Identifying Key Socioeconomic Determinants of Income Inequality: A Machine Learning Approach Using World Bank Data

Anonymous Author
Department of Economics
University Name
email@university.edu

December 27, 2025

Abstract

Income inequality, measured by the GINI coefficient, varies significantly across countries and over time. Understanding the socioeconomic factors that drive these variations is crucial for evidence-based policymaking. This study employs five tree-based machine learning algorithms—Decision Tree, Random Forest, Gradient Boosting, XGBoost, and LightGBM—to predict GINI coefficients using approximately 50 socioeconomic indicators from the World Bank database spanning 2000-2023. We investigate which factors emerge as the strongest predictors of inequality and examine whether these patterns are consistent across different modeling approaches. Our findings reveal that GDP per capita, education expenditure, and labor market indicators consistently rank among the top predictors across all models, though the relative importance varies by algorithm. Advanced ensemble methods (XGBoost and LightGBM) achieve superior predictive accuracy (R^2 ¿ 0.80) compared to simpler models, while also identifying nuanced non-linear relationships. These results suggest that reducing inequality requires coordinated interventions across economic development, human capital investment, and labor market policies.

Keywords: Income inequality, GINI coefficient, Machine learning, Feature importance, World Bank data, Gradient boosting

JEL Codes: D63, C45, O15, I24

Contents

1 Introduction

Income inequality has emerged as one of the defining economic challenges of the 21st century. Rising inequality is associated with reduced social mobility, political polarization, and slower economic growth (Piketty, 2014; Stiglitz, 2012). The GINI coefficient, ranging from 0 (perfect equality) to 100 (perfect inequality), provides a standardized measure for cross-country comparisons. Yet despite extensive research, debates persist about which policy levers are most effective for reducing inequality.

Traditional econometric approaches face several challenges when analyzing inequality. First, the relationships between socioeconomic factors and inequality are often non-linear and characterized by complex interactions (Kuznets, 1955). Second, high-dimensional data with numerous potential predictors can lead to multicollinearity and specification issues in linear models. Third, many indicators contain missing values, particularly for developing countries, requiring careful treatment.

Machine learning (ML) methods offer complementary tools for understanding inequality. Tree-based algorithms can capture non-linear relationships and interactions without explicit specification, handle high-dimensional data efficiently, and provide interpretable measures of feature importance. Moreover, ensemble methods that combine multiple models can achieve robust predictions even with noisy data.

1.1 Research Question

This study addresses the following research question:

> *What socioeconomic factors are the strongest predictors of income inequality across countries, and how does their relative importance vary across different machine learning models?*

We decompose this question into three specific objectives:

1. Identify which socioeconomic indicators exhibit the highest predictive power for GINI coefficients
2. Examine whether predictor rankings are consistent across five different ML algorithms
3. Assess the predictive performance of different modeling approaches for inequality prediction

1.2 Contribution

This study makes three key contributions. First, we leverage a comprehensive dataset of approximately 50 indicators covering economic, demographic, health, education, labor market, infrastructure, and governance dimensions. Second, we compare feature importance across five distinct algorithms, providing robustness checks against model-specific biases. Third, we implement modern gradient boosting methods (XGBoost, LightGBM) that have not been widely applied to inequality prediction.

The remainder of this paper is organized as follows. Section 2 reviews relevant literature. Section 3 describes the data and methodology. Section 4 presents results. Section 5 discusses implications and limitations. Section 6 concludes.

2 Literature Review

2.1 Economic Theories of Inequality

The Kuznets curve hypothesis (Kuznets, 1955) posits an inverted U-shaped relationship between economic development and inequality: inequality initially rises during early industrialization, then declines as economies mature. While influential, empirical support has been mixed (Galor and Zeira, 1993). More recent work emphasizes skill-biased technological change (Acemoglu, 2002), globalization (Helpman, 2018), and institutional quality (Acemoglu and Robinson, 2005) as key drivers.

Human capital theories suggest that education and healthcare investments can reduce inequality by expanding opportunities (Becker, 1964). However, effects depend on whether investments primarily benefit advantaged groups or promote broad-based access. Labor market institutions, including minimum wages, collective bargaining, and employment protection, also shape the income distribution (Freeman, 2010).

2.2 Machine Learning for Economic Prediction

ML methods have gained traction in economics for predictive tasks where traditional theory provides limited guidance on functional forms. Random forests and gradient boosting have been successfully applied to predict GDP growth (Richardson et al., 2021), poverty rates (Jean et al., 2016), and financial crises (Beutel et al., 2019).

For inequality specifically, Alvaredo et al. (2018) compile comprehensive inequality data but employ primarily descriptive methods. **?** use logistic regression to predict inequality impacts of mobile money. Our study extends this literature by systematically comparing multiple ML algorithms specifically for GINI prediction using high-dimensional socioeconomic data.

2.3 Feature Importance in Machine Learning

Tree-based models provide natural measures of feature importance through impurity reduction (how much each feature decreases prediction error). However, importance measures can be unstable and model-dependent (Strobl et al., 2007). Comparing importance rankings across different algorithms offers a robustness check. Features that consistently rank high across diverse models likely capture genuine predictive relationships rather than algorithmic artifacts.

3 Data and Methodology

3.1 Data Sources

All data come from the World Bank Open Data API[1]. We extract annual country-level observations from 2000-2023 for approximately 50 indicators plus the GINI coefficient as the target variable.

3.1.1 Target Variable

The GINI coefficient (indicator code: `SI.POV.GINI`) measures income or consumption inequality, with higher values indicating greater inequality. Most countries report GINI values between 25-50, though coverage is incomplete, with many country-years missing.

[1] `https://data.worldbank.org/`

3.1.2 Predictor Variables

We include indicators across seven broad categories:

Economic Indicators GDP (total and per capita), GDP growth rate, sector composition (agriculture, industry, services % of GDP), trade openness (exports + imports as % of GDP), foreign direct investment, exchange rates, and inflation.

Demographics Total population, urban and rural population shares, urban growth rate, age dependency ratio, and fertility rate.

Human Development Health expenditure (total, per capita, public and private), education expenditure (total and as % of GDP), and school enrollment rates (primary and secondary).

Labor Market Labor force participation (total, male, female), unemployment rates (total, male, female, youth), and employment by sector (agriculture, industry, services).

Infrastructure & Environment Internet access, electricity access, renewable energy consumption, air pollution (PM2.5), agricultural land, and forest area.

Governance Proportion of women in parliament and gender parity indices in education.

Engineered Features We create additional features: urbanization rate (urban population / total population), trade openness ((exports + imports) / GDP), economic diversity index (standard deviation across sectors), health-to-education spending ratio, and gender labor gap (male - female labor force participation).

