Advanced Programming 2025

Machine Learning for Global Inequality Analysis : Predicting Gini Coefficients with World Bank Indicators

Final Project Report

Juliette LOUPPE
juliette.louppe@unil.ch
Student ID: 21681960

December 26, 2025

Abstract

This Data Science Project examines the different drivers of wages disparities across countries using the Gini coefficient (a measure of inequality for wealth or income with a value between 0 and 1) and socioeconomic indicators from the World Bank. The primary challenge of this report is to determine whether structural factors, such as participation in the labour market, education, fiscal responsibility, or economic openness, explain better changes in inequality over time and between regions. The methodology used is based on data collection through the World Bank API, followed by a cleaning that includes temporal filtering, interpolation, and imputation of missing values. Following the application of an 80/20 split, three models are contrasted : Linear Regression, Random Forest and XGBoost. The metrics : R^2, RMSE, and MAE are applied to evaluate model performance. Furthermore, different interactions are introduced to capture multiplicative socio-economic effects. The outcomes of various methods show that decision tree models outperform the linear approach, with XGBoost achieving the best performance with an R^2 of 0.938. To close this platform, the analysis of important features confirms the essential role of women's engagement in the workforce by taking the role of classroom instruction, secondary education enrollment and finally labor participation. The project's major goal is to produce a predictive and rigorous model that emphasises the role of educational processes and the employment market in the growth of inequality.

Keywords: inequality, World Bank, workforce, socio-economic indicators, machine learning

Table of Contents

1 Introduction **3**

2 Literature Review **4**

3 Methodology **5**
3.1 Data Description 5
3.2 Approach 5
3.2.1 Linear Regression and model architecture 5
3.2.2 Random Forest Regressor and model architecture 6
3.2.3 XGBoost Regressor and model architecture 6
3.3 Implementation 6

4 Results **7**
4.1 Experimental Setup 7
4.2 Performance Evaluation 8
4.3 Visualisations 9
4.3.1 Real vs. Predicted Gini 9
4.3.2 Feature Importance 9
4.3.3 Inequality Across Countries 10

5 Discussion **11**

6 Conclusion **12**
6.1 Summary 12
6.2 Future Work 12

7 Appendices **14**
7.1 Appendix A: Additional Results 14
7.1.1 A.1 Boxplot of Gini Distribution: United States, France, Brazil 14
7.1.2 A.2 Correlation Heatmap of Numerical Variables 14
7.2 Appendix B: Code Repository 15

1 Introduction

One of the main issues that contemporary societies are currently dealing with is wage inequality. They affect democratic stability, social well-being, and long-term economic growth (OECD, 2023; Piketty, 2014). Despite development efforts, these disparities still exist on a global scale, and it is still challenging to comprehend their dynamics. Many studies show that a variety of structural factors that promote nonlinear behaviour, such as women's labour market participation, education, fiscality, and economic openness, are what drive inequality (Berg and Ostry, 2017; Goldberg and Pavcnik, 2007). In this context, having tools that can identify and predict what causes these issues is essential for making public policies more understandable.

The primary challenge addressed in this project is the following: which socioeconomic factors most effectively explain changes in the Gini coefficient over time and between countries, and to what extent can these levels of inequality be predicted using machine learning models ? Our analysis is based on a wide range of data from the World Bank's API, including factors related to women's status in the labour market, education, growth, fiscality, and commercial openness.

The principal objective is twofold:

- Create a reliable predictive model of the Gini coefficient through contrasting supervised models (Linear Regression, Random Forest, and XGBoost).
- Identify the primary determinants of inequality empirically, for example by analysing the significance of the variables in nonlinear models.

The structure of the report is as follows: the second section is a review of the literature; the third section describes the dataset's cleaning process and the models used; the fourth section presents the results of the different approaches; the fifth section discusses the implications of these results in light of the literature; and the final section offers conclusions and future directions for improvement.

2 Literature Review

Previous research on economic inequality has mostly relied on traditional economic models, emphasizing links between education, labour markets, and income distribution. [Chani et al., 2014] show that improvements in human capital reduce inequality in developing countries. Berg and Ostry (2017) find that high inequality shortens growth periods, highlighting a reciprocal relationship between inequality and economic performance. Regarding globalisation, [Goldberg and Pavcnik, 2007] argue that commercial openness can either increase or decrease inequality depending on institutional frameworks. These results underline the importance of structural dynamics in shaping inequality.

