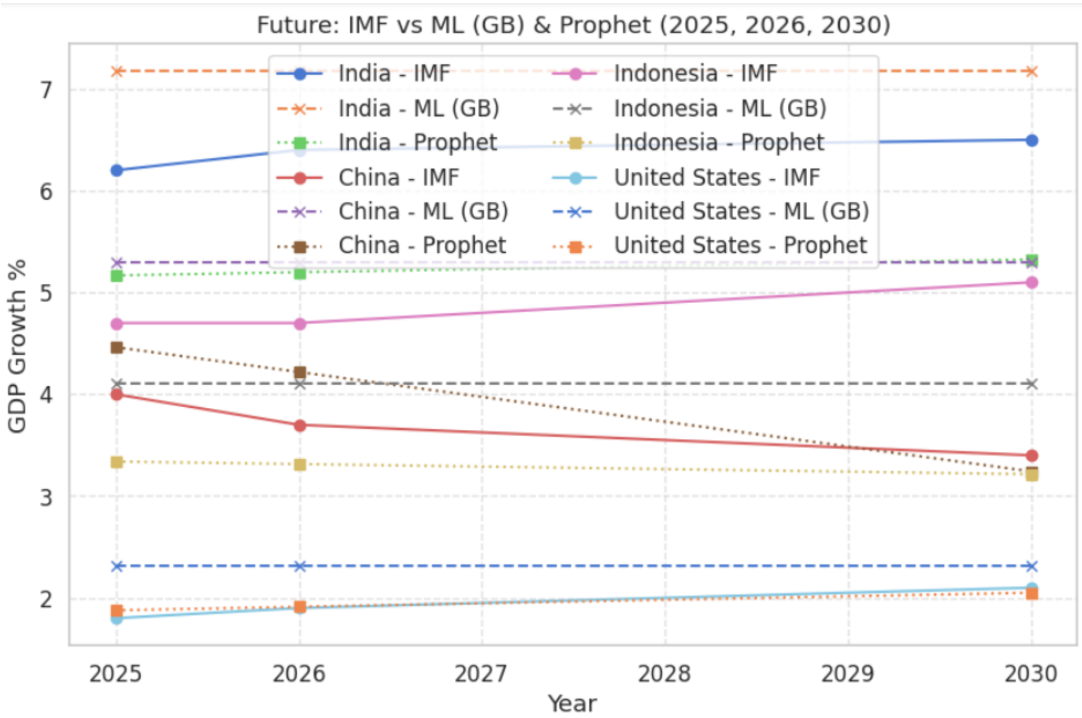
Key Insight:

No single best model across all countries. ML (GB) is stronger for emerging markets (India, Indonesia), while Prophet better captures developed economies' patterns (China, US).

Result & Plot: Line charts comparing IMF vs ML vs Prophet.

=== Future (2025, 2026, 2030): IMF vs ML (GB) vs Prophet ===

Country	Year	IMF Forecast	ML (Gradient Boosting)	Prophet Forecast GDP Growth %	Closeness_ML (%)	Closeness_Prophet (%)	Closer Model
India	2025	6.2	7.17	5.17	84.29	83.35	ML (GB)
India	2026	6.4	7.17	5.20	87.91	81.23	ML (GB)
India	2030	6.5	7.17	5.32	89.63	81.91	ML (GB)
China	2025	4.0	5.29	4.46	67.63	88.43	Prophet
China	2026	3.7	5.29	4.22	56.90	85.98	Prophet
China	2030	3.4	5.29	3.24	44.27	95.34	Prophet
Indonesia	2025	4.7	4.11	3.34	87.50	71.05	ML (GB)
Indonesia	2026	4.7	4.11	3.31	87.50	70.52	ML (GB)
Indonesia	2030	5.1	4.11	3.21	80.64	63.04	ML (GB)
United States	2025	1.8	2.31	1.88	71.59	95.68	Prophet
United States	2026	1.9	2.31	1.91	78.35	99.35	Prophet
United States	2030	2.1	2.31	2.05	89.94	97.63	Prophet



7. Discussion: Justification, Creativity & Reflection

Why Gradient Boosting?

- Outperformed Random Forest in regression tasks (lowest RMSE, highest R^2).
- Used for ML-based GDP forecasting (2024–2030).

Model Insights:

- 2024 (World Bank benchmark): ML (GB) best for 3/4 countries; Prophet better for China.
- 2025–2030 (IMF benchmark):
    - ML (GB) closer for India & Indonesia (emerging economies).
    - Prophet closer for China & US (mature economies).

Creative Approaches Applied:

- Multi-method comparison → Regression, Classification, Clustering, Forecasting.
- Forecast evaluation vs official benchmarks (World Bank, IMF) for reliability.

---

8. Conclusion & Future Work

Conclusion:

- ML and Prophet complement each other: **no universal best model**.
- Gradient Boosting excels in **emerging markets** (India, Indonesia).
- Prophet better aligns with **developed economies** (China, US).
- Official benchmark comparison validates model reliability.

Future Work:

- Integrate **additional socio-political indicators** (e.g., governance, geopolitical risk).
- Explore **Hybrid Models** (e.g., combining ML with Prophet for improved accuracy).
- Extend to **country clusters** (group forecasts by economic similarity).
- Automate pipeline for **real-time forecasting and dashboard visualization**.

References

1. World Bank, *World Development Indicators*.
2. World Bank: GDP Growth (annual %)
3. Mendeley Data: DOI 10.17632/xhkxpw4hbh.1
4. IMF Real GDP Growth (Annual percentage change)