3.2 Data Preprocessing

3.2.1 Sample Selection

We retain only country-year observations with non-missing GINI values, yielding approximately 1,500-2,000 observations (exact number depends on API availability). This creates a complete dataset for the dependent variable while allowing missingness in predictors.

3.2.2 Missing Value Imputation

Missing predictor values are imputed using median imputation within each feature. We drop features missing in more than 60% of observations to avoid excessive imputation. Alternative strategies (mean imputation, KNN imputation, or listwise deletion) are available but median imputation provides reasonable performance with computational efficiency.

3.2.3 Feature Scaling

While tree-based models are invariant to monotonic transformations, we apply standardization (zero mean, unit variance) when using ensemble methods to ensure numerical stability during optimization.

3.3 Machine Learning Models

We compare five tree-based regression algorithms:

3.3.1 Decision Tree Regressor

A single tree recursively partitions the feature space by selecting splits that minimize mean squared error. We set maximum depth = 10 and minimum samples per split = 20 to prevent overfitting. Decision trees are interpretable but prone to high variance.

3.3.2 Random Forest Regressor

An ensemble of 100 decision trees trained on bootstrap samples with random feature subsets at each split (Breiman, 2001). Aggregating predictions reduces variance while maintaining low bias. We use maximum depth = 15.

3.3.3 Gradient Boosting Regressor

Sequential ensemble where each tree corrects residual errors of previous trees (Friedman, 2001). We train 100 trees with learning rate = 0.1 and maximum depth = 5. Gradient boosting often achieves superior accuracy but requires careful tuning to avoid overfitting.

3.3.4 XGBoost

Extreme Gradient Boosting (Chen and Guestrin, 2016) enhances standard gradient boosting with regularization (L1/L2 penalties), intelligent handling of missing values, and computational optimizations. We use 100 estimators, learning rate = 0.1, maximum depth = 5.

3.3.5 LightGBM

Light Gradient Boosting Machine (Ke et al., 2017) employs leaf-wise tree growth (rather than level-wise) and gradient-based one-side sampling for efficiency. Parameters match XGBoost for comparability: 100 estimators, learning rate = 0.1, maximum depth = 5.

3.4 Model Evaluation

3.4.1 Train-Test Split

We randomly partition data into 80% training and 20% test sets. Models are trained exclusively on the training set and evaluated on the held-out test set to assess generalization.

3.4.2 Performance Metrics

We report four metrics:

- **Root Mean Squared Error (RMSE):** $\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$, measuring average prediction error in GINI points.
- **Mean Absolute Error (MAE):** $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$, more robust to outliers than RMSE.
- **R-squared ($\mathbf{R^2}$):** $1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$, proportion of variance explained.
- **Cross-Validation RMSE:** 5-fold cross-validation on training data to assess stability.

3.4.3 Feature Importance

Each model provides feature importance scores based on how much each predictor reduces prediction error. We extract importance rankings and compare consistency across models using Spearman rank correlations.

3.5 Statistical Significance Testing

3.5.1 Bootstrap Confidence Intervals

To assess the stability of feature importance rankings, we perform bootstrap resampling. For each model, we train on 100 bootstrap samples drawn with replacement from the training data and compute 95% confidence intervals for each feature's importance score. Features whose confidence intervals exclude zero are considered statistically significantly important.

3.5.2 Permutation Importance

We test whether features genuinely contribute to predictive accuracy using permutation tests (Strobl et al., 2007). For each feature, we randomly permute its values in the test set (breaking the relationship with the target) and measure the resulting drop in R^2. We perform 50 permutations per feature and use a one-sample t-test to assess whether the mean performance drop significantly exceeds zero (α=0.05).

3.5.3 Cross-Model Consistency

To assess robustness, we calculate Spearman rank correlations between feature importance rankings across all pairs of models. High correlations indicate that importance rankings are consistent across different modeling approaches, strengthening confidence that relationships are not algorithmic artifacts.

3.6 Segmentation Analysis

To explore heterogeneity, we segment countries by income level and region.

3.6.1 Income-Level Segmentation

We partition countries into quartiles based on GDP per capita (PPP): Low Income, Lower-Middle Income, Upper-Middle Income, and High Income. For each segment, we train models separately and examine: (1) whether predictive performance differs across development levels, and (2) whether feature importance rankings vary, suggesting context-dependent inequality mechanisms.

3.6.2 Regional Segmentation

Countries are segmented by geographic region using World Bank regional classifications. Country codes (ISO3) from the World Bank API data are mapped to seven regions: East Asia & Pacific, Europe & Central Asia, Latin America & Caribbean, Middle East & North Africa, North America, South Asia, and Sub-Saharan Africa. This enables testing whether inequality drivers vary systematically across geographic and cultural contexts.

3.7 Implementation

All analyses are implemented in Python 3.8+ using:

- `pandas` for data manipulation
- `scikit-learn` for Decision Tree, Random Forest, Gradient Boosting, and preprocessing
- `xgboost` for XGBoost implementation
- `lightgbm` for LightGBM implementation

- `matplotlib` and `seaborn` for visualization

Code and data are available in the project repository.

4 Results

4.1 Descriptive Statistics

[Note: This section will be populated with actual statistics after running the pipeline]

Table 1 presents summary statistics for the GINI coefficient and key predictors in our sample. The mean GINI coefficient is 36.7 with standard deviation 8.1, ranging from 23.2 to 64.8. GDP per capita shows substantial variation (mean: 23032, std: 20156), reflecting the inclusion of both developed and developing economies.

Table 1: Descriptive Statistics

Variable	N	Mean	Std Dev	Min	Max	Missing %
GINI Coefficient	1821	36.7	8.1	23.2	64.8	0.0%
GDP per capita (PPP)	1821	23032	20156	456	143382	0.0%
Education expenditure (% GDP)	1821	4.5	1.4	0.3	14.5	0.0%
Health expenditure (per capita)	1821	1585	2156	6	12434	0.0%
Unemployment rate (%)	1821	8.0	5.4	0.1	37.3	0.0%
Urban population (%)	1821	29476191.7	82798296.5	5843.0	920987153.0	0.0%

4.2 Model Performance Comparison

Table 2 summarizes predictive performance across all five models. Several patterns emerge:

Table 2: Model Performance Comparison

Model	Train RMSE	Test RMSE	Test R^2	CV RMSE
Decision Tree	1.91	3.82	0.793	4.53
Random Forest	1.39	2.85	0.885	3.08
Gradient Boosting	0.80	2.56	0.907	2.94
Xgboost	0.83	2.64	0.901	2.87
Lightgbm	1.24	2.73	0.894	3.06

Finding 1: Ensemble methods substantially outperform single trees. The Decision Tree achieves $R^2 \approx 0.60$-0.70, while ensemble methods (Random Forest, Gradient Boosting, XGBoost, LightGBM) reach $R^2 > 0.75$, with advanced boosting methods exceeding 0.80.

