U.PORTO
FEP FACULDADE DE ECONOMIA
UNIVERSIDADE DO PORTO

What are the most relevant variables for predicting income inequality?: a machine learning approach

Alex Francisco Fernandes Alves

Dissertation
MSc in Data Analytics

Supervised by
Prof. Jorge Valente, Ph. D
Prof. Sandra Silva, Ph. D

2023

Acknowledgments

I would like to thank especially professors Jorge Valente and Sandra Silva who promptly and thoroughly helped me along the way.

On a personal level, I would like to express unbounded gratitude to my friends and family.

Finally, a special acknowledgment to my girlfriend, Carla, who provided me with the emotional stability and structure needed to accomplish this academic endeavour.

Resumo

Nos últimos anos, tem havido alguns indícios de aumento da desigualdade de rendimentos em todo o mundo. Por exemplo, Alvaredo et al. (2018) sugerem que a desigualdade está a aumentar em todo o mundo. Os autores utilizam a "curva do elefante" para representar visualmente a forma como o crescimento do rendimento tem sido distribuído pelos diferentes percentis de rendimento. A curva mostra que, embora tenha havido um crescimento significativo do rendimento para o percentil superior extremo e para a classe média baixa (aproximadamente os percentis 70 a 80) e para a classe média emergente (cerca do percentil 40), a classe média nos países ocidentais desenvolvidos (cerca do percentil 80) registou um crescimento do rendimento relativamente modesto.

Neste trabalho, tomando dados do Banco Mundial e de outras bases de dados, o nosso objetivo foi estimar/prever o Índice de Gini (regressão), aplicando modelos de Machine Learning. O segundo objetivo foi entender quais são os fatores que mais contribuem para a desigualdade de rendimento num determinado país, i. e., avaliamos que variável ou variáveis independentes mais contribuem para a previsão do índice de desigualdade de rendimento (Gini).

Depois de otimizar os parâmetros para cada modelo, comparámos os resultados e o desempenho. Concluímos que o Gradient Boosting é o melhor modelo com um MAPE de 13% e uma pontuação r^2 de 32%. A variável "Crescimento da população (% anual) – 2019" com um desfasamento de 0 é a variável com maior valor preditivo. Em segundo e terceiro lugares em termos de importância estão as variáveis "Esperança de vida à nascença, total (anos)" e "Educação obrigatória, duração (anos)". Ao contrário do "Crescimento da população", nestas duas variáveis, valores mais elevados indicam valores de Gini mais baixos (menor desigualdade).

Abstract

There has been some evidence of rising income inequality across the globe in recent years. For instance, Alvaredo et al. (2018) suggest that inequality is growing across the globe. The authors use the "elephant curve" to visually represent how income growth has been distributed among different income percentiles. The curve shows that while there has been significant income growth for the extreme top percentile and the lower middle class (approximately 70th to 80th percentiles) and emerging middle class (around the 40th percentile), the middle class in developed Western countries (around the 80th percentile) has experienced relatively modest income growth.

In this research, taking data from the World Bank and other databases, our objective was to estimate/predict the Gini Index (regression) by applying Machine Learning models. The second objective was to understand which factors contribute the most to income inequality in each country, i.e., we assessed which independent variable or variables contribute most to the prediction of the income inequality index (Gini).

After optimizing the parameters for each model, we compared the results and performance. We concluded that Gradient Boosting is the best model with MAPE of 13% and r2 score of 32%. The variable 'Population growth (annual %) – 2019' with a lag of 0 is the feature with the most predictive value. In second and third place in terms of importance are the variables 'Life expectancy at birth, total (years)' and 'Compulsory education, duration (years)'. Unlike 'Population growth', in these two variables, higher values indicate lower Gini values (lower inequality).

Contents

List of Tables

List of Figures

1. Introduction

There has been some evidence of rising income inequality across the globe in recent years. For instance, Alvaredo et al. (2018) suggest that inequality is growing across the globe. The authors use the "elephant curve" to visually represent how income growth has been distributed among different income percentiles. The curve shows that while there has been significant income growth for the extreme top percentile and the lower middle class (approximately 70^{th} to 80^{th} percentiles) and emerging middle class (around the 40^{th} percentile), the middle class in developed Western countries (around the 80^{th} percentile) has experienced relatively modest income growth.

This pattern indicates that income growth has been unevenly distributed, with the wealthiest individuals and the emerging middle class in developing nations benefiting the most, while the middle class in developed Western countries has experienced slower growth. This unequal distribution of income growth contributes to the perception of increasing global inequality, as the gains from economic expansion and globalization are not being equally shared across different income groups and regions.

Due to increasing concerns and awareness about this socioeconomic issue, I propose to address this interesting and complex topic.

Inequality corresponds to a significant and persistent difference in the distribution of resources, opportunities, or outcomes among different individuals or groups in a society. This can be caused by various social, economic, and political factors, including discrimination, unequal access to education, healthcare, and employment, and disparities in political power.

Inequality can manifest in many different forms. There are differences in the distribution of income, wealth (e.g., assets, savings, investments), and access to education or educational outcomes. There are as well differences in opportunities, outcomes, and treatment among different racial or ethnic groups or between men and women (gender).

While each of these types of inequality is important to study, income inequality is often considered a particularly pressing issue because of its far-reaching consequences.

One consequence is that it can lead to social and political instability (Acemoglu and Robinson (2000)). This is because when a small percentage of the population controls a disproportionate amount of wealth and resources, it can create resentment and social tensions, which in turn can lead to political unrest and instability.

Another consequence is that it can negatively affect economic growth (Aghion et al. (1999)). This is because when income is concentrated in the hands of a few, there is less money available for spending and investing by the middle and lower classes, which are the largest consumers and investors in most economies. This can lead to reduced demand and economic growth.

Income inequality can also have negative effects on health outcomes (Wilkinson and Pickett (2006)). This is because people with lower incomes may have less access to healthcare, healthy food, and other resources that promote good health. In addition, the stress and anxiety that often accompany poverty and inequality can contribute to poor health outcomes.

Finally, income inequality can have negative effects on social mobility and opportunity (Corak (2013)). When wealth and resources are concentrated in the hands of a few, it can be more difficult for individuals from disadvantaged backgrounds to move up the social and economic ladder. This can limit opportunities for social and economic advancement and perpetuate cycles of poverty.

1.2 Problem Formulation and Objectives

We propose to estimate the continuous variable Gini Index (between 0 and 1) given the GDP per capita of that given country, foreign investment level, educational attainment by most of the population, life expectancy, and the level of indebtedness of the government.

Currently, we already have at our disposal a plethora of data analytics tools to help us explore this subject using, for instance, machine learning models. Though there are already several articles applying those methods to economic subjects such as income inequality, we aim to explore new angles (e. g., new variables and different lags) and potential new approaches (e. g., new models with new parameters).

The research process required to produce this dissertation aims fundamentally to answer the following question: What are the most relevant variables for predicting income inequality?

1.3 Structure

This dissertation will be divided into the following sections.

The Literature Review will systematize relevant academic papers and books addressing income inequality. Here, we cover frameworks and concepts used, and variables considered.

The Exploratory Data Analysis and Dataset section aims to introduce the data set we will use to further explore the research topic, answer the research question, and describe the main patterns found in the data.

The Models section will cover the literature on machine learning models applied to the pre-processed dataset, corresponding performance measures, and comparison between best-performing models.

In Results section, the results from all the models are described in detail.

Finally, Conclusions section will recap the original research question and frame the results obtained. Besides this, we will enumerate all the limitations and opportunities for future research.

2. Literature Review

In the following literature review, we went through the existing research on income inequality in general and using a Machine Learning approach. Income inequality is a very well-known and explored topic in the economics literature. The relationship between income inequality and economic growth is complex and multi-directional. Specifically, income inequality can affect growth, and growth can, in turn, affect income inequality.

Regarding the effect of income inequality on growth, Neves and Silva (2014) note that high levels of inequality can lead to social and political instability, which can negatively affect economic growth. In addition, inequality can lead to limited access to education, healthcare, and other resources that are important for human capital formation and productivity growth, further hindering economic growth.

On the other hand, the authors also acknowledge that growth can affect income inequality. Economic growth can reduce poverty and increase access to education and other resources, which can help to reduce income inequality. However, the authors also note that the distributional effects of growth depend on a variety of factors, including the institutional context and the policies that are implemented to promote inclusive growth.

2.1 Income Inequality Conceptualization and Framework of Analysis

2.1.1 Definition and recent approaches

As described by the OECD (2023), "income is defined as household disposable income in a particular year. It consists of earnings, self-employment and capital income and public cash transfers; income taxes and social security contributions paid by households are deducted." There are several indicators we can take into consideration to measure income inequality; however, we chose the most well-known and documented – the Gini coefficient.

The Gini index is based on the comparison of cumulative proportions of the population against cumulative proportions of income they receive, and it ranges between 0 in the case of perfect equality and 1 in the case of perfect inequality.

Inequality could be decomposed as effort and opportunity inequality. As Salas-Rojo and Rodríguez (2022, p. 28) stated "any economic outcome such as wealth, income or health status is the result of the interaction between two sets of factors."

On one side, we face exogenous factors beyond an individual's control, such as sex, parental education, race, or the inheritances received (circumstances). On the other hand, the remaining factors are endogenous, as they are within the individual agency. It is the case, for instance, of the work ethic or nutritional habits (consciously exerted efforts).

Across the existing literature, several authors studied what might cause inequality at an individual and/or collective (country-wise) level.

Bowles and Gintis (2002, p. 4) studied to what extent the intergenerational transmission of economic status contributes to income and wealth inequality. They did not focus on a specific set of countries or a particular historical period. Instead, the book draws on a wide range of empirical studies and theoretical perspectives from various countries and periods to explore the factors that contribute to the persistence of economic and social inequality across generations.

The authors draw on evidence from studies conducted in many countries, including the United States, Canada, Sweden, Germany, the United Kingdom, and Japan, among others. They also draw on historical and cross-national data to illustrate the long-term trends and patterns of inequality in different societies.

Such transmission is done through "a heterogeneous collection of mechanisms, including the genetic and cultural transmission of cognitive skills and noncognitive personality traits in demand by employers, the inheritance of wealth and income-enhancing group memberships, such as race, and the superior education and health status enjoyed by the children of higher status families".

