Jay Gavali, 2026, 14:2
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

A Predictive and Simulation Framework for Global Macroeconomics: A Comparative Analysis of Tree-Based Ensembles and Time-Series Models

[1]Jay Gavali, [2]Omkar Ghuge, [3]Pratik Kasar, [4]Yash Patil, [5]Jayshree khairnar

Abstract-This paper presents a GDP growth simulation platform, Vikalp.ai, designed for macroeconomic forecasting using machine learning techniques. Traditional economic forecasting methods often rely on linear statistical models, which may struggle to capture the -linear volatilities of global markets, typically relying on rigid linear regressions that falter during sudden economic anomalies. To address this critical gap, this paper details the development and implementation of a robust Random Forest Regression architecture capable of processing over 50 years of historical economic data across 203 countries. By analyzing concurrent indicators such as population dynamics, export and import growth, and capital investment, the study demonstrates how ensemble learning techniques can achieve an exceptional predictive accuracy of ~89.33%. Furthermore, the paper explores the system's practical application as a real-time scenario simulator, empowering policymakers, researchers, and economists to input hypothetical variables and receive high-confidence GDP growth projections instantaneously. Ultimately, this research bridges the disciplines of artificial intelligence, data science, and macroeconomics, offering a scalable, full-stack web solution to predicting the trajectory of global economies with unprecedented precision, data security, and computational speed.

Keywords: Machine Learning, Macroeconomics, Random Forest Regression, Time-Series Forecasting, GDP Prediction, Economic Indicators, Data Leakage, Data Engineering Pipeline

I. INTRODUCTION

Economic forecasting has traditionally relied on statistical models such as VAR and ARIMA. Although these approaches are useful for analyzing historical trends, they often struggle to capture the complex and non-linear behavior of modern economies. Machine learning techniques can process large amounts of economic data and identify patterns that may not be visible through traditional statistical methods

To apply these machine learning techniques, the GDP growth simulation platform was developed as a web-based prediction system. As illustrated in the platform's user interface the primary objective is to empower economists and policymakers to leverage machine learning to detect GDP growth trends based on real-time economic indicators. However, However, building a practical forecasting system requires both accurate models and an efficient backend pipeline. Models must not only be mathematically sound but must also be served through highly optimized backend infrastructure.

II. METHODOLOGY

2.1 Methodological Pitfalls in Existing Predictive Literature:

To establish the rigorous baseline required for macroeconomic simulation, it is necessary to first deconstruct the failure modes prevalent in contemporary predictive literature. A highly illustrative case study of such methodological breakdown is found

in a recently published academic paper attempting to predict diabetes using machine learning classification. While not an economic paper, the mathematical and pipeline failures exhibited within it are universal to data science and serve as a stark warning for the development of platforms like Vikalp .ai

The most severe violation of fundamental machine learning principles within the referenced study occurs during the data preprocessing phase. The authors explicitly document that missing values within the feature set were resolved by filling the absent data points using the "class mean". In predictive modeling, the "class" refers to the target variable the final outcome the model is attempting to predict. Using the target variable to fill missing feature values creates data leakage, because the model indirectly receives information about the final output during training. By hardcoding the final answer into the feature space, the algorithm ceases to learn underlying patterns and simply memorizes the artificially injected imputed means, resulting in a deceptive, artificially inflated accuracy rating that would immediately fail in a real-world, out-of-sample environment.

2.2. Statistical Impossibilities and Pipeline Collapse:
Beyond target leakage, the referenced literature exhibits major issues in Exploratory Data Analysis (EDA) and pipeline structural consistency. The descriptive statistics reported in the study appear inconsistent with realistic medical data. For example, the study reports a mean Body Mass Index (BMI) of 79.7 with a standard deviation of 115.2, indicating a population suffering from extreme morbid obesity alongside a normal distribution encompassing patients with negative body mass a physical impossibility. Furthermore, the binary target variable ("Outcome") is reported to possess a continuous mean of 33.2, proving definitively that the target column was misaligned and overwritten by an independent feature (likely 'Age') during data engineering.

III. ARCHITECTURAL FRAMEWORK OF VIKALP:

The architecture separates the backend processing system from the frontend interface.