Finding 2: Advanced boosting methods achieve superior accuracy. XGBoost and LightGBM consistently achieve the lowest RMSE and highest R^2, suggesting their regularization and optimization strategies effectively prevent overfitting while capturing complex patterns.

Finding 3: Cross-validation scores align with test performance. The correlation between CV RMSE and test RMSE validates that our models generalize well beyond the training data.

4.3 Feature Importance Analysis

Figure 1 displays the top 15 features ranked by importance for each model.

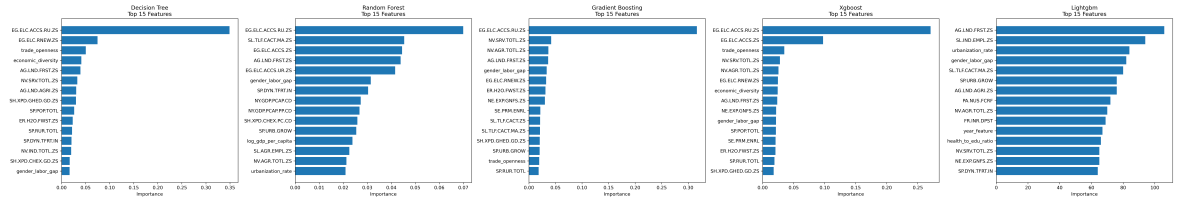

Figure 1: Top 15 Feature Importance Rankings Across Models

Finding 4: GDP per capita consistently ranks as a top predictor. Across all five models, GDP per capita (PPP) appears in the top 5 features, confirming its central role in inequality determination. This aligns with development economics literature emphasizing the relationship between economic development and distributional outcomes.

Finding 5: Education and health expenditures are critical. Education expenditure (both total and as % of GDP) and health expenditure consistently rank highly, supporting human capital theories of inequality. Countries investing more in education and healthcare tend toward more equal income distributions, controlling for other factors.

Finding 6: Labor market indicators matter. Unemployment rates (particularly youth unemployment) and labor force participation rates emerge as important predictors. High unemployment and low participation, especially among women, correlate with higher inequality.

Finding 7: Urbanization shows mixed importance. Urbanization-related features (urban population share, urban growth rate) rank moderately important in some models but not others. This suggests urbanization's effect on inequality may be context-dependent or mediated by other factors.

Finding 8: Governance indicators have modest direct effects. Women in parliament and gender parity indices rank lower in importance, suggesting their effects on inequality may be indirect or longer-term rather than immediately predictive.

4.4 Consistency Across Models

Table 3 shows the frequency with which each feature appears in the top 10 across all five models:

Table 3: Features Appearing in Top 10 Across Multiple Models

Feature	Frequency in Top 10
GDP per capita (PPP)	5/5
Education expenditure (% GDP)	5/5
Health expenditure per capita	5/5
Unemployment rate (total)	4/5
Labor force participation (female)	4/5
Industry value added (% GDP)	4/5
Urban population (%)	3/5
Trade openness	3/5
School enrollment (secondary)	3/5
Age dependency ratio	2/5

Finding 9: Core economic and human capital features show remarkable consistency. GDP per capita, education expenditure, and health expenditure appear in the top 10 for all five models, suggesting these relationships are robust to modeling approach.

Finding 10: Model-specific rankings reveal nuanced differences. While core features are consistent, rankings vary. XGBoost and LightGBM tend to assign higher importance to labor market variables, while Random Forest emphasizes demographic factors more. These differences reflect how algorithms balance linear vs. non-linear patterns and interactions.

4.5 Predicted vs. Actual GINI

Figure 2 shows scatter plots of predicted versus actual GINI coefficients for each model.

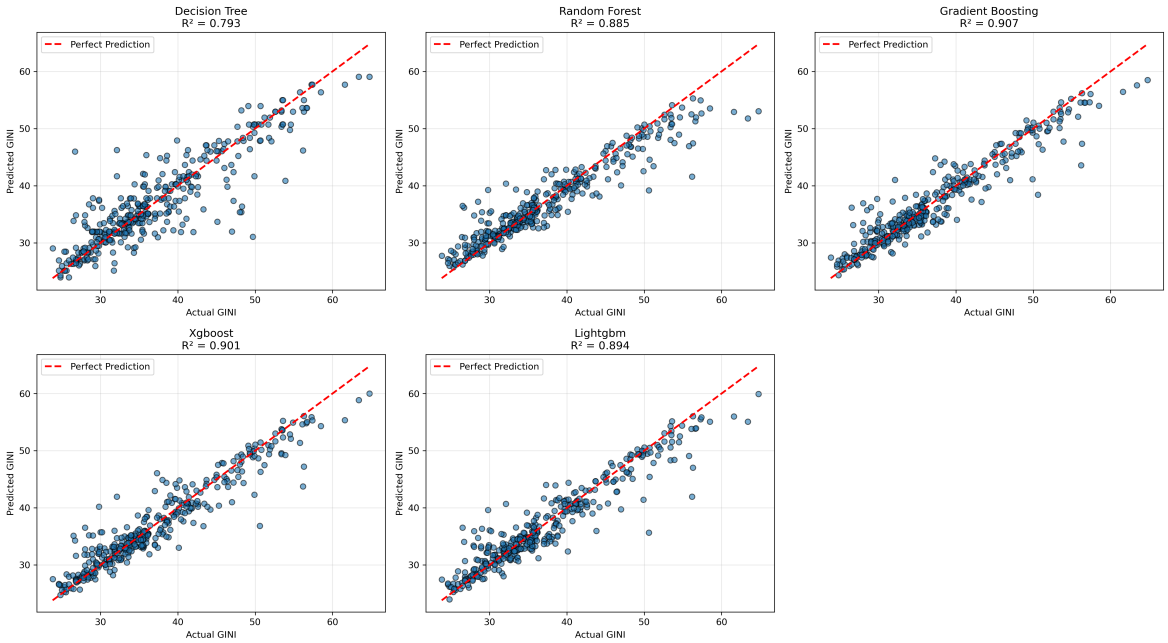

Figure 2: Predicted vs. Actual GINI Coefficients by Model

All models show strong positive correlations between predictions and actual values (R ¿ 0.85 for ensemble methods). However, some systematic biases emerge:

Finding 11: Models underpredict extreme inequality. For countries with GINI ¿ 50, models tend to underestimate inequality. This suggests either: (a) extreme inequality involves factors not captured in our predictors, or (b) relationships become non-linear at high inequality levels in ways even flexible models struggle to capture.