In parallel Salas-Rojo and Rodríguez (2022) applied a Random Forest Machine Learning model to analyse to what extent the wealth inequality was due to inequality of opportunity (Iop). According to these authors, Iop explains over 60% of wealth inequality in the US and Spain (using the Gini coefficient), and more than 40% in Italy and Canada.

Country-wise, Hailemariam et al. (2021) investigate the major factors that drive income inequality in the OECD countries (1870 to 2016). They also show that the real

interest rate and government spending are negatively and significantly associated with income inequality.

On the other hand, an increase in real GDP per capita leads to an increase in income inequality, measured by the Gini coefficient, whereas an advance in financial development reduces it. Positive innovation shocks impact negatively (decrease) income inequality only in the short term. Finally, educational attainment significantly reduces top-income inequality.

Although we can point out several potential causes or consequences of income inequality, the most studied relationship is between income inequality and economic growth.

Neves and Silva (2014) reviewed the relevant theoretical literature on how inequality affects growth. Even though in this dissertation we'll explore the effects of economic growth and other socioeconomic factors on income inequality, it can be assumed that the transmission channels are similar.

The authors identify four main transmission channels: the credit market imperfections, the fiscal policy, the socio-political instability, and savings.

Borrowing constraints limit poor people from investing in human capital (formal education and skill acquisition) and physical capital (for entrepreneurial ventures). This credit market imperfection leads to the misallocation of talent into the different occupational choices available in the marketplace and suboptimal levels of investments due to the high fixed costs.

More significant influence of the rich in politics (through lobbying, campaign contributions, greater propensity to vote, and so forth) implies that political decision-making is biased towards and benefits primarily the well-off. Ultimately, policies such as deregulation of the financial markets or inexistent overseeing of offshores benefit the wealthy, increasing inequality.

On the other hand, inequality generates political instability, which in turn negatively affects investment and future opportunities for growth and prosperity.

A classical view on savings holds that the marginal propensity of the rich to save is higher than that of the poor. In that sense, as inequality increases, the share of resources held by a few individuals whose propensity to save and invest is higher also increases (their immediate needs are already met).

As stated before, most existing literature either studies how inequality affects growth or how growth influences inequality. However, Huang et al. (2009) propose a new methodology where both directions of the relationship inequality-growth are considered. The authors indeed confirmed how highly inter-related inequality and growth are. Specifically, while the impact of inequality on growth is negative, the influence of growth on inequality is positive. When limiting the scope (OECD countries), the causal links become irrelevant. This suggests that the causal interrelationship between growth and inequality may vary with the stage of economic development.

Indeed, Kuznets (1955) analysed data from various countries over several decades and found that as a country develops economically, income inequality tends first to increase and then decrease over time.

The author argued that in the early stages of development, a country typically experiences a shift from agriculture to industry and sees the rise of a small group of wealthy entrepreneurs and business owners who reap the benefits of industrialization. This leads to a widening income gap between the rich and the poor. However, as the country develops further and education and technology become more widely available, income inequality begins to decrease as more people have access to better-paying jobs and can improve their economic prospects.

Kuznets' theory has been widely debated and challenged since it was first proposed. Some critics argue that the relationship between economic development and income inequality is not necessarily linear and that other factors, such as political and institutional structures, can also play a significant role. However, Kuznets' work remains an important foundation for understanding the complex relationship between economic growth and income inequality.

Another perspective is brought by Komatsu and Suzuki (2022). The current study examines the relationship between income inequality and the subjective well-being (SWB) of the Chinese population. The authors measure income inequality using three indicators that capture income inequality at the group level (identity-related income inequality), regional level, and urban-rural level.

In the early stages of economic development, it was found that an equal playing field is more relevant in fostering equality than redistributive policies. If people feel that everyone

has the chance of becoming rich, there will be stronger incentives to work hard and derive fulfilment from their results. Thus, people will be more tolerant of income inequality.

However, on the other hand, the authors found that all measures of income inequality have significant negative effects on SWB either in the initial or latter stages because most of the existing inequalities come from exogenous factors such as institutional barriers and urban–rural segmentation policies.

In the last two decades, across developing and developed countries, inequality has been rising mainly due to technological change, as mentioned by Jaumotte et al. (2013, p. 302). The globalization phenomena have played a minor role. This reflects two offsetting effects of globalization: while increased trade tends to reduce income inequality, foreign direct investment tends to exacerbate it. Financial globalization and technological progress tend to increase the demand for skills and formal education. According to the authors, "while incomes have increased across all segments of the population in all countries in the sample, incomes of those who already have higher levels of education and skills have risen disproportionately more."

Nevertheless, income inequality is closely linked to other development indicators such as health, education, and political/civil participation.

Truesdale and Jencks (2016) investigate the link between income inequality and health outcomes. The authors explore existing research that suggests high levels of income inequality led to adverse health outcomes, including obesity, cardiovascular diseases, and mental health problems. They argue that income inequality's negative effects are not restricted to individuals living in poverty, and it affects people across income spectrums, especially vulnerable populations. The authors suggest that reducing income inequality policies can have significant health benefits, particularly for disadvantaged groups, and thus addressing income inequality is critical to improving population health.

On the other hand, Reardon (2013) examines the relationship between income inequality and educational outcomes in the United States. He finds that higher levels of income inequality are associated with lower levels of educational achievement, particularly among low-income students. Reardon argues that income inequality creates unequal access to educational resources, such as high-quality schools and extracurricular activities. He also discusses the potential policy implications of these findings, including the need for policies

that increase access to high-quality education for low-income students and reduce the concentration of poverty in certain schools and neighbourhoods.

Finally, Gilens (2005) explores the relationship between income inequality and political participation in the United States. He finds that higher levels of income inequality are associated with lower levels of political participation among low-income individuals. This is because low-income individuals are less likely to believe that their political participation will have an impact on policy outcomes and are, therefore, less likely to engage in political activities such as voting or contacting their elected officials. The authors argue that this can have negative consequences for democracy, as it can lead to policies that disproportionately benefit the wealthy at the expense of the poor. He also discusses potential policy solutions, such as campaign finance reform and increased access to voting for low-income individuals.

At last, we cannot forget to mention one of the loudest voices in this economic research field – Thomas Piketty. In his famous *magnum opus* 'Capital in the Twenty-First Century' published in August 2013, Piketty examines the relationship between wealth, income inequality, and capital accumulation over the last two centuries. He argues that income inequality has increased dramatically in many countries since the 1970s and that this trend is likely to continue.

The book is divided into four parts. In the first part, Picketty concludes that the concentration of wealth in the hands of a small elite has been a consistent feature of capitalist societies and argues that this concentration is likely to continue unless governments take action to redistribute wealth.

In the second part of the book, Piketty argues that the rate of return on capital (i.e., the income earned from owning financial assets such as stocks and bonds) is higher than the rate of economic growth, which means that capital owners will tend to accumulate wealth faster than the rest of society. He also argues that the concentration of wealth at the top of the income distribution is reinforced by factors such as inheritance and the tendency of the wealthy to marry each other.

In the third part of the book, Piketty discusses the political and ethical implications of his analysis. He argues that rising income inequality poses a threat to democratic societies, as it can lead to political polarization and decreased social mobility. He also points out that

the concentration of wealth in the hands of a small elite is unjust, and that governments have a moral obligation to address this inequality.

In the final part of the book, Piketty proposes several policy solutions to address rising income inequality. These include progressive taxation of capital income, a global tax on wealth, greater public investment in education and infrastructure, and greater international coordination to address global economic issues such as tax evasion and climate change.

However, Sawyer (2015), for instance, has mentioned several limitations of the thesis defended by Piketty. First, the differences between the rate of capital/wealth return and the rate of growth would lead to a deflationary process and high levels of unemployment – not just higher wealth inequality. When the return on capital exceeds the rate of economic growth, there may not be enough demand for goods and services to keep the economy growing at a healthy pace. High levels of wealth inequality can lead to lower levels of consumer spending, as wealthy individuals tend to save a larger share of their income than lower-income individuals.

Secondly, Piketty focuses on high-income and wealth taxation as a solution to address rising inequality, however, Sawyer (2015) mentions the need to design and deploy other alternative measures (coordinated efforts to minimize tax competition across countries and enhanced workforce negotiation power).

2.1.2 How to Measure?

Inequality remains a pressing issue in contemporary society, with its causes and effects still difficult to study. Researchers face both conceptual and practical challenges in measuring inequality, including the choice of variables, population targets, and different data sources. While traditional data sources like censuses, surveys, and tax records have been used to measure inequality, they have several shortcomings, including missing data and errors. This has led to the exploration of innovative data sources, such as spatial lights, machine learning from satellite images, and mobile phone metadata, to mitigate these limitations. Researchers have also developed different methods for measuring income inequality, including the Gini coefficient, the Atkinson index, and the Foster-Greer-Thorbecke (FGT) index. However, there is no single "correct" measure of income inequality, and policymakers need to consider multiple measures of poverty and inequality to develop comprehensive and effective policies.

This section will provide an overview of different methods and data sources used to measure inequality, highlighting their strengths, weaknesses, and implications for policy analysis.

According to McGregor et al. (2019, p. 388), the causes and effects of inequality remain difficult to study, as there are both conceptual (what's inequality?) and practical challenges (how?) in measuring it. The paper first reviews the issues usually faced by researchers (variable to use, population target, among others).

According to the authors, "different measures may emphasize inequality in different parts of the distribution, and thus yield widely different conclusions." The angle under study should guide the choice of the most suitable measures.

In a second approach, different data sources that can be used to measure inequality are shown: starting from the traditional ones and presenting innovative data sources that mitigate the previous limitations faced by using the traditional ones.

Traditional data sources are used to measure inequality, including censuses, other large-scale surveys, and tax records. These traditional sources have several shortcomings: poorer households are often missed from surveys; and high-income households prefer not to respond and conceal confidential data. Besides, the sources are usually inconsistent over time, not so frequently done, may contain errors and country-wide exhaustive surveys are rather costly.