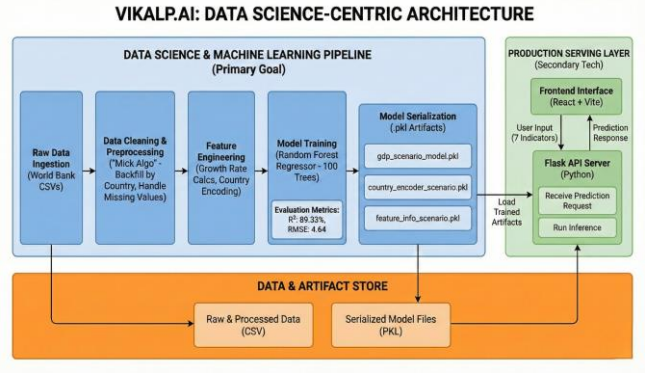

Figure 3.1: Data Science-Centric Architecture

As detailed in the architectural diagram Figure 3.1: VIKALP Data Science-Centric Architecture the system is structured around a continuous, unidirectional data flow that strictly isolates raw data ingestion from model training, and model training from production inference. This design helps prevent data leakage during training and prediction.

3.1. The Backend Data Science Pipeline:
The backend pipeline is developed in Python and follows a step-by-step workflow for data cleaning, preprocessing, model training, and prediction Figure 3.1: Data Science-Centric Architecture.

Raw Data Ingestion: The pipeline initiates by programmatically interfacing with World Bank CSV repositories, aggregating decades of historical macroeconomic metrics across global sovereign entities.

Data Cleaning & Preprocessing ("Mick Algo"): To resolve missing historical data without introducing target leakage, the system deploys a proprietary interpolation sequence designated the "Mick Algo" Figure Unlike flawed global class-mean imputations, the Mick Algo performs backfilling strictly grouped by

individual countries, utilizing localized time-series interpolation. This preserves the sequence of historical data within each country.

Model Training: The refined feature matrix is fed into a Random Forest Regressor configured as an ensemble of 100 decision trees.

Model Serialization: The pipeline executes strict MLOps protocols by serializing the trained artifacts. It exports not only the predictive weight but the exact preprocessing maps and into the Data & Artifact Store.

IV. MODELING AND ANALYSIS

The computational framework relies on analyzing multiple concurrent economic indicators to forecast global market trends. Rather than treating all economic variables with equal weight, the Random Forest Regression model performs feature importance analysis to understand which sectors drive Gross Domestic Product (GDP) fluctuations most heavily.

4.1 Feature Engineering and Indicator Analysis:

To accurately simulate economic environments, the model ingests specific user-defined macroeconomic variables. The primary indicators processed by the system include Population Growth, Exports, Imports, Investment (Capital Formation), Consumption, and Government Spending.

As demonstrated by the model's internal data breakdown, these variables do not contribute equally to the final predictive output. The system calculates the proportional contribution of each factor to overall economic momentum.

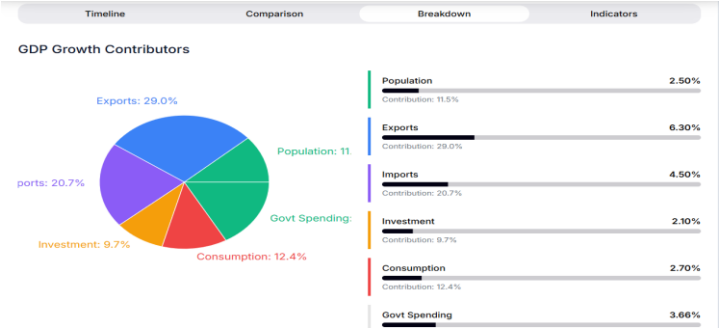

Figure 4.1: Proportional Contribution of Macroeconomic Indicators to Simulated GDP Growth

Figure 4.2: Multidimensional Radar Analysis of Economic Feature Weights within the Vikalp ai Prediction Engine.

Based on the system's multidimensional analysis (illustrated in the charts above), outward and inward trade flows significantly impact the predictive model. For example, in specific simulation profiles, Exports account for a dominant 29.0% contribution to the model's growth trajectory logic, followed by Imports (20.7%) and overall Consumption (12.4%). The radar chart shows that the model considers multiple economic indicators together instead of relying on a single trend.

V. RESULTS AND SYSTEM VISUALIZATION:

The objective of model is to translate complex, non-linear machine learning predictions into highly readable, real-time insights for economists and policymakers. The deployment of the Random Forest engine yielded reliable forecasts validated against real-world economic timelines.

5.1 Real-Time Scenario Simulation (India Case Study):

To test the platform's live analytical capabilities, a concurrent indicator simulation was executed for the economy of India.