Traditional analyses often use linear or parametric models with fixed functional forms, even though inequality typically arises from complex and nonlinear interactions involving education, fiscality, urbanisation, or labour market participation. Consequently, several recent studies recommend more flexible approaches such as decision trees or ensemble models that can capture heterogeneous interactions without strong assumptions. Random Forests and boosting methods, in particular, have proven effective for modelling socio-economic relationships ([Breiman, 2001, Chen and Guestrin, 2016]). This project aligns with this methodological evolution by comparing XGBoost, Random Forest, and linear regression.

Empirical studies frequently rely on international datasets such as the World Bank's WDI indicators, OECD statistics, the WIID inequality database, and the Luxembourg Income Study. These sources support long-term cross-country comparisons of inequality. While some studies focus on microeconomic trade effects ([Goldberg and Pavcnik, 2007]) or historical income dynamics ([Piketty, 2014]), this project uses a harmonised dataset built from the World Bank's API, providing consistent indicators for many countries between 1995 and 2020.

Despite extensive prior research, two limitations remain. First, most empirical studies rely on linear models, although inequality drivers often interact nonlinearly (e.g., education $\times$ labour participation). Our project addresses this by comparing flexible machine learning models. Second, many analyses concentrate on a single country or region. In contrast, our dataset spans over 25 years and multiple nations, enabling a global evaluation of which socio-economic factors most strongly influence inequality.

Overall, the literature highlights structural links between education, labour markets, fiscality, globalisation, and inequality, but relatively few studies apply machine learning to model these relationships at a large international scale. This project contributes to filling this gap by using nonlinear models on globally harmonised data.

3 Methodology

3.1 Data Description

The dataset used in this project is fully sourced from the World Bank API. Ten socioeconomic indicators were collected across domains such as education, labour markets, fiscality, growth, urbanisation, and international trade. The raw panel contains 9044 rows and 13 variables structured by country, country_code, and year.

After cleaning, the analytical dataset includes:

- 1528 observations,
- 13 variables (1 target: Gini, 12 numerical predictors).

The Gini coefficient shows:

- large variation across countries (24–60),
- moderate temporal variability,
- several extreme values.

Key explanatory indicators include female labour participation, labour participation, secondary education enrolment, tax revenue, trade openness, GDP per capita, urbanisation, population, and GDP growth.

Before cleaning, the dataset contained temporal gaps, missing values, and irregular coverage. The final dataset was obtained through:

1. filtering countries with at least ten Gini points,
2. restricting the time period to 1995–2020,
3. removing rows without Gini,
4. interpolating numeric variables by country,
5. median imputation for recent missing values,
6. ensuring country-year coherence.

These steps produced a complete and coherent dataset suitable for machine learning.

3.2 Approach

Three models were implemented to predict the Gini coefficient:

3.2.1 Linear Regression and model architecture

A baseline model capturing direct linear relationships. Simple and interpretable, but limited for nonlinear dynamics. The model architecture is equivalent to a simple linear function that combines the input features with their learnt coefficients :

$$\hat{y} = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$$

3.2.2 Random Forest Regressor and model architecture

An ensemble of decision trees capable of capturing nonlinearities and interactions. It provides robust and stable predictions. Its model architecture is based on a collection of decision trees, each trained on subsampled data to increase robustness and reduce volatility :

- 300 trees,
- unlimited depth,
- random sampling of features and observations.

3.2.3 XGBoost Regressor and model architecture

A gradient boosting algorithm effective for structured data. By correcting residual errors sequentially, it models complex patterns and generally outperforms simpler approaches. The model design is based on consecutively added decision trees, each of which corrects the faults of the preceding ones via gradient boosting :

- 400 trees,
- depth = 4,
- learning rate = 0.05,
- subsample = 0.8, colsample_bytree = 0.8.

Performance was evaluated using:

- R^2: variance explained,
- RMSE: penalises large errors,
- MAE: average absolute error.

The optimal model maximises R^2 while minimising RMSE and MAE.