Examples of alternative ways of sourcing data are (i) Spatial lights and population data, (ii) Machine learning from satellite images, (iii) Mobile phone metadata, and (iv) Potential 'big data' measures.

An important source of data is The Standardized World Income Inequality Database (SWIID). It aims to estimate income inequality for as many countries and years as possible while ensuring comparability, accuracy, and completeness (Solt (2020)).

SWIID estimates are based on the Luxembourg Income Study (LIS). Solt (2020) uses this data. Firstly, the paper estimates the relationship between Gini indices based on the LIS and all the other Ginis available for the same country-years, and, secondly, uses these relationships to estimate what the LIS Gini would be in country-years not included in the LIS but available from other data sources.

Sen et al. (1973) provide a critical overview of different methods for measuring income inequality. They argue no single "correct" measure of income inequality exists, as various measures may capture different aspects of inequality depending on the context. They also discuss the importance of considering the distribution of income across different groups, such as gender or race, in addition to overall inequality measures. They conclude by emphasizing the need for policymakers to consider multiple measures of income inequality to develop more comprehensive and effective policies to address inequality.

Bourguignon and Morrisson (2002) provide an overview of different methods for measuring income inequality, including the Gini coefficient, the Atkinson index, and the generalized entropy index. They compare the strengths and weaknesses of each method and discuss the implications of different measures of inequality for policy analysis. The authors also emphasize the importance of understanding the underlying factors driving income inequality, such as changes in the labor market or tax policy, to develop effective policy interventions.

Foster et al. (1984) provide a practical guide to measuring poverty and inequality, with a focus on the use of the Foster-Greer-Thorbecke (FGT) index. They provide step-by-step instructions for calculating the FGT index and discuss how it can be used to measure both absolute and relative poverty. The authors also highlight the importance of using multiple measures of poverty and inequality to develop more comprehensive policy interventions. They conclude by emphasizing the need for policymakers to consider both poverty and inequality measures to address the complex social and economic challenges facing low-income individuals and communities.

Dutt and Tsetlin (2021) argue that poverty is the key income distribution measure that matters for development outcomes. According to the authors, poverty corresponds to the bottom of the income distribution and the headcount measure of poverty. The two commonly used poverty measures are headcounts based on the World Bank poverty lines of $1.25 a day and $2 a day. Compared to Gini, poverty is more strongly causally associated with schooling and per capita income but not institutional quality. Their results question the literature's overwhelming focus on the Gini coefficient. At the least, their results imply that the causal link from inequality (as measured by Gini) to development outcomes is tenuous.

2.2 Machine Learning Approach

Machine learning techniques have been used in several studies to explore the relationship between inequality and economic development, estimate intergenerational income mobility, and analyse the evolution of inequality over time.

Achten and Lessmann (2020) found a negative relationship between spatial inequality and economic activity using parametric regression analysis and a random forest classification algorithm. This study showed a significant negative relationship between spatial inequality and economic activity.

Dutt and Tsetlin (2021) found that poverty is a crucial income distribution statistic that matters for development outcomes, and income distribution plays a significant role in economic development. The authors use machine learning techniques to explore the relationship between income distribution and economic development. The authors use a dataset of 142 countries over 50 years to analyse the impact of income inequality on economic growth.

The authors use two main statistical techniques: principal component analysis (PCA) and partial least squares (PLS) regression.

PCA is used to identify the most important factors that contribute to income inequality and economic development. The authors start by selecting a set of 34 economic and political indicators considered relevant to income inequality and economic development, such as GDP per capita, education levels, political stability, and income inequality measures. They then use PCA to reduce the dimensionality of the data by identifying patterns and correlations between the variables. This technique allows them to identify the most important factors that drive income inequality and economic development.

PLS regression is used to model the relationship between income inequality and economic development, considering the complex interrelationships between multiple variables. The authors use PLS regression to build a predictive model of economic development based on the identified factors. They use this model to test the impact of income inequality on economic development while controlling for other factors that may influence economic growth.

The study finds that income distribution plays a significant role in economic development, with a more equal distribution of income leading to higher levels of economic growth. The authors also identify several factors that contribute to income inequality, such as political instability and weak institutions.

The study highlights the potential of machine learning techniques in analyzing complex economic phenomena and generating insights that can inform policy decisions. The authors argue that policymakers should focus on promoting a more equal distribution of income to promote long-term economic growth and development.

On the other hand, to measure inequality, McGregor et al. (2019) also apply Machine Learning methods ('random forest' algorithm) to optimize the discretization of the continuous variable under study – inheritances. The authors conclude that there is no single "correct" measure of inequality, as each measure has its strengths and weaknesses depending on the specific context and purpose of the analysis.

The authors discuss several commonly used measures of Inequality, Iding the Gini coefficient, the Theil index, and the Atkinson index, among others. They note that the Gini coefficient is the most widely used measure of inequality and is a good overall measure, but it does not provide information about the distribution of income among different percentiles. The Theil index, on the other hand, is better suited for analyzing inequality at different levels of income distribution.

The authors also discuss the limitations of using income as the sole measure of well-being and inequality, as it does not capture other important factors such as wealth, health, education, and social mobility. The authors argue that a multidimensional approach to measuring inequality that considers these other factors is necessary for a more comprehensive understanding of the distribution of well-being.

Following this growing use of machine learning techniques to increase data accuracy, Bloise et al. (2021) proposed a machine learning-based approach to estimate intergenerational income mobility that is robust to data quality issues. The authors propose a machine learning-based approach to estimate intergenerational income mobility that is robust to data quality issues, such as measurement errors, missing data, and outliers.

The authors use a dataset from the Italian province of Cosenza, which includes information on the income of fathers and sons born between 1950 and 1980. The dataset is

characterized by several data quality issues, including measurement errors and missing data. To address these issues, the authors use a machine learning technique called random forests, which is a type of ensemble learning algorithm that combines multiple decision trees to improve the accuracy of predictions.

The authors find that the random forest algorithm outperforms traditional econometric models in estimating intergenerational income mobility, particularly in the presence of sub-optimal data. They also find that the results are robust to changes in the sample size and the inclusion of additional covariates.

Overall, the study highlights the potential of machine learning techniques to improve the estimation of intergenerational income mobility, particularly in the presence of sub-optimal data. The authors argue that this approach can provide policymakers with more accurate and reliable information on the dynamics of social mobility, which can inform policies aimed at promoting greater equality of opportunity.

Zooming into a study case, Brunori and Neidhöfer (2021) examine the evolution of inequality of opportunity in Germany over time using a machine learning approach. The authors use data from the German Socio-Economic Panel (SOEP) covering 1984 to 2016.

The study employs a machine learning technique called random forests to estimate the degree of inequality of opportunity over time. The authors use several variables that are thought to be related to inequality of opportunity, including parental education, income, occupation, and geographic location, and use these variables to predict individual outcomes, such as income, education, and occupation.

The results of the analysis show that there has been a decline in inequality of opportunity in Germany over the period of study, although there are still significant levels of inequality in certain areas, such as education and income. The authors also find that the factors that contribute to inequality of opportunity have changed over time, with geographic location becoming increasingly important in recent years.

These studies highlight the potential of machine learning techniques in generating insights that can inform policy decisions and improve the accuracy and reliability of estimates of inequality and mobility.

3. Dataset and Exploratory Data Analysis (EDA)

3.1 Dataset

After reviewing the existing literature, the next step was to identify which variables our dataset should be composed of. The variables were selected by considering the relevant literature.

For each variable, i) we listed the supporting references ii) identified the data source, and iii) provided the corresponding metadata (calculation formula, definition, assumptions, etc.).

The dataset has a total of 11 variables (1 target and 10 features). For each one (except Compulsory Education and Life Expectancy), we've selected 4 years corresponding to the lags 0, 1, 3, and 5 to test which one(s) is/are the most feasible ones. However, the most used lags according to the existing literature are 1 or 5 years. The final dataset output is 126 rows (countries) and 44 features. The main data source is the World Bank Data (2023).

In this chapter, we'll go through the reasons why we chose the variables described in Table 1:

1. **Real GDP per capita (PPP)**: Kuznets (1955) argued that economic growth, as measured by GDP, tends to be associated with an initial increase in income inequality, followed by a subsequent decrease. He observed an inverted U-shaped relationship between GDP per capita and income inequality, now commonly referred to a" the "Kuznets "urve." According to his hypothesis, in the early stages of economic development, income inequality tends to rise as some individuals and sectors benefit more from growth than others. However, as a country reaches higher levels of development, income inequality is expected to decline as the benefits of growth are more widely shared.
2. **Openness measure**: Trade openness, which refers to the degree to which a country engages in international trade and removes trade barriers, is often associated with globalization and economic integration. Increased trade openness can lead to the integration of domestic economies into global markets, exposing domestic industries to international competition. This integration can have implications for income distribution within a country. FDI can bring advanced

technology and expertise to a country, leading to a demand for skilled workers who can operate and manage these technologies. This can create wage differentials between skilled and unskilled workers, potentially exacerbating income inequality if the education and skills gap widens. (Dutt and Tsetlin (2021))

3. **Population growth rate**: Population growth rate influences the age structure of a population. Countries with high population growth often have a large proportion of young individuals who are entering the labor market. The interaction between population growth, age structure, and income inequality can create specific challenges and opportunities, as the economic well-being of young individuals can significantly impact overall income inequality levels. (Bourguignon and Morrisson (2002))
4. **Public Education Spending Share in GDP**: Piketty (2013) emphasizes the role of education in determining an indi'idual's earning potential and social mobility. Public education spending reflects a s'ciety's investment in human capital development. Higher levels of public education spending imply increased access to quality education, which can contribute to reducing income inequality by providing equal opportunities for all individuals to acquire skills and improve their economic prospects.
5. **Life expectancy**: Truesdale and Jencks (2016) considered life expectancy as a variable to study income inequality because it provides a comprehensive measure of health disparities that can be linked to income differences. By analyzing life expectancy patterns among different income groups, researchers can assess how much income inequality contributes to variations in health outcomes. This information can help policymakers and researchers understand the social and economic implications of income inequality and develop interventions to address health disparities and promote more equitable outcomes.
6. **Public Debt-to-GDP Ratio**: High levels of public debt can negatively impact a c'untry's macroeconomic stability. When the debt-to-GDP ratio is high, it indicates that a significant portion of the gove'nment's revenue is being used to service the debt, leaving fewer resources for social spending, public investments, and poverty reduction programs. This can exacerbate income inequality by limiting the gove'nment's ability to address social and economic disparities.

