A Survey on Assessing Global Economic Vitality: A Machine Learning Approach to GDP Classification

K.Tamilselvi[1], N.Suvaathi[2], G.Kanisha[3], M.Anandha Harini[4]
[1]*Associate Professor and Head, Department of Computer Science, Shri Nehru Maha Vidyalaya College of Arts and Science, Coimbatore*
[2,3,4]*II M.Sc Computer Science, Shri Nehru Maha Vidyalaya College of Arts and Science, Coimbatore.*

Abstract: In an era of volatile global markets, the accurate assessment of economic health is paramount for policymakers and investors. Traditional econometric models, while foundational, often struggle to capture the non-linear complexities inherent in global economic interactions. This paper presents a machine learning framework utilizing the Random Forest classifier to categorize national economic health based on Gross Domestic Product (GDP) growth. Utilizing a comprehensive dataset of over 200 countries (2010–2025), we engineer features from key fiscal and monetary indicators—including inflation, public debt, and unemployment—to predict discrete growth categories (High, Moderate, Low). Our tuned Random Forest model achieves an accuracy of 0.6561, distinguishing itself from baseline models and validating the potential of ensemble learning methods to provide robust, scalable risk analysis tools for the global economy.

Keywords: Random Forest, Economic Health, GDP Growth, Ensemble Learning, Macroeonomic Indicators, Risk Analysis, Machine Learning

I. INTRODUCTION

The global economy is a dynamic system influenced by a myriad of interconnected factors, from fiscal policy and public debt to inflation and labor market dynamics. Traditionally, forecasting GDP growth has relied heavily on linear regression models and time-series econometrics (ARIMA, VAR). While effective in stable environments, these models often fail to account for the stochastic and non-linear nature of modern economic shocks [1]. The emergence of Machine Learning (ML) and Data Mining offers a paradigm shift. By processing high dimensional datasets and identifying complex patterns without rigid distributional assumptions, ML models provide a more nuanced lens for economic forecasting [2]. Recent literature suggests that techniques such as Random Forest and Granger Causality tests can significantly outperform standard autoregressive models in identifying turning points in economic cycles [3]. This study leverages these advancements to build a classification system that simplifies complex continuous GDP data into actionable "economic health" categories.

II. LITERATURE REVIEW & RELATED WORK

The application of non-parametric algorithms to economic data has gained significant traction. Research indicates that ML approaches, particularly ensemble methods, frequently outperform traditional models in

2.1 Machine Learning in Macroeconomics

Recent studies highlight the efficacy of algorithms like Support Vector Machines (SVM) and Neural Networks in capturing non-linear economic relationships. For instance, forecasting models integrating global trade networks and social media sentiment have shown superior predictive capability over standard linear benchmarks [4]. The ability of these models to handle "big data"—incorporating diverse indicators from varying sources—allows for a more holistic view of economic vitality.

2.2 Comparative Analysis of Classifiers

A critical area of research is the comparison between interpretable models like Logistic Regression and "black-box" ensembles. Comparisons in fiscal stress prediction have found that while Logistic Regression offers parameter interpretability, Random Forest consistently yields higher accuracy (up to 80% in some stress events) by effectively modeling high-order interactions between variables [5]. Similar findings in credit scoring and bank churn prediction corroborate that Random Forest is more robust to noise and multicollinearity, common issues in macroeconomic datasets [6][7].

III. METHODOLOGY

The proposed system follows a rigorous data mining pipeline designed to transform raw economic data into predictive insights.