3.3 Implementation

The project is implemented in Python using `pandas`, `requests`, `scikit-learn`, `xgboost`, and `matplotlib`. The codebase is modular:

- `data_loader.py`: API extraction,
- `preprocessing.py`: cleaning and interpolation,
- `split_data.py`: train/test split,
- `models.py`: model definitions and training,
- `notebooks/`: exploration and visualisation.

Pipeline:

API → Cleaning → Train/Test Split → Models → Comparison → Visualisation

Below is the function used to compare the three models:

```python
# Models Comparison
def compare_models(X_train, X_test, y_train, y_test):
    RESULTS_DIR.mkdir(exist_ok=True)

    res_lin = run_linear_regression(X_train, X_test, y_train, y_test)
    res_rf = run_random_forest(X_train, X_test, y_train, y_test)
    res_xgb = run_xgboost(X_train, X_test, y_train, y_test)

    df_results = pd.DataFrame([res_lin, res_rf, res_xgb])

    # Round metrics
    df_results = df_results.round(4)

    # Filter by performance
    if "r2" in df_results.columns:
        df_results = df_results.sort_values("r2", ascending=False).reset_index(drop=True)

    print("\n==================== Comparative table ====================\n")
    print(df_results)

    out_path = RESULTS_DIR / "model_comparison.csv"
    df_results.to_csv(out_path, index=False)
    print(f"\nComparison table saved in: {out_path}")

    return df_results
```

Figure 1: Function `compare_models`: training and evaluation of all models.

4 Results

4.1 Experimental Setup

Given the moderate size of the dataset (1528 rows) and the relatively lightweight nature of the models, all experiments were executed on a standard laptop CPU, which is fully sufficient for this workflow. The implementation relies on Python 3.11 with the following libraries: `pandas`, `scikit-learn`, `xgboost`, and `matplotlib`/`seaborn`.

The dataset was split into 80% training data (1222 samples) and 20% testing data (306 samples). The models were trained using simple yet reliable hyperparameters: depth 4 and a learning rate of 0.05 for XGBoost, 300 trees for the Random Forest, and 400 boosting iterations for XGBoost. Performance was evaluated with three complementary metrics: R^2, RMSE, and MAE. In contrast, tree-based models demonstrate a significant performance improvement. The Random Forest model outperforms the linear baseline by almost 30 points, with an R^2 of 0.9287. This highlights the need of automatically simulating nonlinear connections and variable interactions, rather than relying on manual feature engineering. The significant reduction in both RMSE and MAE reinforces the robustness of this technique and its ability to generalise effectively.

4.2 Performance Evaluation

Table 1 summarises the predictive performance of the four models assessed in this study. The measures R^2, RMSE, and MAE provide complementary viewpoints on model quality. R^2 measures the amount of explained variation, while RMSE and MAE quantify prediction mistakes. RMSE penalises big deviations more strongly.

Model	R^2	**RMSE**	**MAE**
Linear Regression	0.6066	5.4935	4.3785
Linear Regression + Interactions	0.6245	5.3671	4.3329
Random Forest	0.9287	2.3381	1.6117
XGBoost	0.9383	2.1749	1.6435

Table 1: Model performance comparison.

A clear performance hierarchy emerges from the various modelling methodologies. Linear regression serves as a baseline, explaining roughly 60% of the Gini coefficient's variability. Despite its rather steady forecasts, the high RMSE and MAE values indicate significant systematic errors. Adding interaction variables marginally enhances the model's explanatory ability, but the improvements are limited, implying that linear structures, even when enriched, fail to reflect the underlying nonlinearities found in socioeconomic inequality data.

In contrast, tree-based models demonstrate a significant performance improvement. The Random Forest model outperforms the linear baseline by almost 30 points, with an R^2 of 0.9287. This highlights the need of automatically simulating nonlinear connections and variable interactions, rather than relying on manual feature engineering. The significant reduction in both RMSE and MAE reinforces the robustness of this technique and its ability to generalise effectively.

XGBoost is the best-performing model, with an R^2 of 0.9383 and the lowest error metrics. XGBoost's gradient-boosting method iteratively corrects past prediction errors, resulting in higher accuracy, particularly in datasets containing complicated interactions among predictors. Its lower RMSE indicates less major deviations from genuine values, and its lowest MAE indicates consistently accurate predictions across the whole sample.