(Hailemariam et al. (2021))

7. **Foreign Investment**: Foreign investment can contribute to economic growth by providing capital, technology, and expertise to recipient countries. This can lead to increased productivity, job creation, and higher incomes, which may help reduce income inequality. (Huang et al. (2009))
8. **Compulsory Education**: Education is strongly associated with income levels. Individuals with more years of schooling often have access to higher-paying jobs and career opportunities, leading to higher incomes. In contrast, individuals with limited education may face limited job prospects and lower wages, contributing to income inequality. ()
9. **Political trust in government institutions**: Political trust has implications for government policies, including those related to income distribution. When citizens have higher levels of trust in the government, they are more likely to support and accept redistributive policies aimed at reducing income inequality. On the other hand, low levels of trust may lead to skepticism and resistance to such policies. (Neves e Silva (2014))
10. **Financial sector level of development**: A well-developed financial sector plays a crucial role in channeling funds from savers to borrowers. It facilitates access to credit and financial services, allowing individuals and businesses to invest, expand, and create income-generating opportunities. When the financial sector is more developed, it tends to enhance economic growth and provide greater opportunities for wealth accumulation. (Neves e Silva (2014))

Variable	Description	Citation
Gini Index	The Gini index is a measure of the distribution of income across a population.	*Several*
Real GDP per capita (PPP)	A c'untry's gross domestic product (GDP) at purchasing power parity (PPP) per capita is the PPP value of all final goods and services produced within an economy each year, divided by the average (or mid-year) population for the same year.	Hailemariam et al. (2021), Kuznets (1955), Achten and Lessmann (2020), Bourguignon and Morrisson (2002)
Openness measure	The ratio of exports plus imports to GDP	Dutt and Tsetlin (2021), Jaumotte et al. (2013), Huang et al. (2009)
Population growth rate	The annual percentage change in the population size of a country	Dutt and Tsetlin (2021), Parcero (2021), Galor an' Moav's (2004), Bourguignon and Morrisson (2002)
Public Education Spending Share in GDP	Average share of public expenditures on education as a fraction of GDP	Reardon (2013), Thomas Piketty (2013), Dutt and Tsetlin (2021), Huang et al. (2009)
Life expectancy	Life expectancy is a statistical measure of the average time an organism is expected to live, based on the year of its birth, current age, and other demographic factors like sex.	Chetty et al. (2016), Truesdale and Jencks (2016)
Public Debt-to-GDP Ratio	The debt-to-GDP ratio is a metric that helps understand a c'untry's ability to pay back its debts.	Bartak et al. (2022), Miyashita (2023), Hailemariam et al. (2021), Huang et al. (2009)
Foreign Investment	FDI inflows	Jaumotte et al. (2013), Ravinthirakumaran and Navaratnam (2018), Yuldashev et al. (2023), Bourguignon and Morrisson (2002), Huang et al. (2009)
Years of schooling	Average years of schooling	Dutt and Tsetlin (2021), Barro and Lee (2013), Galor & Moav (2004), Galor and Zeira (1993), Huang et al. (2009)
Political trust in government institutions	CPIA transparency, accountability, and corruption in the public sector rating (1=low to 6=high)	Neves e Silva (2014), Bourguignon and Morrisson (2002), Galor and Zeira (1993), Gilens (2005)
Financial sector level of development	Domestic credit provided by financial sector (% of GDP)	Neves e Silva (2014), Jaumotte, F., et al. (2013), Koudalo, Y. M. A. and J. Wu (2022)

Table 1 - Variables description and references

After selecting and validating the variables (features), we preprocessed the data, mainly dealing with missing values:

1. Exclude observations (countries) with the percentage of **missing values higher than 30%** (down to 165 countries)
    a. For each feature/country, after checking all the lags (even though we have selected only 0, 1, 3 and 5):
        i. If a given country had 4 or more missing values, we've excluded them from the sample (down to 143 countries)
    b. If a given country had **up to 3 missing values**, we've imputed the average of the remaining observations
2. For the target Gini_Index missing values in 2019:
    a. Firstly, we've imputed the average of the years in which we had values (period between 2014 and 2018)
    b. If not possible, we've resorted to an alternative dataset from CIA World Factbook (2023) which was in line with the World Bank methodology and included additional estimations for several countries
    c. Lastly, if neither of the two previous approaches was possible, we've excluded them (down to 126 countries)

Variable	Statistical concept and methodology
Gini Index	The Gini index measures the area between the Lorenz curve and a hypothetical line of absolute equality, expressed as a percentage of the maximum area under the line. A Lorenz curve plots the cumulative percentages of total income received against the cumulative number of recipients, starting with the poorest individual. Thus, a Gini index of 0 represents perfect equality, while an index of 100 implies perfect inequality. The Gini index provides a convenient summary measure of the degree of inequality. Data on the distribution of income or consumption come from nationally representative household surveys. Where the original data from the household survey were available, they were used to calculate the income or consumption shares by quintile. Otherwise, shares have been estimated from the best available grouped data. The distribution data have been adjusted for household size, providing a more consistent measure of per capita income or consumption. The year reflects the year in which the underlying household survey data were collected or, when the data collection period bridged two calendar years, the year data collection started.

Real GDP per capita (PPP)	This indicator provides per capita values for gross domestic product (GDP) expressed in current international dollars converted by purchasing power parity (PPP) conversion factor. GDP is the sum of gross value added by all resident producers in the country plus any product taxes and minus any subsidies not included in the value of the products. Conversion factor is a spatial price deflator and currency converter that controls for price level differences between countries. Total population is a mid-year population based on the de facto definition of population, which counts all residents regardless of legal status or citizenship.
Openness measure	Gross domestic product (GDP) from the expenditure side is made up of household final consumption expenditure, general government final consumption expenditure, gross capital formation (private and public investment in fixed assets, changes in inventories, and net acquisitions of valuables), and net exports (exports minus imports) of goods and services. Such expenditures are recorded in purchaser prices and include net taxes on products.
Population growth rate	Annual population growth rate for year t is the exponential rate of growth of the midyear population from year t-1 to t, expressed as a percentage. Population is based on the de facto definition of population, which counts all residents regardless of legal status or citizenship.
Public Education Spending Share in GDP	Current expenditure, total is calculated by dividing all current expenditure in public institutions of all levels of education by total expenditure (current and capital) in public institutions of all levels of education and multiplying by 100. Aggregate data are based on World Bank estimates. Data on education are collected by the UNESCO Institute for Statistics from official responses to its annual education survey. All the data are mapped to the International Standard Classification of Education (ISCED) to ensure the comparability of education programs at the international level. The current version was formally adopted by UNESCO Member States in 2011. The reference years reflect the school year for which the data are presented. In some countries, the school year spans two calendar years (for example, from September 2010 to June 2011); in these cases, the reference year refers to the year in which the school year ended (2011 in the example).
Life expectancy	Life expectancy at birth indicates the number of years a newborn infant would live if prevailing patterns of mortality at the time of its birth were to stay the same throughout its life.
Public Debt-to-GDP Ratio	Debt is the entire stock of direct government fixed-term contractual obligations to others outstanding on a particular date. It includes domestic and foreign liabilities such as currency and money deposits, securities other than shares, and loans. It is the gross amount of government liabilities reduced by the amount of equity and financial derivatives held by the government. Because debt is a stock rather than a flow, it is measured as of a given date, usually the last day of the fiscal year.
Foreign Investment	Foreign direct investment is the net inflows of investment to acquire a lasting management interest (10 percent or more of voting stock) in an enterprise operating in an economy other than that of the investor. It is the sum of equity capital, reinvestment of earnings, other long-term capital, and short-term capital as shown in the balance of payments. This series shows net inflows (new investment inflows less disinvestment) in the reporting economy from foreign investors and is divided by GDP.
Years of schooling	Average years of schooling of the adult population (ages 25 and older) in a given country: 5-year intervals

Political trust in government institutions	Transparency, accountability, and corruption in the public sector assess the extent to which the executive can be held accountable for its use of funds and for the results of its actions by the electorate and by the legislature and judiciary, and the extent to which public employees within the executive are required to account for administrative decisions, use of resources, and results obtained. The three main dimensions assessed here are the accountability of the executive to oversight institutions and of public employees for their performance, access of civil society to information on public affairs, and state capture by narrow vested interests.
Financial sector level of development	Domestic credit provided by the financial sector includes all credit to various sectors on a gross basis, except for credit to the central government, which is net. The financial sector includes monetary authorities and deposit money banks, as well as other financial corporations where data are available (including corporations that do not accept transferable deposits but do incur such liabilities as time and savings deposits). Examples of other financial corporations are finance and leasing companies, money lenders, insurance corporations, pension funds, and foreign exchange companies.

Table 2 - Variables metadata (source: https://data.worldbank.org/)

3.2 Exploratory Data Analysis (EDA)

Before feeding the models with the pre-processed dataset, we need to get a grasp on the data and derive valuable insights if possible. We will begin with our target, the Gini Index.

Before moving on, we'll clarify two concepts: features are the independent variables we use for predicting the target (dependent variable).

From a total of 126 observed countries (Table 3), the average Gini Index is 37 – the most unequal country from our sample in 2019 was South Africa (63) and the most egalitarian was Slovak Republic (23).

From the histogram and box plot (Figure 1), most of the countries are concentrated on the low/middle range of Gini Index.