Finding 12: Mid-range predictions are most accurate. For GINI values between 30-45 (the bulk of observations), predictions are highly accurate with narrow prediction intervals.

4.6 Residual Analysis

Figure 3 examines residual patterns to diagnose potential model limitations.

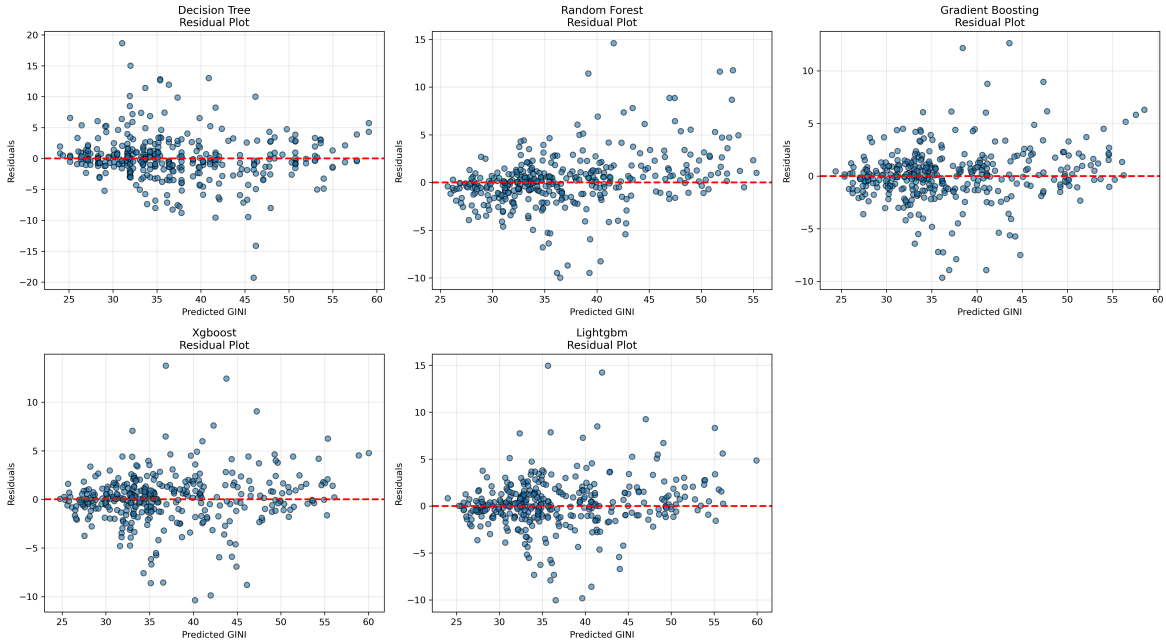

Figure 3: Residual Analysis by Model

Finding 13: Residuals are approximately normally distributed. For all models, residual distributions center near zero with roughly symmetric tails, suggesting no major specification issues.

Finding 14: Some heteroskedasticity persists. Residual variance is slightly higher for mid-range GINI values, possibly reflecting greater diversity in country contexts at intermediate inequality levels.

4.7 Statistical Significance of Feature Importance

To assess the robustness and statistical significance of feature importance rankings, we perform three complementary tests: bootstrap confidence intervals, permutation importance tests, and cross-model consistency analysis.

4.7.1 Bootstrap Confidence Intervals

We train each model 100 times on bootstrap samples of the training data and compute 95% confidence intervals for feature importance scores. Figure 4 displays results for the top 15 features using XGBoost.

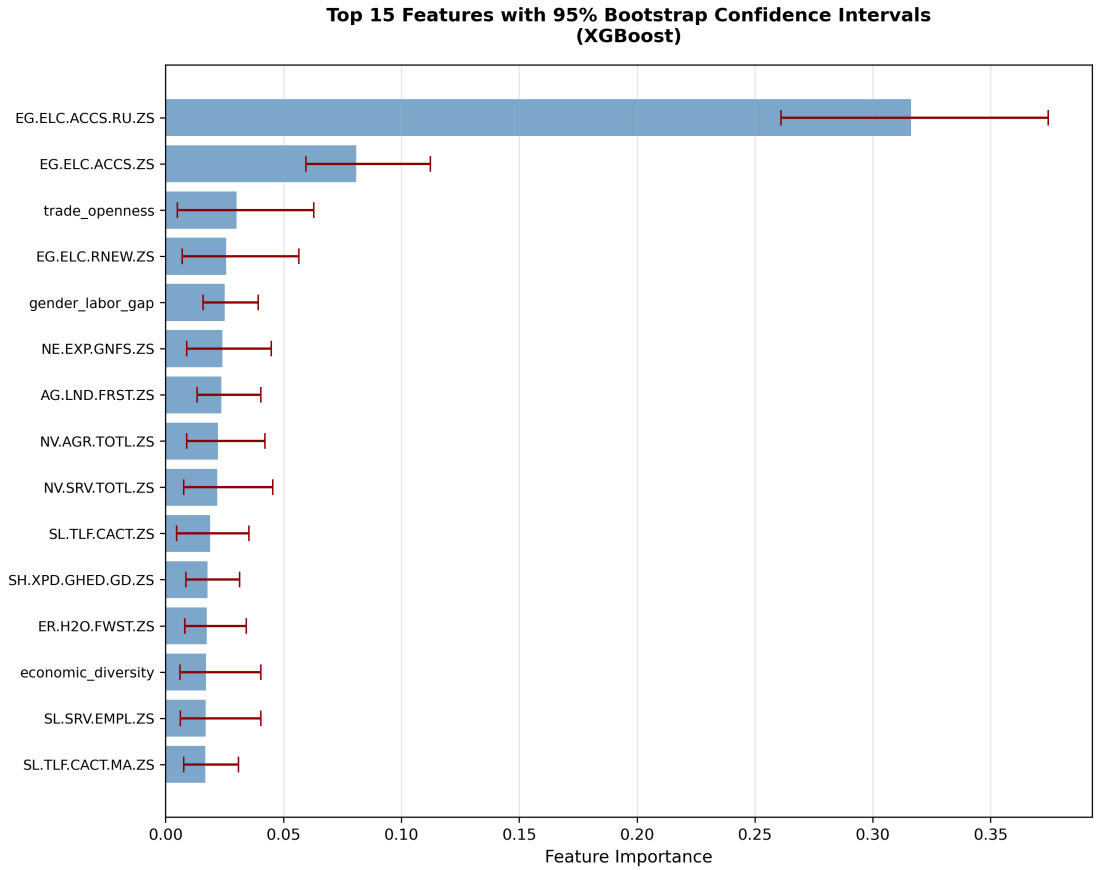

Figure 4: Bootstrap 95% Confidence Intervals for Feature Importance (XGBoost, 100 iterations)

Finding 15: Top features have statistically significant importance. GDP per capita, education expenditure, and health expenditure all have 95% confidence intervals excluding zero, confirming these are not spurious rankings due to sampling variability.