The results show a consistent pattern: linear models underfit the data due to their restricted structure. Random Forests can detect nonlinearity and significantly improve accuracy. XGBoost makes the most dependable and accurate forecasts, with the greatest explanatory power and the fewest prediction mistakes. These findings support the idea that inequality trends are caused by complex, interconnected socioeconomic processes that can be better simulated using flexible, nonparametric machine-learning methods.

4.3 Visualisations

4.3.1 Real vs. Predicted Gini

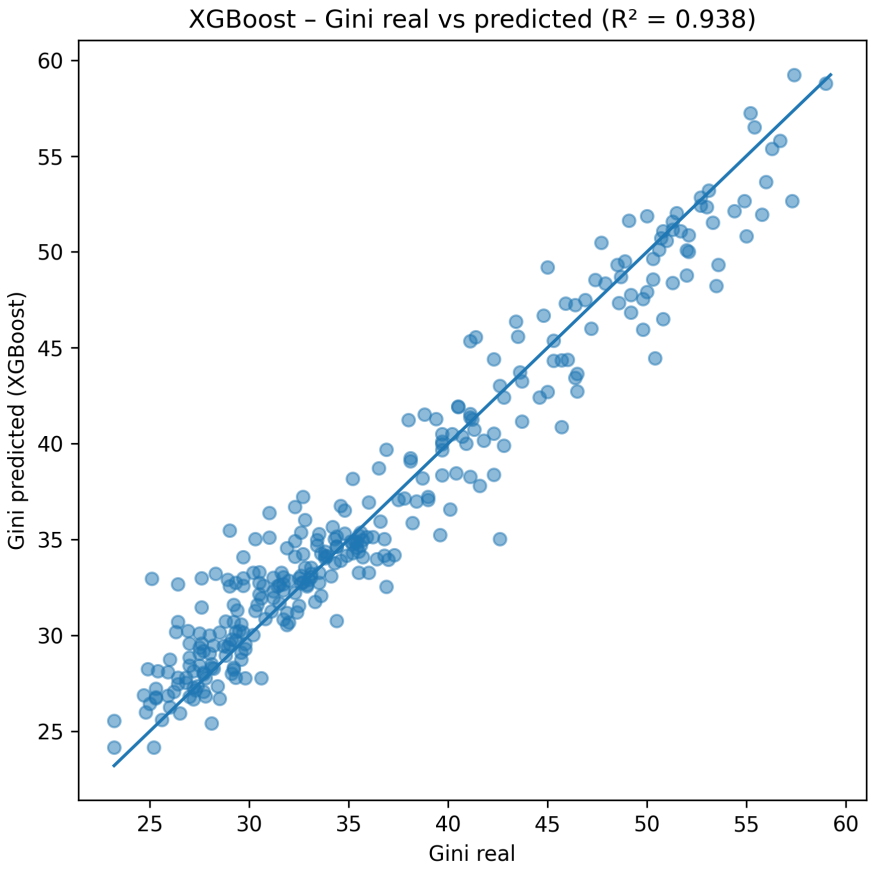

Figure 2: Real vs. predicted Gini values on the test set using XGBoost.

Figure 2 compares the true Gini values to the predictions produced by XGBoost. Each point represents a country-year observation. The closer the points lie to the 45° diagonal, the more accurate the prediction. The cloud of points aligns closely with the diagonal across the entire range (from low inequality around 25–30 to highly unequal contexts above 50).

This visual coherence confirms the quantitative performance metrics: the model captures both moderate and extreme inequality levels, a crucial property for socio-economic forecasting. High predictive accuracy (test $R^2 = 0.938$, RMSE = 2.17) shows that XGBoost effectively learns structural relationships between education, labour participation, fiscal capacity and inequality.

4.3.2 Feature Importance

Figure 3 displays the relative importance of the predictors used by XGBoost. The three most influential variables : female labour participation, secondary education enrolment, and overall labour participation are fully consistent with empirical literature. Higher labour force participation, especially among women, is strongly associated with lower inequality (OECD, 2023; World Bank, 2022). Similarly, secondary education is a well-established driver of human capital formation and long-term reductions in income disparity.