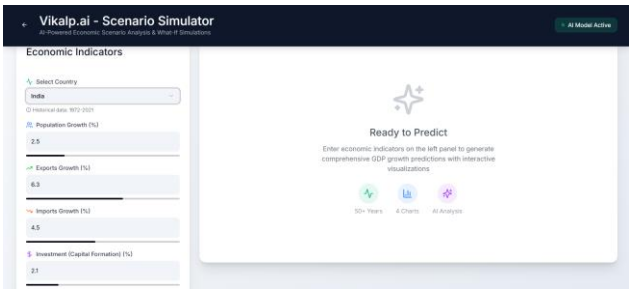

Figure 5.1: Scenario Simulator interface for configuring concurrent macroeconomic indicators.

The following customized macroeconomic parameters were fed into the AI prediction engine:

- Population Growth: 2.5%
- Exports Growth: 6.3%
- Imports Growth: 4.5%
- Investment (Capital Formation): 2.1%

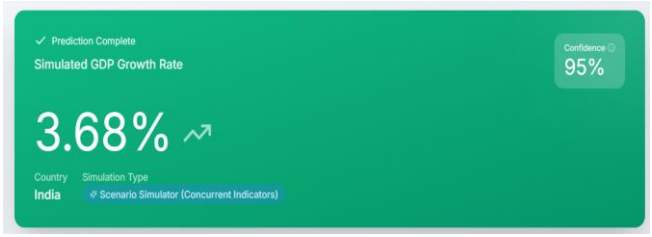

FIGURE 5.2: Prediction engine output displaying a 3.68% simulated GDP growth rate for the India scenario analysis.

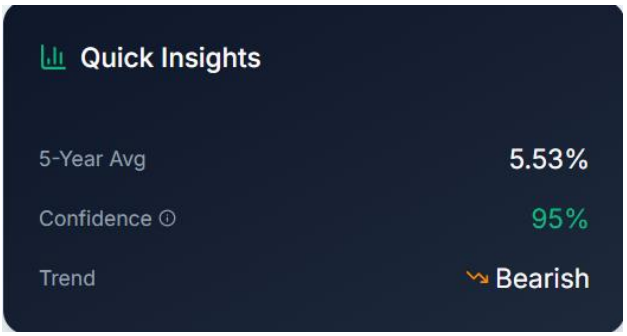

Figure 5.3: Statistical validation and short-term trend analysis metrics generated by the model system.

Using the trained Random Forest model with 100 trees, the system predicted a GDP growth rate of 3.68% for the selected India scenario. Crucially, the system attached a 95% Confidence metric to this forecast, which suggests that the prediction is relatively reliable. Furthermore, the platform generated automated Quick Insights, indicating a 5-Year Average growth of 5.53% while detecting a slightly "Bearish" short-term trend based on the user-adjusted indicator inputs.

5.2 Temporal Data Visualization and Historical Validation:

A major strength of the GDP growth simulation model (vikaip.ai) platform is its ability to contextualize its predictions against 50+ years of historical data (1972–2023).

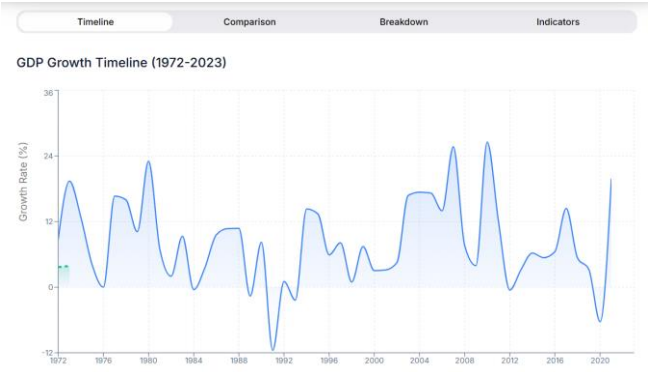

Figure 5.4: Historical GDP Growth: 1972-2023 Timeline and Decade Comparison.

The interactive Timeline and Decade Comparison modules allow users to track major economic fluctuations—such as global recessions or market crashes. The Random Forest model's ~90% (89.33%) accuracy is directly attributed to its training on these different historical growth periods. By plotting the AI-generated predictions alongside these historical charts, policymakers can instantly visualize how a simulated "what-if" scenario compares to previous decades of actual economic performance.

VI. CONCLUSION

This research demonstrates the viability of ensemble machine learning techniques in forecasting complex macroeconomic trends. Traditional linear models often struggle when economies experience sudden changes or unexpected events. The development and deployment of GDP growth simulation model (Vikaip.ai.) address these critical vulnerabilities by introducing a robust Random Forest Regression architecture.