3.1 System Architecture

The following diagram illustrates the high-level data flow, from raw global indices to the final classification of economic health.

```
graph LR
    A[Raw Economic Data<br/>(200+ Countries)] --> B(Preprocessing &
<br/>Imputation)
    B --> C{Feature Engineering}
    C -->|Debt/Revenue| D[Fiscal Metrics]
    C -->|Inflation*Unemployment| E[Monetary Metrics]
    D --> F[Random Forest<br/>Classifier]
    E --> F
    F --> G((Economic Health<br/>Prediction))
    G --> H[High / Moderate / Low]
```

3.2 Data Acquisition and Preprocessing

We constructed a dataset spanning 16 years (2010–2025) across 200+ countries. Key attributes include:

Monetary: Inflation (CPI & Deflator) [12], Real Interest Rates.
Fiscal: Government Revenue/Expense, Tax Revenue, Public Debt (% of GDP).

Labor/Production: Unemployment Rate, GDP per Capita.

Preprocessing addressed real-world data challenges:

Imputation: Missing values were handled via country-specific median imputation to preserve local economic characteristics.

Outlier Management: The Interquartile Range (IQR) method capped extreme anomalies to prevent model skew [13].

Normalization: Log-transformations (log(1+x)) were applied to highly skewed variables (e.g., Inflation) to stabilize variance.

3.3 Feature Engineering and Selection

To capture interaction effects, we engineered specific domain-relevant features. The "inflation unemployment" interaction serves as a proxy for the misery index, while the "debt-to-revenue" ratio provides a more direct measure of fiscal sustainability than raw debt figures. Feature selection was performed using Chi-Squared and ANOVA F-tests to retain only the most predictive variables, ensuring model parsimony.

We framed the problem as a multi-class classification task with three target labels:
High Growth (≥ 4%)
Moderate Growth (2% - 4%)
Low Growth (< 2%)

IV. RESULTS AND DISCUSSION

The tuned Random Forest model yielded a classification accuracy of 0.6561, significantly outperforming baseline models.

Model	Accuracy	Precision	Recal	F1-Score
Decision Tree	0.5827	0.6061	-	-
k-Nearest Neighbors	0.5482	0.5603	-	-
Random Forest (Tuned)	0.6561	0.6743	0.6561	0.6575

The superior performance of Random Forest (F1-Score: 0.6575) confirms its ability to disentangle the noisy relationships in macroeconomic data. Unlike single Decision Trees which are prone to overfitting, the ensemble approach generalizes better to unseen data [14]. The precision score of 0.6743 suggests that when the model predicts a specific health category, it is correct markedly often, a critical trait for risk management applications.

V. CONCLUSION

This study demonstrates that utilizing machine learning—specifically Random Forest—enables a scalable, data-driven approach to categorizing global economic health. By effectively integrating diverse fiscal and monetary indicators, the model offers a viable alternative to traditional forecasting methods. The inclusion of engineered features like debt sustainability ratios proved crucial, aligning with recent findings on the importance of non-linear fiscal comparisons [15]. Future work should focus on integrating temporal dynamics via LSTM networks and enhancing explainability using SHAP values.

REFERENCES

[1] Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. Journal of Economic Perspectives, 31(2), 87-106.
[2] Chakraborty, C., & Joseph, A. (2017). Machine learning at central banks. Bank of England Staff Working Paper No. 674.
[3] Bok, B., et al. (2018). Nowcasting GDP with Machine Learning and Big Data. Liberty Street Economics.
[4] Nyman, R., et al. (2021). News and narratives in financial systems: Exploiting big data for systemic risk assessment. Bank of England Staff Working Paper.
[5] Alessi, L., et al. (2021). Comparing Machine Learning and Logistic Regression for Fiscal Stress Prediction. European Commission Discussion Paper 146.
[6] Alaoui, A. E., et al. (2019). Random Forest for Financial Distress Prediction. International Journal of Finance & Economics.
[7] Kruppa, J., et al. (2013). Risk estimation and prediction with random forest and logistic regression. Biometrical Journal.
[8] World Bank Group. (2020). Global Economic Prospects, June 2020. Washington, DC: World Bank.
[9] International Monetary Fund (IMF). (2023). World Economic Outlook: Navigating Global Divergences. Washington, DC.