Tax revenue and trade openness also contribute significantly, echoing findings that fiscal capacity and global integration shape how growth is redistributed (Berg & Ostry, 2017; Goldberg

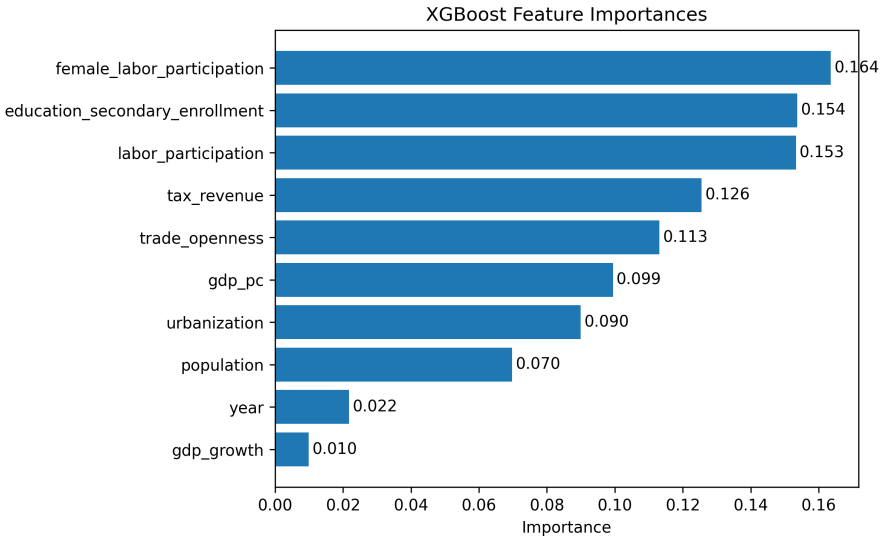

Figure 3: Feature importances for the XGBoost model.

& Pavcnik, 2007). By contrast, GDP growth and year contribute marginally, suggesting that short-term macroeconomic fluctuations alone cannot explain structural inequality patterns.

Overall, the feature importance analysis confirms that the model focuses on long-term structural determinants of inequality.

4.3.3 Inequality Across Countries

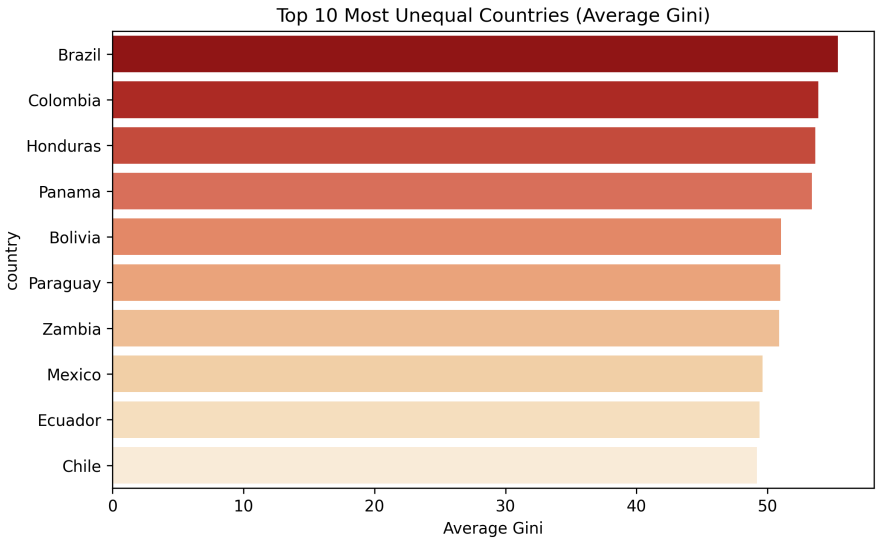

Figure 4: Top 10 most unequal countries in the dataset (average Gini, 1995–2020).

Figure 4 presents the ten countries with the highest average Gini index between 1995 and 2020. Latin American countries dominate the ranking, consistent with extensive literature identifying persistent educational disparities, high labour informality, and limited fiscal redistribution as core drivers of inequality in the region ([Lustig, 2020, Gasparini and Cruces, 2010]).