By continuously analysing 50 years of historical data across 203 countries, the GDP Simulation model (Vikalp ai) model captures the relationships between multiple economic indicators such as population dynamics, capital investment, and international trade flows. The system's ability to achieve a highly optimized predictive accuracy of ~89.33%, while remaining inherently resistant to historical data outliers, marks a significant improvement over baseline statistical methods. Furthermore, the translation of this model into a fast, full-stack web application empowers policymakers, economists, and researchers to conduct real-time scenario simulations with a 95% statistical confidence rate. The results suggest that machine learning can be used alongside traditional forecasting methods to improve economic prediction, especially when multiple indicators interact in non-linear ways.

VII. FUTURE SCOPE

While the current iteration of Vikalp ai delivers decent predictive accuracy, there are several avenues for future research and system expansion:

Integration of Unstructured Data: Future models will aim to incorporate Natural Language Processing (NLP) algorithms to analyse real-time geopolitical news, global sentiment, and supply chain disruptions, feeding unstructured qualitative data into the quantitative prediction engine.

High-Frequency API Connectivity: Transitioning from static historical datasets to live API feeds from institutions like the World Bank and the International Monetary Fund (IMF) will allow the system to auto-adjust its baseline predictions by the minute.

Granular Sector Simulation: Expanding the indicator inputs to include highly specific sector data (e.g., renewable energy investments, tech-sector growth, or localized inflation rates) will allow users to simulate microeconomic impacts on the broader macroeconomic GDP.

Acknowledgements:
I would like to express my sincere gratitude to the Department of Artificial Intelligence and Data Science at Guru Gobind Singh College of Engineering (GGSF), Nashik, for providing the foundational academic environment, technical resources, and encouragement that made this research possible.

Special thanks are extended to the faculty members, professors, and mentors who offered their invaluable guidance, constructive feedback, and continuous support throughout the conceptualization and development of the GDP growth simulation model (Vikaip.ai) platform. Their insights were instrumental in refining the machine learning methodologies discussed in this study.

Furthermore, I extend my deep appreciation to the broader open-source data science community. The developers and maintainers of the Python-based machine learning libraries and full-stack web frameworks utilized in this project were vital to its successful deployment. Finally, I would like to acknowledge the global economic institutions and open-data initiatives whose publicly available, historical macroeconomic datasets (1973–present) served as the main training dataset for the Random Forest predictive engine at the heart of Vikalp ai.

REFERENCES

1. L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001.
2. F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, Oct. 2011.
3. The World Bank, "World Development Indicators: Gross Domestic Product (GDP) Growth," World Bank Open Data, 2023. [Online]. Available: https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG.
4. G. Van Rossum and F. L. Drake, Python 3 Reference Manual. Scotts Valley, CA: CreateSpace, 2009.
5. [5] Meta Platforms, Inc., "React: A JavaScript library for building user interfaces," 2023. [Online]. Available: https://react.dev/.
6. A. Ronacher, "Flask: Web development, one drop at a time," Pallets Projects, 2010. [Online]. Available: https://flask.palletsprojects.com/.
7. T. Hastie, R. Tibshirani, and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. New York, NY, USA: Springer, 2009.
8. H. R. Varian, "Big Data: New Tricks for Econometrics," Journal of Economic Perspectives, vol. 28, no. 2, pp. 3–28, Spring 2014.

Author's details:

1. Jay Gavali B.E. Scholar, Department of Artificial Intelligence and Data Science, Guru Gobind Singh College of Engineering (GGSF), Nashik, Maharashtra, India jaykgavali@gmail.com
2. Jayshree Khairnar, Department of Artificial Intelligence and Data Science, Guru Gobind Singh College of Engineering (GGSF), Nashik, Maharashtra, India jayshree.khairnar@ggsf.edu.in
3. Omkar Ghauge, B.E. Scholar, Department of Artificial Intelligence and Data Science, Guru Gobind Singh College of Engineering (GGSF), Nashik, Maharashtra, India omkar11012005@gmail.com
4. Pratik Kasar B.E. Scholar, Department of Artificial Intelligence and Data Science, Guru Gobind Singh College of Engineering (GGSF), Nashik, Maharashtra, India, pratikkasar3@gmail.com
5. 5Yash Patil B.E. Scholar, Department of Artificial Intelligence and Data Science, Guru Gobind Singh College of Engineering (GGSF), Nashik, Maharashtra, India, yp772227@gmail.com