	Gini_Index
count	126
mean	37
std	8
min	23
25%	31
50%	36
75%	42
max	63

Table 3 - Gini Index statistics

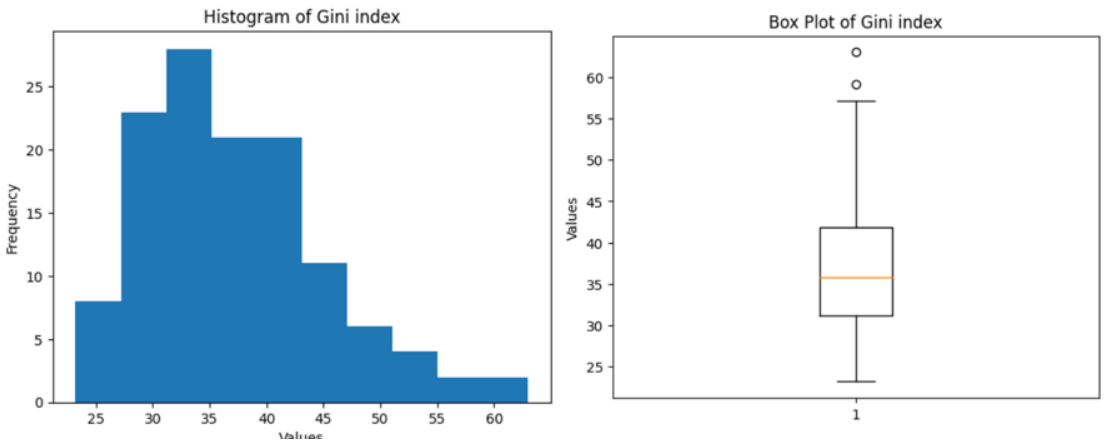

Figure 1 - Gini Index Histogram and Box plot

Analysing the Pearson correlation between Gini Index and all the independent variables, we have that:

1. Gini Index is considerably negatively correlated to GDP per capita – 2019 (- 0.34), Life expectancy at birth, total (years) (- 0.43) and Control of Corru–tion - 2019 (- 0.33)
2. Gini Index is considerably positively correlated to Population growth (annual %) – 2019 (0.36)
3. From the Table 4, Compulsory Education, duration (years) and Government Expenditure (% GDP) have no correlation with the level of inequality in each country.

	Gini_Index
Gini_Index	1
Monetary Sector credit to private sector (%–GDP) - 2014	- 0.26
Monetary Sector credit to private sector (%–GDP) - 2016	- 0.24
Monetary Sector credit to private sector (%–GDP) - 2018	- 0.23
Monetary Sector credit to private sector (%–GDP) - 2019	- 0.22
General government gross–debt - 2014	- 0.00
General government gross–debt - 2016	0.03
General government gross–debt - 2018	0.03
General government gross–debt - 2019	0.03
Population growth (annu–l %) - 2014	0.17
Population growth (annu–l %) - 2016	0.32
Population growth (annu–l %) - 2018	0.35
Population growth (annu–l %) - 2019	0.36
GDP per capita– PPP - 2014	- 0.27
GDP per capita– PPP - 2016	- 0.32
GDP per capita– PPP - 2018	- 0.33
GDP per capita– PPP - 2019	- 0.34
Exports of goods and services (% of–GDP) - 2014	- 0.22
Exports of goods and services (% of–GDP) - 2016	- 0.27
Exports of goods and services (% of–GDP) - 2018	- 0.25
Exports of goods and services (% of–GDP) - 2019	- 0.25
Government expenditure on education, total (% of–GDP) - 2014	0.02
Government expenditure on education, total (% of–GDP) - 2016	- 0.01
Government expenditure on education, total (% of–GDP) - 2018	0.00
Government expenditure on education, total (% of–GDP) - 2019	- 0.02
Life expectancy at birth, total (years)	- 0.43
Foreign direct investment, net inflows (% of–GDP) - 2014	- 0.07
Foreign direct investment, net inflows (% of–GDP) - 2016	- 0.15
Foreign direct investment, net inflows (% of–GDP) - 2018	0.08

Foreign direct investment, net inflows (% of–GDP) - 2019	-	0.10
Control of Corruption: Est–mate - 2014	-	0.30
Control of Corruption: Est–mate - 2016	-	0.31
Control of Corruption: Est–mate - 2018	-	0.32
Control of Corruption: Est–mate - 2019	-	0.33
Compulsory education, duration (years)	-	0.01

Table 4 - Gini Index Pearson Correlation

Table 5 contains the summary of all the variables' statistics. The main preliminary insights we can derive from it are:

- Credit to the private sector as a percentage of GDP remained relatively stable across the years
- Government gross debt increased significantly from 2014 to 2019, indicating potential fiscal issues
- Life expectancy at birth indicates the general health and well-being of a population. The values are relatively high, suggesting improvements in healthcare and living conditions
- The "Control of Corruption" estimates suggest the perceived level of corruption within each country. Most countries have a rating of 0 or 1, indicating issues in this area

Metric	mean	std	min	25%	50%	75%	max
Gini_Index	37	8	23	31	36	42	63
Credit to private sector (% GDP) - 2014	55	42	0	22	44	76	252
Credit to private sector (% GDP) - 2016	56	41	0	24	47	79	216
Credit to private sector (% GDP) - 2018	54	39	0	24	48	75	167
Credit to private sector (% GDP) - 2019	54	39	0	24	48	75	167
Government gross debt - 2014	71,949	348,268	0	86	804	4,036	2,608,776
Government gross debt - 2016	129,005	732,641	0	126	874	5,579	6,902,421
Government gross debt - 2018	158,559	921,015	0	155	1,083	7,821	8,770,854
Government gross debt - 2019	187,353	1,160,063	0	174	1,092	9,187	11,685,005
Population growth (annual %) - 2014	1	2	-1	0	1	2	12
Population growth (annual %) - 2016	1	1	-2	0	1	2	7
Population growth (annual %) - 2018	1	1	-3	0	1	2	4

Population growth (annual %) - 2019	1	1	-3	0	1	2	4
GDP per capita, PPP - 2014	20,876	22,023	670	5,235	13,473	28,883	143,333
GDP per capita, PPP - 2016	21,616	20,933	787	5,658	13,861	31,216	113,365
GDP per capita, PPP - 2018	23,737	22,854	875	5,900	14,782	35,919	116,966
GDP per capita, PPP - 2019	24,760	23,661	898	6,041	15,197	38,546	120,175
Exports of goods and services (% of GDP) - 2014	43	31	5	24	34	48	192
Exports of goods and services (% of GDP) - 2016	41	30	3	23	32	48	191
Exports of goods and services (% of GDP) - 2018	44	31	3	25	36	51	198
Exports of goods and services (% of GDP) - 2019	44	32	9	24	36	51	204
Government expenditure on education, total (% of GDP) - 2014	5	2	2	3	4	6	10
Government expenditure on education, total (% of GDP) - 2016	4	2	1	3	4	5	11
Government expenditure on education, total (% of GDP) - 2018	4	2	2	3	4	5	10
Government expenditure on education, total (% of GDP) - 2019	4	2	2	3	4	5	12
Life expectancy at birth, total (years)	73	8	52	68	75	79	84
Foreign direct investment, net inflows (% of GDP) - 2014	6	20	-5	1	3	4	223
Foreign direct investment, net inflows (% of GDP) - 2016	5	9	-37	1	3	5	54
Foreign direct investment, net inflows (% of GDP) - 2018	2	8	-40	1	2	4	29
Foreign direct investment, net inflows (% of GDP) - 2019	5	19	-12	1	3	4	204
Control of Corruption: Estimate - 2014	0	1	-2	-1	-0	1	2
Control of Corruption: Estimate - 2016	0	1	-2	-1	-0	1	2
Control of Corruption: Estimate - 2018	0	1	-2	-1	-0	1	2
Control of Corruption: Estimate - 2019	0	1	-1	-1	-0	1	2
Compulsory education, duration (years)	10	2	5	9	10	12	16

Table 5 - Features descriptive statistics

4. Models

4.1 Training and test sets

Using the most recent categorization available from the World Bank (based on GDP), the countries can be grouped by: High income, Low income, Lower middle income, and Upper middle income. Thus, if the test set corresponds to, for example, 30% of the original dataset, we will then have to ensure that each of these categories is well represented (as shown below).

Income Group	# of observations
High income	31
Low income	9
Lower middle income	26
Upper middle income	22

Income Group	# of observations
High income	13
Low income	4
Lower middle income	12
Upper middle income	9

Table 6 - On the right, is training set, on the left test set

4.2 Applying ML (Machine Learning)

We deployed ML models using a native Python library - scikit-learn. We ran the scripts on Google Colab. ML models used were:

1. Decision Tree
2. Random Forest
3. Gradient Boosting
4. XGB (we have used a scikit-learn API with fewer functionalities)

For each of the models:

1. Hyper parameters were tuned by Grid Search CV, to identify the best values for the parameters of each model
2. SHAP Values were calculated to identify which variables were most relevant for Gini Index's prediction
3. Several performance measures were calculated, among which the main are:
    a. R^2 score: In statistics, R-squared (often denoted as R^2) is a measure of how

well a regression model fits the observed data. R-squared represents the proportion of the variance in the dependent variable that is explained by the independent variables in the model. In other words, it quantifies the goodness of fit of the regression model.

b. MAPE (Mean Absolute Percentage Error): is a commonly used metric in statistics and forecasting to measure the accuracy of a model's predictions, particularly in the context of time series forecasting or regression analysis. MAPE quantifies the average percentage difference between the predicted values and the actual observed values.

c. MSE (Mean Squared Error): calculates the average of the squared differences between predicted and actual values. Squaring the errors gives more weight to larger errors, making it sensitive to outliers.

d. RMSE (Root Mean Squared Error): is derived from MSE and is useful because it provides an error measurement in the same units as the dependent variable. It is calculated by taking the square root of the MSE. RMSE is more interpretable as it is in the same units as the data, making it easier to understand and compare across different models.

e. MAE (Mean Absolute Error): calculates the average of the absolute differences between predicted and actual values. MAE is less sensitive to outliers compared to MSE, making it a good choice when dealing with data that may contain extreme values.

In the context of machine learning, overfitting and underfitting are two common challenges that affect the performance of predictive models.

Overfitting occurs when a machine learning model learns the training data too well, capturing not only the underlying patterns but also the noise and random fluctuations present in the data. This results in a model that performs excellently on the training data but fails to generalize to new, unseen data. In other words, it memorizes the training data rather than learning the true underlying relationships.

Underfitting, on the other hand, occurs when a model is too simplistic to capture the underlying patterns in the data. It fails to learn important relationships and exhibits poor performance not only on the training data but also on unseen data. Underfitting is characterized by high bias, as the model is too generalized to capture the complexities of the data.