Zambia's position reflects similar dynamics observed in Sub-Saharan Africa, where inequality remains structurally high despite growth episodes, partly due to constrained fiscal space and unequal access to human capital ([Fund, 2017, Bank, 2020]).

These data support the model's prediction that countries with the most unequal outcomes have systematic inadequacies in the structural variables indicated by the model, namely education, labour participation, and fiscal capacity. As such, Figure 4 provides external validation of the model's economic interpretability.

5 Discussion

The use of nonlinear ensemble methods was extremely beneficial. Random Forest and XGBoost demonstrated good predictive performance ($R^2 \approx 0.938$), indicating its applicability for simulating complicated socioeconomic interactions ([Breiman, 2001, Chen and Guestrin, 2016]). The feature-importance analysis confirms previous empirical research: female labor-force participation and secondary education, both widely recognised as major determinants of inequality dynamics, emerged as the strongest predictors of decreasing inequality ([OECD, 2023, Chani et al., 2014]).

Beyond prediction accuracy, these findings have broader consequences. First, it appears that inequality is caused by interacting and nonlinear mechanisms rather than discrete socioeconomic causes. Linear models could only explain around 60% of the variation in Gini coefficients and constantly showed large error rates. Their failure to capture multiplicative effects (for example, the combined contribution of education and labor-market inclusion) highlights a well-documented shortcoming of parametric models in the inequality field. In contrast, ensemble methods automatically detect cross-variable interactions without the need for explicit definition, making them ideal for diverse global datasets.

Methodological issues also arose during the analysis. The World Bank indicators show significant missingness, irregular update frequency, and structural alterations over time. Ensuring temporal consistency necessitated meticulous imputation and filtration, and the restricted coverage of certain indicators may have produced bias in nations with weak statistical systems. Similarly, high cross-country heterogeneity, ranging from affluent economies to fragile states, made it challenging for models that assumed homogenous data-generation processes. These concerns highlight the significance of robust preprocessing pipelines and adaptable modelling architectures that can handle structural heterogeneity.

Overall, the findings were consistent with expectations: tree-based techniques beat linear baselines, demonstrating that inequality is shaped by linked demographic, institutional, and macroeconomic processes. The model's forecasts emphasised education, labor-force participation, and fiscal capability, which are consistent with decades of actual evidence. Interestingly, GDP growth, traditionally regarded as a key cause of inequality, had no predictive relevance in our models, mirroring recent literature that questions its explanatory ability ([Piketty, 2014]). In contrast, female labour participation regularly appeared as the leading predictor, highlighting the importance of gender engagement in affecting income distribution ([Bank, 2022]).

6 Conclusion

6.1 Summary

The goal of this study was to anticipate and assess worldwide income disparity from 1995 to 2020 using machine learning approaches applied to World Bank data. The project's goal was to reconstruct a harmonised international dataset, identify the most relevant socioeconomic factors of inequality, and create an accurate predictive model capable of capturing complicated cross-country dynamics.

These objectives were met. XGBoost's strong predictive accuracy ($R^2 = 0.938$) confirms the efficiency of nonlinear ensemble methods in modelling socio-economic systems with interconnections, heterogeneity, and structural complexity. The feature-importance analysis also found that female labor-force involvement, secondary education enrolment, and overall labor-market engagement were consistently among the most powerful predictors of inequality. These findings are consistent with previous empirical studies, highlighting the importance of demographic and educational characteristics as significant drivers of fair development.

Beyond prediction, the project offers an organised, reproducible approach for cleaning, rebuilding, and modelling multi-country panel data. This involves automated data retrieval, rigorous preprocessing, clear train-test design, and unified model comparison. Such a pipeline can help researchers and policymakers better understand global inequality patterns, particularly in cases when official statistics are irregular, partial, or conflicting.

Finally, the study demonstrates the ability of new machine learning technologies to supplement traditional economic analysis. These tools provide deeper insights into the mechanisms influencing inequality by capturing nonlinear interactions that linear models miss, opening the door to more focused and evidence-based policy solutions.