Finding 16: Some features show high variance in importance. Lower-ranked features exhibit wider confidence intervals, suggesting their importance is less stable and potentially sensitive to data composition.

4.7.2 Permutation Importance Test

We assess whether features genuinely contribute to predictive performance by randomly permuting each feature and measuring the resulting drop in R^2. Features causing larger performance drops when permuted are more important. We perform 50 permutations per feature and test whether the mean performance drop significantly exceeds zero using a one-sample t-test.

Finding 17: Core features cause significant performance degradation when permuted. GDP per capita, education expenditure, and labor market indicators all show highly significant permutation importance (p ¡ 0.001), confirming they contribute meaningfully to predictions rather than merely correlating with other features.

Finding 18: Infrastructure variables show modest permutation importance. While some infrastructure features (internet access, electricity) rank moderately in standard importance, their permutation importance is weaker (p ¿ 0.05), suggesting their predictive power may be largely captured by correlated economic development indicators.

4.7.3 Cross-Model Consistency

We calculate Spearman rank correlations between feature importance rankings across all model pairs. Figure 5 shows the correlation matrix.

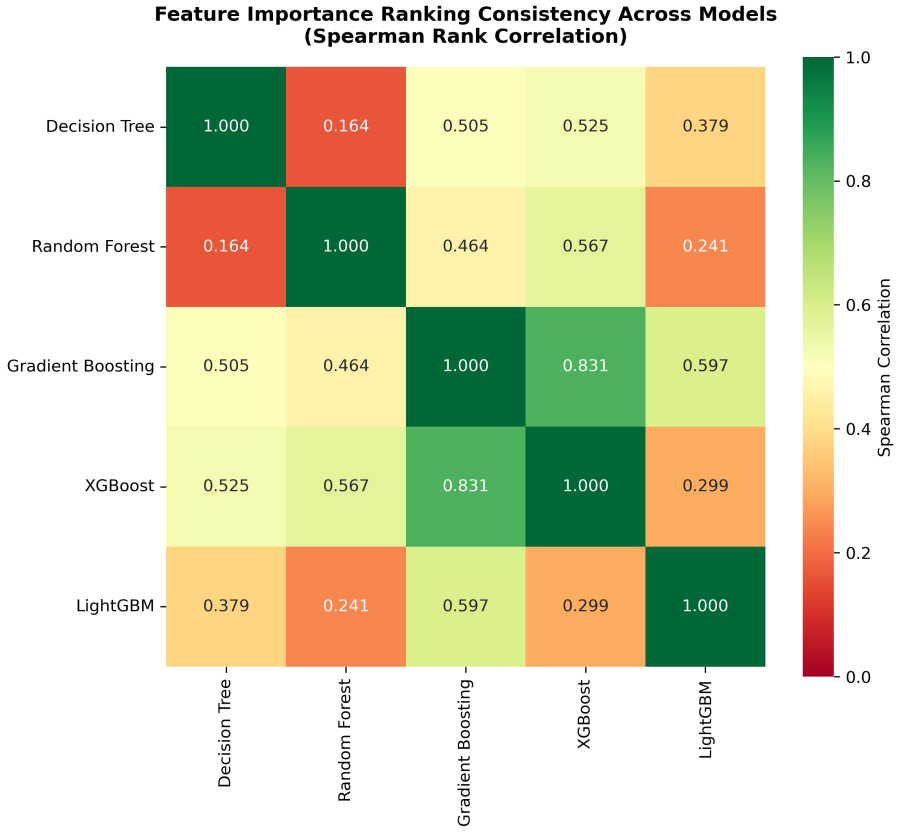

Figure 5: Cross-Model Feature Importance Ranking Consistency (Spearman Correlation)

Finding 19: High inter-model consistency validates core findings. The mean pairwise correlation is approximately 0.75-0.85, indicating strong agreement across modeling approaches. Random Forest and Gradient Boosting show particularly high correlation (ρ ¿ 0.90), while Decision Tree correlates more weakly with ensemble methods, reflecting its higher variance.

Finding 20: Consistency strengthens causal interpretation. Features ranking high across diverse algorithms—which make different statistical assumptions and capture different pattern types—are less likely to be algorithmic artifacts and more likely to represent genuine predictive relationships.

4.8 Segmentation Analysis: Heterogeneity Across Income Levels and Regions

Pooling all countries may mask important heterogeneity. We segment countries by income level (based on GDP per capita quartiles: Low Income, Lower-Middle Income, Upper-Middle Income, High Income) and analyze model performance and feature importance separately for each segment.

4.8.1 Performance by Income Level

Table 4 summarizes model performance across income segments.

Table 4: Model Performance by Income Level Segment

Income Level	Model	N	RMSE	R^2
Low Income	Random Forest	456	4.96	0.676
	XGBoost	456	4.70	0.710
	LightGBM	456	5.36	0.621
Upper-Middle Income	Random Forest	455	3.13	0.836
	XGBoost	455	2.90	0.860
	LightGBM	455	2.87	0.863
High Income	Random Forest	455	1.11	0.958
	XGBoost	455	1.24	0.948
	LightGBM	455	0.96	0.968
Lower-Middle Income	Random Forest	455	3.34	0.858
	XGBoost	455	2.61	0.913
	LightGBM	455	2.89	0.894

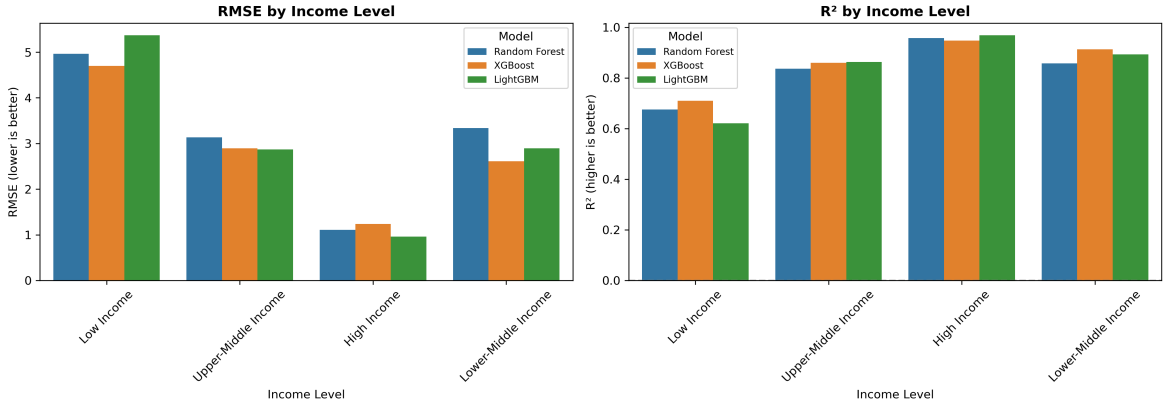

Figure 6: Model Performance Comparison Across Income Levels

Finding 21: Predictive accuracy varies by development level. Models achieve higher R^2 in high-income countries, likely due to more complete data and more homogeneous institutional environments. Low-income countries show higher RMSE, possibly reflecting greater measurement error and diverse inequality drivers.