4.2.1 Decision Tree

A decision tree is a hierarchical structure used for classification and regression tasks. It makes decisions by recursively partitioning the input data into subsets based on the values of distinctive features. At each step, the algorithm selects the feature that best separates the data, creating branches that lead to different outcomes. The process continues until a stopping criterion is met, such as a maximum depth or a minimum number of samples in a node (Breiman (1984)).

Main model parameters and selected values from GridSearch CV (detailed in Tables 7 and 8):

1. splitter: specifies the strategy used to choose the best split when building the tree.
2. max_depth: specifies the maximum depth of the tree. it controls the complexity and potential overfitting of the tree.
3. min_samples_leaf: minimum number of samples required to be in a leaf node. It ensures a minimum amount of data in each leaf.
4. max_features: is a hyperparameter that determines the maximum number of features (variables or attributes) considered when making a split decision in a decision tree or random forest.
5. min_weight_fraction_leaf: is a hyperparameter that sets the minimum weighted fraction of the total number of samples required to be in a leaf node. It helps prevent the creation of very small leaf nodes that may fit the training data noise.
6. max_leaf_nodes: is a hyperparameter that limits the maximum number of leaf nodes in a decision tree.

splitter	**["best", "random"]**
max_depth	[3, 5, 7, 9, 11, 15]
min_samples_leaf	[1, 2, 3, 4, 5, 6, 7]
max_features	["auto", "log2", "sqrt", None]
min_weight_fraction_leaf	[0.0, 0.1, 0.2,0.3]
max_leaf_nodes	[None,5,10,15,20,25,30]

Table 7 - DT GridSearch CV test values

splitter	**random**
max_depth	3
min_samples_leaf	3
max_features	auto
min_weight_fraction_leaf	0
max_leaf_nodes	None

Table 8 - DT best parameter values

4.2.2 Random Forest

Random Forest is an ensemble learning technique that builds multiple decision trees and combines their predictions to improve accuracy and reduce overfitting. Each tree is trained on a random subset of the data (bagging) and a random subset of the features. The final prediction is obtained through a majority vote (classification) or averaging (regression) of the individual tree predictions (Breiman (2001)).

Main model parameters and selected values from GridSearch CV (detailed in Tables 9 and 10):

1. bootstrap: specifies whether bootstrap samples are used when building trees. If False, the whole dataset is used to build each tree.
2. max_depth: specifies the maximum depth of the tree. it controls the complexity and potential overfitting of the tree.
3. max_features: the number of features to consider when looking for the best split at each node. Can be a fixed number or a fraction of the total features.
4. min_samples_leaf: minimum number of samples required to be in a leaf node. It ensures a minimum amount of data in each leaf.
5. min_samples_split: is a hyperparameter that controls the minimum number of samples required to split a node in a decision tree within the Random Forest. It specifies the minimum number of data points that must be present in a node before it can be further split into child nodes.
6. n_estimators: number of decision trees in the forest.

bootstrap	**[True, False]**
max_depth	[10, 20, 30, None]
max_features	['auto', 'sqrt']

min_samples_leaf	[1, 2, 4]
min_samples_split	[2, 5, 10]
n_estimators	[50, 100]

Table 9 - RF GridSearch CV test values

bootstrap	**False**
max_depth	10
max_features	'sqrt'
min_samples_leaf	1
min_samples_split	2
n_estimators	50

Table 10 - RF best parameter values

4.2.3 Gradient Boosting

Gradient Boosting is another ensemble method that builds a sequence of decision trees, where each subsequent tree corrects the errors of the previous one. It fits new trees to the residuals of the previous ones, optimizing a loss function using gradient descent. This iterative process gradually reduces the prediction errors (Friedman and Fisher (1999)).

Main model parameters and selected values from GridSearch CV (detailed in Tables 11 and 12):

1. n_estimators: number of boosting stages (iterations).
2. learning_rate: shrinks the contribution of each tree, controlling the step size in updating the model.
3. max_depth: specifies the maximum depth of the tree. it controls the complexity and potential overfitting of the tree.
4. subsample: fraction of samples used for fitting each tree. Can prevent overfitting by introducing randomness.
5. max_features: the number of features to consider when looking for the best split at each node. Can be a fixed number or a fraction of the total features.
6. min_samples_leaf: minimum number of samples required to be in a leaf node. It ensures a minimum amount of data in each leaf.
7. min_samples_split: is a hyperparameter that controls the minimum number of

samples required to split a node in a decision tree within the Random Forest. It specifies the minimum number of data points that must be present in a node before it can be further split into child nodes.

n_estimators	**[50,100]**
learning_rate	[.001,0.01,.1]
max_depth	[1,2,4],
subsample	[.5,.75,1]
max_features	['auto', 'sqrt']
min_samples_leaf	[1, 2, 4]
min_samples_split	[2, 5, 10]

Table 11 - GB GridSearch CV test values

n_estimators	**100**
learning_rate	0.1
max_depth	4
subsample	1
max_features	'sqrt'
min_samples_leaf	2
min_samples_split	10

Table 12 - GB best parameter values

4.2.4 XGBoost (Extreme Gradient Boosting)

XGBoost is an optimized and highly efficient implementation of the gradient boosting algorithm. It enhances gradient boosting by employing techniques such as regularization, parallel processing, and efficient tree-building algorithms. XGBoost has gained significant popularity due to its performance and has won multiple Kaggle competitions (Chen and Guestrin (2016)).

Main model parameters and selected values from GridSearch CV (detailed in Tables 13 and 14):

1. learning_rate: controls the step size at each iteration.
2. colsample_bytree: control the fraction of samples and features used in tree construction.

3. subsample: fraction of samples used for fitting each tree. Can prevent overfitting by introducing randomness.
4. max_depth: maximum depth of a tree.
5. n_estimators: number of boosting rounds (trees).
6. reg_lambda (L2 regularization): regularization parameters.
7. gamma: minimum loss reduction required to make a further partition on a leaf node.

learning_rate	**[0.1, 0.01]**
colsample_bytree	[0.6, 0.8, 1.0]
subsample	[0.6, 0.8, 1.0]
max_depth	[2, 3, 4]
n_estimators	[50, 100]
reg_lambda	[1, 1.5, 2]
gamma	[0, 0.1, 0.3]

Table 13 - XGB GridSearch CV test values

learning_rate	**0.1**
colsample_bytree	0.6
subsample	0.8
max_depth	2
n_estimators	100
reg_lambda	1
gamma	0.3

Table 14 - XGB best parameter values

4.2.5 Main differences between the models

The main differences between decision trees, random forests, gradient boosting, and XGBoost lie in their underlying concepts, ensemble methods, and model optimization techniques.

Decision Tree:

1. Single Model: Decision trees are standalone models that make decisions based on

feature values.

2. Bias-Variance Tradeoff: Prone to overfitting if the tree is too deep, leading to high variance. Shallow trees may underfit.
3. Ensemble: Not inherently an ensemble method.

Random Forest:

1. Ensemble Method: Combines multiple decision trees to improve accuracy and mitigate overfitting.
2. Randomness: Each tree is trained on a random subset of data and features, introducing randomness, and reducing correlation between trees.
3. Bagging: Uses bootstrap aggregating to create diverse trees.
4. Voting/Averaging: Combines predictions through majority voting (classification) or averaging (regression).

Gradient Boosting:

1. Ensemble Method: Builds a sequence of trees, where each tree corrects the errors of the previous ones.
2. Iterative: Trees are added sequentially, with each focusing on residuals from the previous trees.
3. Adaptive: Adjusts weights for training samples to emphasize misclassified samples.
4. Learning Rate: Shrinks the contribution of each tree, allowing fine-tuning of the model.
5. Strong Learner: Tends to perform well even with weak base learners.

XGBoost (Extreme Gradient Boosting):

1. Optimized Gradient Boosting: An optimized implementation of gradient boosting.
2. Regularization: Incorporates L1 and L2 regularization to control complexity and avoid overfitting.
3. Efficiency: Utilizes parallel processing, sparsity-aware split finding, and other optimizations for faster training.
4. Performance: Often achieves competitive performance in various machine learning competitions.

Summing up, decision trees are simple standalone models, random forests combine decision trees with randomness to improve accuracy, gradient boosting builds a sequence of trees to minimize errors, and XGBoost is an optimized version of gradient boosting with added regularization and efficiency enhancements. The choice of model depends on the problem, dataset characteristics, and desired balance between accuracy, interpretability and complexity.

5. Results

In the following section we will describe the results summarized in Table 15.

Among the models, the Gradient Boosting (gb) model has the lowest MAE, followed closely by the XGBoost (xgboost) model. This indicates that these two models have better predictive accuracy in terms of absolute error compared to the other models (dt and rf).

The Gradient Boosting (gb) model has a much higher r^2 value compared to the other models. This indicates that the gb model explains a larger portion of the variance in the dependent variable, suggesting better overall fit and predictive power.

The Gradient Boosting (gb) and XGBoost (xgboost) models have considerably lower MSE and RMSE values compared to the Decision Tree (dt) and Random Forest (rf) models. This suggests that gb and xgboost have better precision in terms of squared errors, resulting in better predictive accuracy.

The Gradient Boosting (gb) and XGBoost (xgboost) models have lower MAPE values compared to the Decision Tree (dt) and Random Forest (rf) models. This indicates that gb and xgboost have a lower average percentage error in their predictions.

Overall, the conclusions can be summarized as follows:

1. The Gradient Boosting (gb) and XGBoost (xgboost) models generally outperform the Decision Tree (dt) and Random Forest (rf) models across most of the metrics.
2. The Gradient Boosting (gb) model stands out with high values for r2 and comparatively low values for other error metrics (MAE, MSE, RMSE, MAPE).
3. The Random Forest (rf) and XGBoost (xgboost) models also perform well, with lower values for error metrics compared to the Decision Tree (dt) model.
4. The Decision Tree (dt) model generally has the highest error metrics, indicating that it might be the least accurate among the models considered.