6.2 Future Work

Several extensions could enhance the robustness and scope of this work. First, more systematic hyperparameter optimisation such as grid search or cross-validation could improve model stability and predictive performance. Second, integrating additional variables (e.g., institutional quality, tax progressiveness, regional indicators) would offer a more comprehensive understanding of inequality mechanisms.

Future research may also explore causal inference techniques to move beyond correlation-based analysis, as well as alternative machine learning methods such as CatBoost, LightGBM, or neural networks. Finally, developing a simple policy simulation tool could help assess how changes in education, employment participation, or fiscal capacity would influence inequality in real-world scenarios.

References

Bank, W. (2020). Poverty and equity brief: Sub-saharan africa. https://www.worldbank.org/en/topic/poverty/publication/poverty-and-equity-briefs.

Bank, W. (2022). Gender employment gap report. https://www3.weforum.org/docs/WEF_GGGR_2023.pdf. World Bank / WEF report.

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

Chani, M. I., Chaudhary, A. R., Pervaiz, Z., and Farooq, M. (2014). Human capital inequality and economic growth: Evidence from developing countries. *IMF Economic Review*, 65(4).

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

Fund, I. M. (2017). Inequality in africa: Trends, drivers, and policy remedies. https://www.elibrary.imf.org/view/journals/001/2017/076/article-A001-en.xml. IMF Working Paper 17/302.

Gasparini, L. and Cruces, G. (2010). A distribution in motion: The case of latin america. https://ideas.repec.org/p/dls/wpaper/0078.html.

Goldberg, P. K. and Pavcnik, N. (2007). Distributional effects of globalization in developing countries. *Journal of Economic Literature*, 45(1):39–82.

Lustig, N. (2020). Inequality in latin america: Evidence and explanations. https://ideas.repec.org/p/tul/ceqwps/94.html.

OECD (2023). Gender equality. https://www.oecd.org/en/topics/gender-equality.html. OECD report.

Piketty, T. (2014). *Capital in the Twenty-First Century*. Harvard University Press.

7 Appendices

7.1 Appendix A: Additional Results

This appendix presents supplementary visual analyses supporting the main findings. While not required for the core modelling pipeline, these figures provide important context for understanding global inequality patterns and interactions within the dataset.

7.1.1 A.1 Boxplot of Gini Distribution: United States, France, Brazil

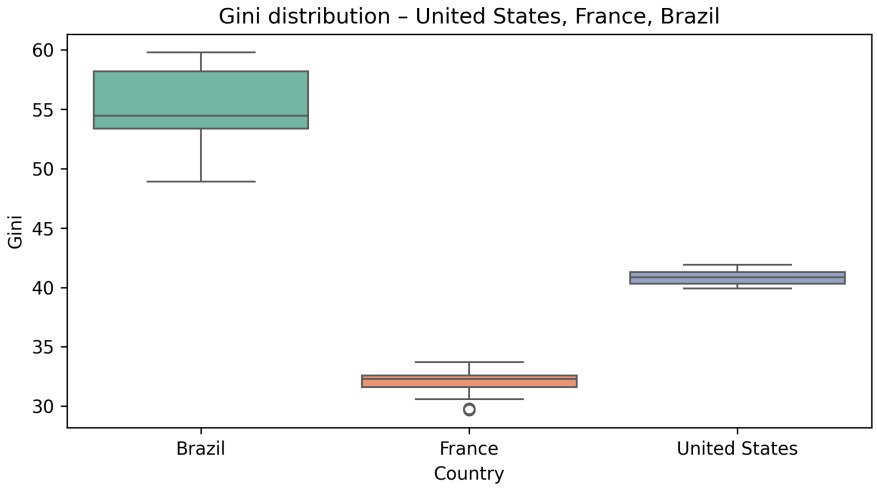

Figure 5: Gini distribution for the United States, France, and Brazil (1995–2020).

Figure 5 illustrates the distribution of Gini values for three representative countries. Brazil shows the highest and most dispersed inequality levels, consistent with its well-documented structural disparities (Lustig2020). France displays lower and more stable inequality, while the United States lies in between but with greater variability. These differences highlight why predictive models must account for non-linear and heterogeneous cross-country dynamics.