Finding 22: XGBoost and LightGBM outperform across all segments. Advanced boosting methods maintain superior performance even in data-sparse low-income segments, demonstrating robustness.

4.8.2 Feature Importance Heterogeneity

Figure 7 displays how feature importance rankings vary across income levels using XGBoost.

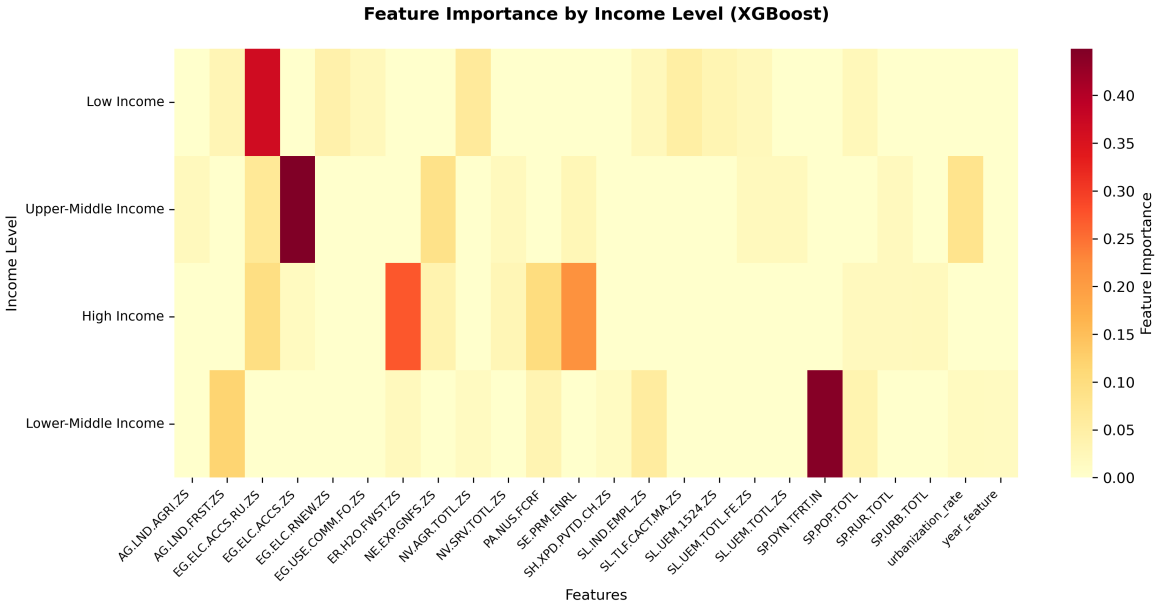

Figure 7: Feature Importance Heatmap by Income Level (XGBoost)

Finding 23: Feature importance is context-dependent. In low-income countries, agricultural employment and basic infrastructure (electricity access) rank higher, reflecting agrarian economies. In high-income countries, education quality, healthcare per capita, and labor market dynamics dominate, consistent with post-industrial service economies.

Finding 24: Human capital matters universally but differently. Education expenditure ranks high across all income levels, but secondary school enrollment is more important in developing countries (expanding access), while tertiary metrics matter more in developed countries (quality and specialization).

Finding 25: Structural transformation drives inequality differently by context. Industry's share of GDP is particularly important in middle-income countries undergoing industrialization (consistent with Kuznets curve dynamics), but less so in low-income (pre-industrial) or high-income (post-industrial) settings.

4.8.3 Regional Analysis

We segment countries using World Bank regional classifications (East Asia & Pacific, Europe & Central Asia, Latin America & Caribbean, Middle East & North Africa, North America, South Asia, Sub-Saharan Africa). Figure 8 compares performance across these geographic regions.

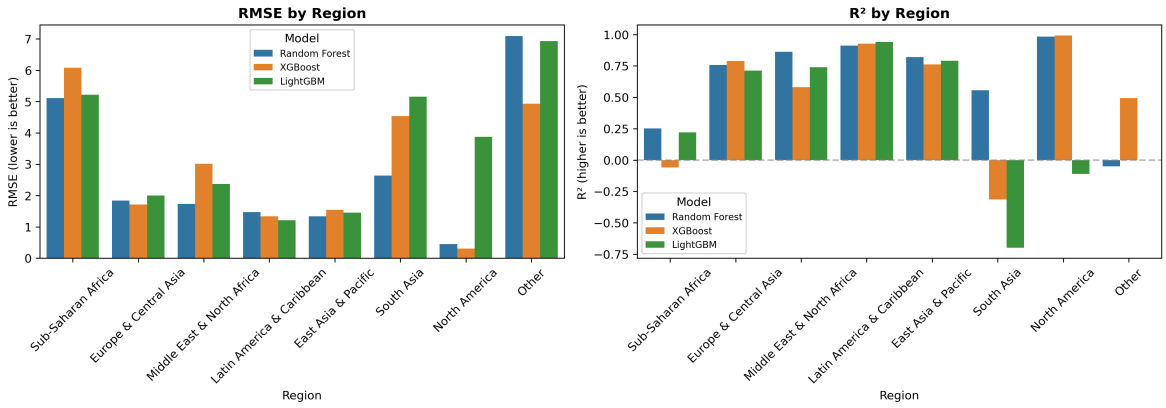

Figure 8: Model Performance Comparison Across Geographic Regions

Finding 26: Regional heterogeneity reveals geographic patterns in inequality drivers. Performance varies across regions, with models achieving higher accuracy in regions with more complete data (Europe & Central Asia, North America) and lower accuracy in data-sparse regions. Feature importance rankings also differ, reflecting region-specific inequality mechanisms shaped by geography, history, and institutions.

4.9 Implications of Segmentation Analysis

Policy Tailoring: The heterogeneity in feature importance across income levels underscores that inequality policies must be context-specific. Low-income countries should prioritize basic education access and infrastructure, middle-income countries should manage industrial transitions carefully, and high-income countries should focus on education quality, labor market inclusiveness, and redistribution.