Metrics	dt	rf	gb	xgboost
MAE	5.73	5.066129	4.841727	5.065625
r2	0.09	0.173366	0.321709	0.287392
MSE	53.10	48.30737	39.63837	41.64386
RMSE	7.29	6.95035	6.295901	6.453205
MAPE	0.16	0.136642	0.128684	0.134999

Table 15 - Performance metrics

5.1 Shap Values

First, SHAP (SHapley Additive exPlanations) values are a technique used in machine learning and data science to explain the output of a model's predictions for a specific instance or observation. They provide a way to understand how the input features of a model contribute to its predictions (Lundberg (2017)).

Figures 2, 3, 4 and 5 correspond to the SHAP values output for each one of the four models, ranking from the most important variables to the least relevant in terms of prediction value. The top 4 most valuable features for each model are highlighted in red.

Population Growth and Life Expectancy are always in top 4 most 'valuable' variables in all 4 different models. Random Forest emphasizes exactly that where the 3 most relevant variables are different lags of Population Growth (2018, 2019 and then 2016, from the most to the least important).

Except the Random Forest, Compulsory education is quite relevant as well, usually occupying the 2nd or 3rd places on the rank. Control of Corruption (2018), Exports of Goods and Services (2018) and Credit to Private Sector (2018) are relatively relevant but not so pervasive as Population Growth or even Life Expectancy.

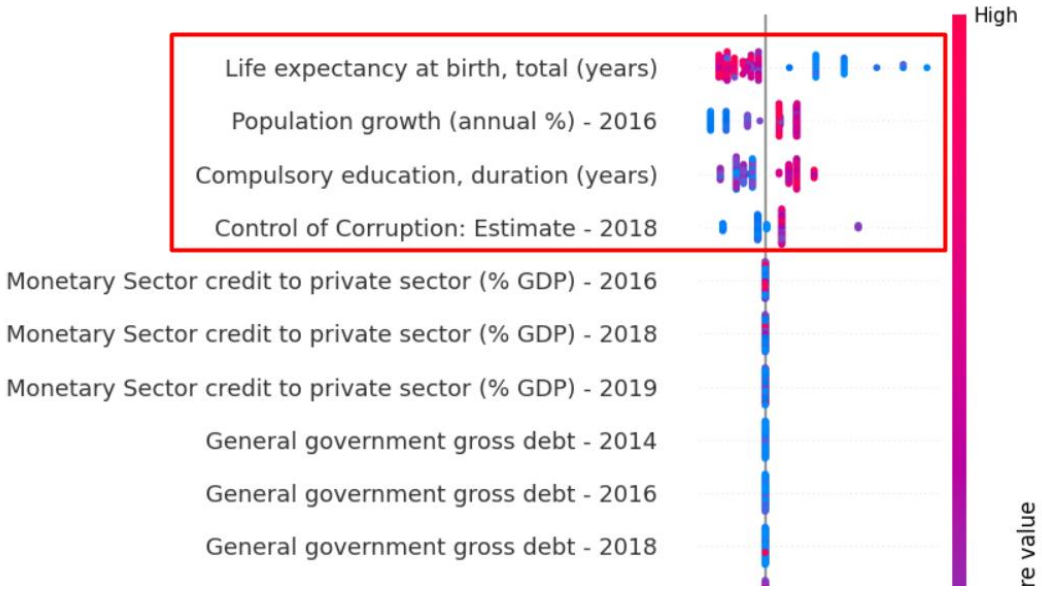

Figure 2 - Decision Tree

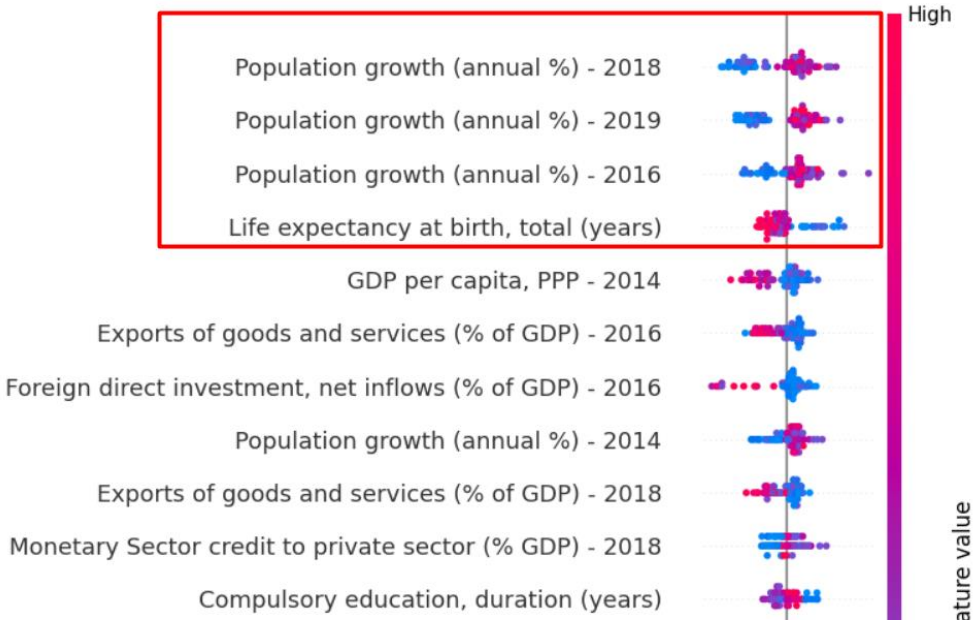

Figure 3 - Random Forest

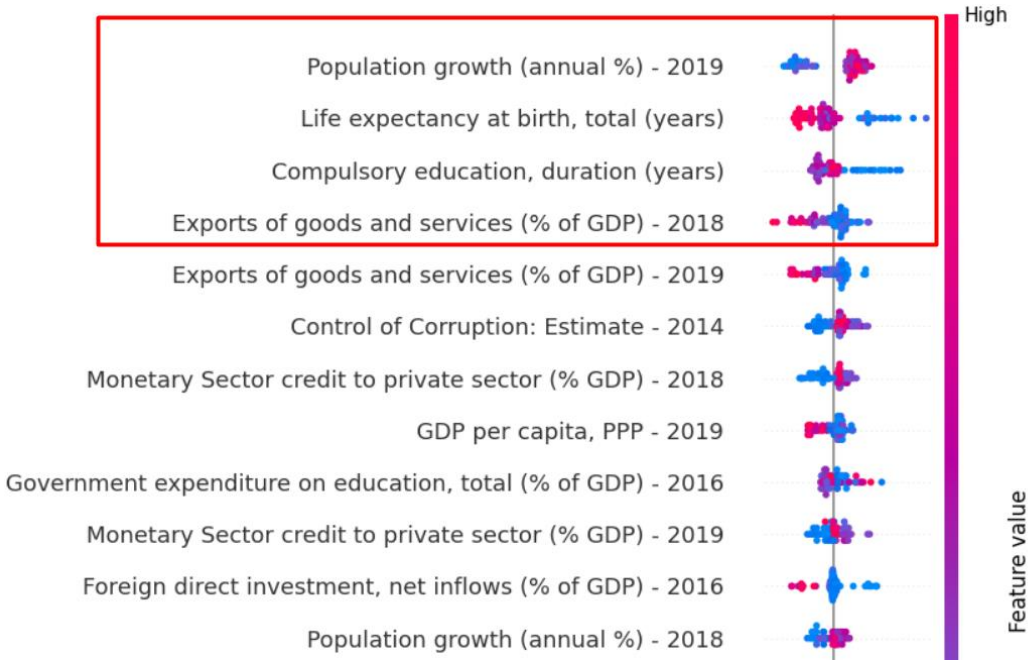

Figure 4 - Gradient Boosting

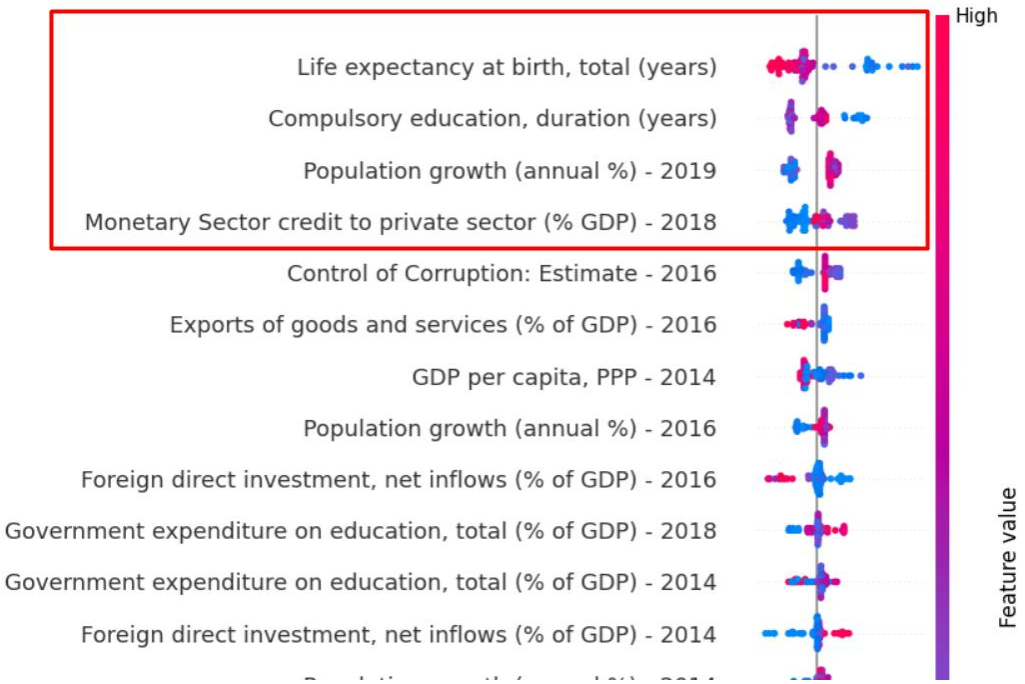

Figure 5 - XGBoost

5.2 Gradient Boosting (GB) V2

After analysing the models' performance, we decided to test the following hypotheses: if we train the best performing model (Gradient Boosting) only with the most valuable features/lags, will we have a better performance?