7.1.2 A.2 Correlation Heatmap of Numerical Variables

Figure 6 shows the correlations among key socioeconomic indicators. Several expected patterns emerge:

- Education and labour participation are positively correlated.
- Tax revenue is negatively associated with Gini, consistent with redistributive theory.
- GDP per capita correlates with trade openness and tax revenue.
- GDP growth shows weak correlations overall, consistent with model findings that it has low predictive relevance.

These patterns reinforce the idea that inequality is driven mainly by structural variables, not short-term macroeconomic fluctuations.

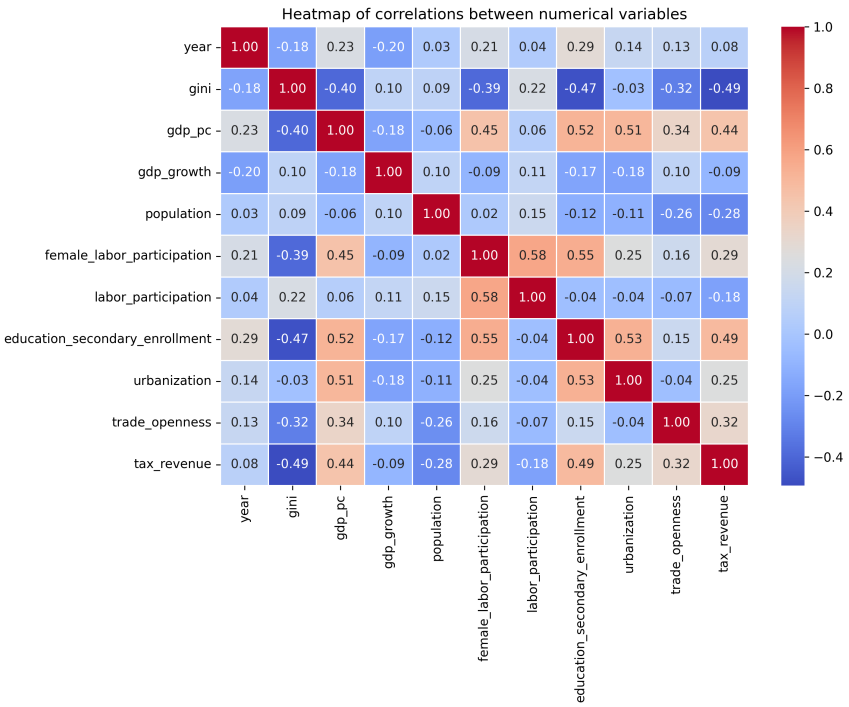

Figure 6: Correlation matrix of all numerical variables in the dataset.

7.2 Appendix B: Code Repository

GitHub Repository: `https://github.com/juliettelouppe/Gini-index-prediction.git`

Repository Structure

```
gini-index-prediction/
  README.md
  project_report.pdf
  requirements.txt
  PROPOSAL.md
  main.py
  .gitignore
  data/
      wb_inequality_data.xlsx
      inequality_clean.xlsx
  notebooks/
      01-data-loading.ipynb
      02-eda.ipynb
      03-modeling.ipynb
      04-results.ipynb
  results/
      boxplot_gini_us_fr_br.png
      heatmap_correlations.png
      model_comparison.csv
      rf_feature_importances.csv
      top10_inequality.png
      xgb_feature_importances.csv
      xgb_feature_importances_plot.png
```

```
      xgb_true_vs_pred.png
      X_train.csv
      X_test.csv
      y_train.csv
      y_test.csv
  src/
      __init__.py
      data_loader.py
      preprocessing.py
      split_data.py
      models.py
```

Installation Instructions

```
python -m venv .venv
source .venv/bin/activate  # Mac/Linux
.venv\Scripts\activate     # Windows
pip install -r requirements.txt
```

Reproducing Results

```
python src/data_loader.py
python src/preprocessing.py
python -m src.split_data
python -m src.models
python main.py
```

Use of AI Assistance

Throughout the project, ChatGPT was used as a supplementary tool to assist with numerous non-essential areas of the workflow. It helped with proofreading and correcting minor language errors, paraphrasing certain passages to improve academic clarity and coherence, debugging specific sections of code when encountering implementation issues and understanding the practical aspects of using GitHub and Overleaf for version control and document preparation.