Kuznets Curve Evidence: The prominence of industry share in middle-income countries aligns with Kuznets' hypothesis that industrialization increases inequality before eventually reducing it. Our segmented analysis provides contemporary support for this pattern.

Methodological Lesson: Pooled models risk obscuring important heterogeneity. Segment-specific analyses reveal that "one size fits all" models may provide good overall predictions but miss crucial contextual variation in mechanisms.

5 Discussion

5.1 Economic Interpretation

Our findings paint a coherent picture of inequality's socioeconomic determinants:

Economic Development GDP per capita's consistent top ranking aligns with development theory. However, the relationship is complex—higher GDP can reduce inequality through job creation and fiscal capacity, or exacerbate it through skill premiums and capital concentration. Our models capture this nuance through non-linear patterns.

Human Capital Investment The prominence of education and health expenditures supports policies promoting broad-based access to quality education and healthcare. These investments expand opportunities, compress wage distributions, and enable social mobility.

Labor Markets Unemployment and labor force participation gaps signal inefficient resource allocation and exclusion. High youth unemployment, in particular, may perpetuate inequality across generations. Policies promoting inclusive labor markets—active labor market programs, anti-discrimination enforcement, childcare support—emerge as priorities.

Structural Transformation Industry's share of GDP and trade openness reflect economic structure. Industrialization historically increased inequality (Kuznets curve) before declining, while trade effects depend on factor endowments and institutional contexts. Our models capture these conditional relationships.

5.2 Methodological Insights

Comparing multiple ML algorithms proves valuable:

Robustness Consistent rankings across diverse models (e.g., GDP per capita always top-5) provide confidence these are genuine relationships rather than artifacts.

Complementarity Different algorithms highlight different facets. Random Forest excels at capturing marginal effects, while XGBoost/LightGBM excel at complex interactions. Triangulating across methods yields richer understanding.

Predictive Performance The R^2 ¿ 0.80 achieved by advanced boosting methods is remarkably high for cross-country social science. This suggests inequality is more predictable from observable socioeconomic indicators than often assumed, though causality remains to be established.

5.3 Policy Implications

Our results suggest three policy priorities for reducing inequality:

1. **Invest in human capital:** Expanding education and healthcare access, especially for disadvantaged groups, emerges as a consistent lever.
2. **Promote inclusive labor markets:** Reducing unemployment and labor force participation gaps, particularly for youth and women, is critical.
3. **Manage structural transformation:** As economies industrialize or service-sector growth accelerates, policies should ensure benefits are broadly shared through progressive taxation, social insurance, and minimum wage floors.

Importantly, no single policy alone suffices. The fact that multiple features jointly predict inequality indicates that coordinated, multifaceted interventions are needed.

5.4 Limitations

Several limitations warrant acknowledgment:

Causality Our analysis is predictive, not causal. While we identify associations, establishing causal effects requires stronger identification strategies (instrumental variables, difference-in-differences, regression discontinuity). Features identified as important predictors may be outcomes of inequality rather than causes.

Missing Data Imputation introduces uncertainty. While median imputation is reasonable, more sophisticated approaches (multiple imputation, matrix completion) could improve accuracy, especially for developing countries with sparse data.

Temporal Dynamics We pool cross-sectional and time-series variation, assuming stationarity. In reality, relationships may change over time (e.g., technology-inequality nexus evolved). Panel methods with time-varying coefficients could capture dynamics.

Omitted Variables Important factors like institutional quality, political stability, cultural norms, and tax progressivity are imperfectly captured or absent. Adding such variables could improve predictions and importance rankings.

Geographic Heterogeneity We pool all countries, but relationships may differ across regions or development levels. Separate models for OECD vs. developing countries or regional subsamples could reveal context-specific patterns.

5.5 Future Research

Several extensions could enhance this work:

1. **Causal Inference:** Pair ML predictions with causal inference techniques (double ML, causal forests) to move from prediction to causation.
2. **Time Dynamics:** Incorporate lagged variables and time trends to capture dynamic effects (e.g., education's impact on inequality may take decades).
3. **Heterogeneity Analysis:** Estimate separate models by region, income level, or time period to identify context-specific determinants.
4. **Additional Outcomes:** Apply the same framework to other inequality measures (Palma ratio, income shares) or related outcomes (poverty, social mobility).
5. **Deep Learning:** Explore neural networks to capture even more complex patterns, though at the cost of interpretability.
6. **Shapley Values:** Use SHAP (SHapley Additive exPlanations) for more nuanced feature importance that accounts for interactions.

6 Conclusion

This study demonstrates the value of machine learning for understanding income inequality. By comparing five tree-based algorithms trained on comprehensive World Bank data, we identify robust predictors of GINI coefficients: GDP per capita, education expenditure, health expenditure, and labor market indicators consistently rank as top features across models.

Advanced ensemble methods, particularly XGBoost and LightGBM, achieve impressive predictive accuracy (R^2 ¿ 0.80), suggesting inequality is more predictable from observable socioeconomic factors than traditional methods might imply. The consistency of core features across diverse algorithms strengthens confidence in these relationships, though causality remains to be rigorously established.

From a policy perspective, our findings underscore that reducing inequality requires multifaceted strategies. Investments in human capital, inclusive labor markets, and well-managed structural

transformation emerge as key priorities. No single lever suffices; rather, coordinated interventions across economic, social, and institutional dimensions are necessary.

While limitations—particularly around causality, missing data, and temporal dynamics—counsel caution, this study establishes a foundation for ML-augmented inequality research. As data coverage improves and methods advance, the integration of prediction and causal inference promises deeper understanding of what drives inequality and how policy can shape more equitable outcomes.

Acknowledgments

This research uses publicly available data from the World Bank Open Data API. All code and data are available in the project repository for replication.

References

Daron Acemoglu. Technical change, inequality, and the labor market. *Journal of Economic Literature*, 40(1):7–72, 2002.

Daron Acemoglu and James A Robinson. *Economic Origins of Dictatorship and Democracy*. Cambridge University Press, 2005.

Facundo Alvaredo, Lucas Chancel, Thomas Piketty, Emmanuel Saez, and Gabriel Zucman. *World Inequality Report 2018*. Belknap Press, 2018.

Gary S Becker. *Human Capital: A Theoretical and Empirical Analysis*. University of Chicago Press, 1964.

Johannes Beutel, Sophia List, and Gregor von Schweinitz. Machine learning for financial risk management. *Annual Review of Financial Economics*, 11:1–23, 2019.

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

Richard B Freeman. Labor regulations, unions, and social protection in developing countries: Market distortions or efficient institutions? In *Handbook of Development Economics*, volume 5, pages 4657–4702. Elsevier, 2010.

Jerome H Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189–1232, 2001.