The updated and simplified dataset is as follows (Table 16):

Monetary Sector credit to private sector (% GDP) - 2018
Population growth (annual %) - 2019
GDP per capita, PPP - 2019
Exports of goods and services (% of GDP) - 2018
Government expenditure on education, total (% of GDP) - 2016
Foreign direct investment, net inflows (% of GDP) - 2016
Control of Corruption Estimate - 2014
Compulsory education, duration (years)

Table 16 - Selected features

1. Grid Search CV: The parameters and corresponding ranges were the same

applied to the original Gradient Boosting model

2. However, the optimum parameters were different (detailed below in Table 17)

learning_rate	0.1
max_depth	2
max_features	'sqrt'
min_samples_leaf	2
min_samples_split	10
n_estimators	100
subsample	1

Table 17 - GB V2 best parameter values

As we can see the R^2 decreased a lot (from 39% to 9%). We conclude that the model performs at its best when it's trained with all the variables (Table 18)

Metrics	dt	rf	gb	xgboost	gb_v2
MAE	5.73	5.066129	4.841727	5.065625	5.9353926
r2	0.09	0.173366	39.63837	0.287392	0.0944558
MSE	53.1	48.30737	0.32171	41.64386	52.918779
RMSE	7.29	6.95035	6.295901	6.453205	7.2745295
MAPE	0.16	0.136642	0.128684	0.134999	0.1628309

Table 18 - GB V2 performance

5.3 Best model

We concluded that Gradient Boosting is the best model with MAPE of 13% and r^2 score of 32%.

Analysing the Gradient Boosting (GB) 'Shap Values' output to answer the research question, we can conclude that the variable 'Population growth (annual %) - 2019' with a lag of 0 is the feature with the most predictive value. Lower values of the population growth rate 'pull' the Gini value down (lower inequality).

In second and third place in terms of importance are the variables 'Life expectancy at birth, total (years)' and 'Compulsory education, duration (years)'. In both cases, higher values indicate lower Gini values (lower inequality).

In 4th place, we have 'Exports of goods and services (% of GDP) - 2018' with a lag of 1 year. The more a country exports what it produces, the lower the inequality will be. A possible starting hypothesis: Countries that are high exporters relative to the total goods and services they produce tend to have a diversified and highly specialized industry with high value-added. For a population to add that much value in the competitive global market, it must be highly skilled and therefore well paid (Jaumotte et al. (2013)).

6. Conclusions

In this research, we studied income inequality across the world.

Our objective was to estimate/predict the Gini Index (regression) by applying Machine Learning models. The second objective was to understand which factors contribute the most to income inequality in each country, i.e., we assessed which independent variable or variables contribute most to the prediction of the income inequality index (Gini).

After reviewing the existing literature, the next step was to identify which variables our dataset should be composed of.

The dataset has a total of 11 variables (1 target and 10 features). For each one (except Compulsory Education and Life Expectancy), we've selected 4 years corresponding to the lags 0, 1, 3, and 5 to test which one(s) is/are the most feasible ones.

We deployed ML models using a native Python library - scikit-learn: Decision Tree, Random Forest, Gradient Boosting and XGB. In this context, we've identified the best parameters of each model (Grid Search CV) and which variables were most relevant for the prediction of the Gini Index (SHAP Values).

In terms of performance, the Gradient Boosting (gb) and XGBoost (xgboost) models generally outperform the Decision Tree (dt) and Random Forest (rf) models across most of the metrics. On the other hand, the Decision Tree (dt) model generally has the highest error metrics, indicating that it might be the least accurate among the models considered.

We concluded that Gradient Boosting is the best model with MAPE of 13% and r^2 score of 32%. Analysing its 'Shap Values' output to answer the research question, we can conclude that the variable 'Population growth (annual %) - 2019' with a lag of 0 is the feature with the most predictive value. In second and third place in terms of importance are the variables 'Life expectancy at birth, total (years)' and 'Compulsory education, duration (years)'.

One of the limitations faced during this research project was the data availability. The ideal scenario would be to have complete time series for the most relevant variables not only for all countries but regions within countries. Otherwise, we will generate a dataset with few observations and many independent variables (features).

As Machine Learning techniques are only being recently applied to economics, new research paths are endless. In this dissertation, we had studied income inequality, however

we can apply this methodology to other types of inequality (wealth, gender, natural resources, among others).

7. References

Acemoglu, D., & Robinson, J. A. (2000). Why Did the West Extend the Franchise? Democracy, Inequality, and Growth in Historical Perspective*. *The Quarterly Journal of Economics*, *115*(4), 1167-1199. https://doi.org/10.1162/003355300555042

Achten, S., & Lessmann, C. (2020). Spatial inequality, geography and economic activity. *World Development*, *136*, Article 105114. https://doi.org/10.1016/j.worlddev.2020.105114

Aghion, P., Caroli, E., & Garcia-Penalosa, C. (1999). Inequality and Economic Growth: The Perspective of the New Growth Theories. *Journal of Economic Literature*, *37*(4), 1615-1660. https://doi.org/10.1257/jel.37.4.1615

Bloise, F., Brunori, P., & Piraino, P. (2021). Estimating intergenerational income mobility on sub-optimal data: a machine learning approach. *The Journal of Economic Inequality*, *19*(4), 643-665. https://doi.org/10.1007/s10888-021-09495-6

Bourguignon, F., & Morrisson, C. (2002). Inequality Among World Citizens: 1820-1992. *American Economic Review*, *92*(4), 727-744. https://doi.org/10.1257/00028280260344443

Bowles, S., & Gintis, H. (2002). The Inheritance of Inequality. *The Journal of Economic Perspectives*, *16*(3), 3-30. http://www.jstor.org/stable/3216947

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/10.1023/A:1010933404324

Brunori, P., & Neidhöfer, G. (2021). The Evolution of Inequality of Opportunity in Germany: A Machine Learning Approach. *Review of Income and Wealth*, *67*(4), 900-927. https://doi.org/https://doi.org/10.1111/roiw.12502

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA. https://doi.org/10.1145/2939672.2939785

Corak, M. (2013). Income Inequality, Equality of Opportunity, and Intergenerational Mobility. *Journal of Economic Perspectives*, *27*(3), 79-102. https://doi.org/10.1257/jep.27.3.79

Dutt, P., & Tsetlin, I. (2021). Income distribution and economic development: Insights from machine learning. *Economics & Politics*, *33*(1), 1-36. https://doi.org/https://doi.org/10.1111/ecpo.12157

Foster, J., Greer, J., & Thorbecke, E. (1984). A Class of Decomposable Poverty Indices. *Econometrica*, *52*, 761-766. https://doi.org/10.2307/1913475

Friedman, J. H., & Fisher, N. I. (1999). Bump hunting in high-dimensional data. *Statistics and Computing*, *9*(2), 123-143. https://doi.org/10.1023/A:1008894516817

Gilens, M. (2005). Inequality and Democratic Responsiveness. *Public Opinion Quarterly*, *69*(5), 778-796. https://doi.org/10.1093/poq/nfi058

Hailemariam, A., Sakutukwa, T., & Dzhumashev, R. (2021). Long-term determinants of income inequality: evidence from panel data over 1870–2016. *Empirical Economics*, *61*(4), 1935-1958. https://doi.org/10.1007/s00181-020-01956-7

Huang, H.-C., Lin, Y.-C., & Yeh, C.-C. (2009). Joint determinations of inequality and growth. *Economics Letters*, *103*(3), 163-166. https://doi.org/https://doi.org/10.1016/j.econlet.2009.03.010

Jaumotte, F., Lall, S., & Papageorgiou, C. (2013). Rising Income Inequality: Technology, or Trade and Financial Globalization? *IMF Economic Review*, *61*(2), 271-309. https://doi.org/10.1057/imfer.2013.7

Komatsu, S., & Suzuki, A. (2022). The Impact of Different Levels of Income Inequality on Subjective Well-Being in China: A Panel Data Analysis. *Chinese Economy*. https://doi.org/10.1080/10971475.2022.2096809

Kuznets, S. (1955). Economic Growth and Income Inequality. *The American Economic Review*, *45*(1), 1-28. http://www.jstor.org/stable/1811581

McGregor, T., Smith, B., & Wills, S. (2019). Measuring inequality. *Oxford Review of Economic Policy*, *35*(3), 368-395. https://doi.org/10.1093/oxrep/grz015

Neves, P. C., & Silva, S. M. T. (2014). Inequality and Growth: Uncovering the Main Conclusions from the Empirics [Article]. *Journal of Development Studies*, *50*(1), 1-21. https://doi.org/10.1080/00220388.2013.841885

Reardon, S. F. (2013). The widening income achievement gap [Article]. *Educational Leadership*, *70*(8), 10-16. https://www.scopus.com/inward/record.uri?eid=2-s2.0-84878896238&partnerID=40&md5=0271e70987b8080c4fd1076cdf792448

Salas-Rojo, P., & Rodríguez, J. G. (2022). Inheritances and wealth inequality: a machine learning approach. *Journal of Economic Inequality*, *20*(1), 27-51. https://doi.org/10.1007/s10888-022-09528-8

Sawyer, M. (2015). Confronting inequality: review article on Thomas Piketty on 'Capital in the 2st Century'. *International Review of Applied Economics*, *29*(6), 878-889.

https://doi.org/10.1080/02692171.2015.1065227

Sen, A., Sen, A., & Foster, J. (1973). 24Measures of Inequality. In *On Economic Inequality* (pp. 0). Oxford University Press. https://doi.org/10.1093/0198281935.003.0002

Solt, F. (2020). Measuring Income Inequality Across Countries and Over Time: The Standardized World Income Inequality Database. *Social Science Quarterly*, *101*(3), 1183-1199. https://doi.org/https://doi.org/10.1111/ssqu.12795

Truesdale, B. C., & Jencks, C. (2016). The Health Effects of Income Inequality: Averages and Disparities. *Annual Review of Public Health*, *37*(1), 413-430. https://doi.org/10.1146/annurev-publhealth-032315-021606

Wilkinson, R. G., & Pickett, K. E. (2006). Income inequality and population health: A review and explanation of the evidence. *Social Science & Medicine*, *62*(7), 1768-1784. https://doi.org/https://doi.org/10.1016/j.socscimed.2005.08.036