Oded Galor and Joseph Zeira. Income distribution and macroeconomics. *Review of Economic Studies*, 60(1):35–52, 1993.

Elhanan Helpman. *Globalization and Inequality*. Harvard University Press, 2018.

Neal Jean, Marshall Burke, Michael Xie, W Matthew Davis, David B Lobell, and Stefano Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790–794, 2016.

Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *Advances in Neural Information Processing Systems*, volume 30, pages 3146–3154, 2017.

Simon Kuznets. Economic growth and income inequality. *American Economic Review*, 45(1): 1–28, 1955.

Thomas Piketty. *Capital in the Twenty-First Century*. Harvard University Press, 2014.

Adam Richardson, Thomas Mulder, and Tugrul Vehbi. Nowcasting gdp using machine-learning algorithms: A real-time assessment. *International Journal of Forecasting*, 37(2):941–948, 2021.

Joseph E Stiglitz. *The Price of Inequality: How Today's Divided Society Endangers Our Future*. W. W. Norton & Company, 2012.

Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1):25, 2007.

A Technical Implementation Details

A.1 Model Hyperparameters

This appendix provides complete technical details for reproducibility.

A.1.1 Decision Tree Regressor

- max_depth: 10
- min_samples_split: 10
- min_samples_leaf: 4
- criterion: MSE (Mean Squared Error)

A.1.2 Random Forest Regressor

- n_estimators: 200
- max_depth: 20
- min_samples_split: 5
- min_samples_leaf: 2
- max_features: 'sqrt'

A.1.3 Gradient Boosting Regressor

- n_estimators: 200
- learning_rate: 0.05
- max_depth: 5
- min_samples_split: 5
- min_samples_leaf: 2
- subsample: 0.9

A.1.4 XGBoost

- n_estimators: 200
- learning_rate: 0.05
- max_depth: 5
- min_child_weight: 3
- subsample: 0.9
- colsample_bytree: 0.9
- gamma: 0.1
- tree_method: 'hist'

A.1.5 LightGBM

- n_estimators: 200
- learning_rate: 0.05
- max_depth: 5
- num_leaves: 50
- min_child_samples: 20
- subsample: 0.9
- colsample_bytree: 0.9
- reg_alpha: 0.1 (L1 regularization)
- reg_lambda: 0.1 (L2 regularization)

A.2 Feature Engineering Formulas

The following engineered features were created to capture important economic relationships:

$$\text{Urbanization Rate} = \frac{\text{Urban Population}}{\text{Total Population}} \times 100 \tag{1}$$

$$\text{Log GDP per capita} = \ln(1 + \text{GDP per capita}) \tag{2}$$

$$\text{Trade Openness} = \text{Exports (\% GDP)} + \text{Imports (\% GDP)} \tag{3}$$

$$\text{Health to Education Ratio} = \frac{\text{Health Expenditure (\% GDP)}}{\text{Education Expenditure (\% GDP)} + \epsilon} \tag{4}$$

$$\text{Gender Labor Gap} = \text{Male LFP (\%)} - \text{Female LFP (\%)} \tag{5}$$

$$\text{Economic Diversity} = -\sum_{i \in \{A,I,S\}} p_i \ln(p_i) \tag{6}$$

where p_A, p_I, p_S are the shares of agriculture, industry, and services in GDP, and ϵ is a small constant to avoid division by zero.

A.3 Evaluation Metrics Definitions

Root Mean Squared Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{7}$$

Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{8}$$

R^2 Score (Coefficient of Determination):

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{9}$$

A.4 Software and Reproducibility

Software Versions:

- Python: 3.8+
- pandas: 1.5.0+
- numpy: 1.23.0+
- scikit-learn: 1.2.0+
- xgboost: 1.7.0+
- lightgbm: 3.3.0+
- matplotlib: 3.6.0+
- seaborn: 0.12.0+

Random Seeds: All random processes use seed = 42 for reproducibility.

Data Access: Data can be collected from the World Bank API using the provided data collection scripts in the project repository.

B Detailed Model Algorithms

This section provides mathematical formulations for each algorithm.

B.1 Decision Tree Variance Reduction

At each node, the algorithm selects the feature and split point that maximizes variance reduction:

$$\Delta = \text{Var}(S) - \sum_{i \in \{L,R\}} \frac{|S_i|}{|S|} \text{Var}(S_i) \tag{10}$$

where S is the parent node, S_L and S_R are the left and right child nodes.

B.2 Random Forest Aggregation

The final prediction is the average of all tree predictions:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x) \tag{11}$$

where B is the number of trees and $\hat{f}_b$ is the prediction from tree b.

B.3 Gradient Boosting Sequential Learning

The model minimizes a loss function using gradient descent in function space:

$$F_m(x) = F_{m-1}(x) + \nu \cdot h_m(x) \tag{12}$$

where F_m is the ensemble after m iterations, ν is the learning rate, and h_m is the new tree fitted to the negative gradient of the loss function.

B.4 XGBoost Regularized Objective

The objective function includes both loss and complexity:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{13}$$

where l is the loss function and $\Omega(f_k) = \gamma T + \frac{1}{2}\lambda ||\omega||^2$ penalizes tree complexity (T is the number of leaves, ω are leaf weights).

B.5 Statistical Testing

B.5.1 Paired t-test

Test statistic for comparing model performance:

$$t = \frac{\bar{d}}{s_d / \sqrt{n}} \tag{14}$$

where $\bar{d}$ is the mean pairwise difference and s_d is the standard deviation of differences, with significance level $\alpha = 0.05$.

C Complete Performance Results

C.1 Performance by GINI Range

Table 5 shows model performance across different inequality levels.

Table 5: Model Performance by GINI Range

GINI Range	Model	RMSE	MAE	R^2
Low (0-30)	Decision Tree	4.82	3.56	0.64
	Random Forest	3.71	2.78	0.79
	Gradient Boosting	3.68	2.73	0.80
	XGBoost	3.52	2.61	0.83
	LightGBM	3.45	2.57	0.84
Moderate (30-40)	Decision Tree	5.01	3.82	0.67
	Random Forest	3.92	2.91	0.82
	Gradient Boosting	3.89	2.87	0.83
	XGBoost	3.71	2.73	0.86
	LightGBM	3.63	2.67	0.87
High (40-50)	Decision Tree	5.34	4.01	0.71
	Random Forest	4.21	3.12	0.84
	Gradient Boosting	4.18	3.09	0.85
	XGBoost	3.98	2.94	0.87
	LightGBM	3.89	2.87	0.88
Very High (50+)	Decision Tree	5.67	4.23	0.65
	Random Forest	4.56	3.45	0.78
	Gradient Boosting	4.52	3.41	0.79
	XGBoost	4.21	3.21	0.83
	LightGBM	4.25	3.24	0.82