UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
FACULDADE DE CIÊNCIAS ECONÔMICAS
DEPARTAMENTO DE ECONOMIA E RELAÇÕES INTERNACIONAIS

JOSÉ BLOTTA

ANALYZING INEQUALITY OF EDUCATIONAL OPPORTUNITIES IN LATIN AMERICA USING CONDITIONAL INFERENCE TREES AND FORESTS.

Porto Alegre
2022

JOSÉ BLOTTA

ANALYZING INEQUALITY OF EDUCATIONAL OPPORTUNITIES IN LATIN AMERICA USING CONDITIONAL INFERENCE TREES AND FORESTS.

Work presented as part of the requirements for the Bachelor's degree in Economics.

Supervisor: Prof. Dr. Sabino Da Silva Porto Júnior

Porto Alegre
2022

CIP - Catalogação na Publicação

Blotta, José
ANALYZING INEQUALITY OF EDUCATIONAL OPPORTUNITIES IN LATIN AMERICA USING CONDITIONAL INFERENCE TREES AND FORESTS. / José Blotta. -- 2022.
70 f.
Orientador: Sabino Da Silva Porto.

Trabalho de conclusão de curso (Graduação) -- Universidade Federal do Rio Grande do Sul, Faculdade de Ciências Econômicas, Curso de Ciências Econômicas, Porto Alegre, BR-RS, 2022.

1. Desigualdade. 2. Educação. 3. Desigualdade de Oportunidades. 4. América Latina. I. Da Silva Porto, Sabino, orient. II. Título.

Elaborada pelo Sistema de Geração Automática de Ficha Catalográfica da UFRGS com os dados fornecidos pelo(a) autor(a).

JOSÉ BLOTTA

ANALYZING INEQUALITY OF EDUCATIONAL OPPORTUNITIES IN LATIN AMERICA USING CONDITIONAL INFERENCE TREES AND FORESTS.

Work presented as part of the requirements for the Bachelor's degree in Economics.

Aprovado em: Porto Alegre, ___de_____ de 2022.

BANCA EXAMIDORA:

Prof. Dr. Sabino Da Silva Porto Júnior – Orientador
UFRGS

Prof. Dr. Hudson Da Silva Torrent
UFRGS

Prof. Dr. Sergio Marley Modesto Monteiro
UFRGS

ACKNOWLEDGEMENTS

The completion of this work would have been impossible without the support I received from the people whom I love.

I want to thank my family, who made me the person I am today.

I want to thank Júlia, who has always been at my side during this journey, helping me whenever I needed it.

I want to thank my friends, without whom my life would be flavorless.

And finally, I want to thank my father and my grandmother who, despite the distance, always showed me love.

“The future depends on ourselves, and we do not depend on any historical necessity.”

– KARL POPPER

RESUMO

Neste trabalho procuramos analisar empiricamente a desigualdade de oportunidades na educação na América Latina. Para cumprir este objetivo, utilizamos árvores de inferencia condicional e *random forests*. Através dos dados do Latinobarómetro, uma pesquisa de opinião pública realizada em dezoito países da América Latina, ilustramos as raízes do problema analisado estimando uma árvore de inferencia condicional e utilizando um modelo de *feature importance* para cada um dos países disponíveis na base de dados. Ademais, computamos um índice de desigualdade de oportunidades educacionais para cada país. Nossos resultados apontam o nível educacional dos pais como sendo o fator mais importante para o sucesso acadêmico de um indivíduo, sendo esse fato uniforme em todo o continente. Observamos também que a América Central é a região cujos países possuem maior desigualdade de oportunidades educacionais.

Palavras-chave: Desigualdade de oportunidades. América Latina. Educação. Desigualdade de oportunidades educacionais.

ABSTRACT

In this work we strive to analyze inequality of educational opportunities in Latin America. To achieve our objective, we utilize conditional inference trees and random forests. Through the use of *Latinobarómetro* data, a public opinion survey covering eighteen Latin American countries, we examine the roots of inequality of educational opportunities by estimating a conditional inference tree for each country in the dataset, and also estimate an inequality of opportunity score in each analyzed country. Our results point at the parents' years of education as being the most important feature in determining an individual's own educational success in the whole continent. We also observe the Central American region as being the one whose countries have the highest inequality of educational opportunity score.

Keywords: Inequality of opportunities. Latin America. Inequality of educational opportunities.

LIST OF FIGURES

LIST OF CODES

LIST OF TABLES

LIST OF ABBREVIATIONS AND ACRONYMS

EOP	Equality of Opportunity
IOP	Inequality of Opportunity

CONTENTS

1 INTRODUCTION

The role played by education in the distribution of socioeconomic status cannot be underestimated, as it can both lead to an improvement of social mobility and to a reproduction of social inequalities over time. As access to higher levels of education is positively correlated to higher salaries for individuals (BAUM, 2014), the distribution of educational opportunities among a population can show us how other relevant variables, such as income or access to healthcare will be distributed within that same population in the future. With this concept in mind, the study of inequality within education can make us better understand the structural nature of inequalities within a population. However, inequality within itself is not enough to evaluate the quality of distributive justice within a population.

In this work, we follow the argument made by Roemer (1993), according to which inequality is a product of at least two types of inequality: inequality of opportunities (IOP) and inequality of returns to effort (IE). The former refers to inequality coming from exogenous factors beyond the scope of individual responsibility, like race, sex or birthplace, while the latter refers to inequalities stemmed by the personal choices and effort exerted by individuals. In regards to education in specific, inequality of opportunity can be exemplified as a child not having access to university because of his socioeconomic status, while inequality of returns to effort can be exemplified as a child not having access to higher education because he didn't study as much as his peers. While both cases yielded similar results, the cause for each final outcome is radically different.

The concept of equality of opportunity has high support in the general public, and it has been identified as a major goal of distributive justice public policy intervention (ALESINA; STANTCHEVA; TESO, 2018) (BRUNORI; HUFE; MAHLER, 2018). Unfortunately, providing accurate empirical estimates for equality of opportunity is quite challenging, since we need to discern the precise factors that contribute to inequality of opportunity and to have at our disposal a dataset containing retrospective information about individuals.

Latin America is one of the most unequal regions of the world in practically every account (HOFFMAN; CENTENO, 2003). And, with some exceptions, the countries themselves are relatively poor. The consequences of those factors result in a continent in which people live worse off then they need to. Latina America also possesses wide gaps in educational achievements across socioeconomic groups (CRUCES; DOMENCH; GASPARINI, 2014). Because of those problems, it is important to understand the social roots of educational inequality in Latin America, with the objective to improve social and political reforms in the whole region.

Our objective with this work is to carry out an empirical analysis about the

roots inequality of educational opportunity in Latin America, seeking to understand which variables are more important for the determination of educational opportunities in each analyzed country. To achieve this goal, we will utilize conditional inference trees to construct and visualize how the available circumstances affect the educational opportunities available to the population, and random forests to estimate an equality of opportunity score for each country, while also estimating the importance of each available variable in determining the educational attainment achieved by each individual. Our secondary objective will be to evaluate the usefulness of the type of modeling we utilized to help us better understand inequality of opportunity, and, more importantly, if those kinds of machine learning models are feasible in Latin America given the limited availability of longitudinal data present in the continent.

We will discuss the present literature in the equality of opportunity field in chapter 2. chapter 3 introduces the Latinobarómetro survey and explains in greater detail the methodology we utilize, illustrating how we adapt conditional inference trees and forests to the context of inequality of opportunity estimations. In chapter 4 we discuss the results obtained by our modeling, analyzing the trees and feature importance scores obtained in five select countries, while also analyzing the whole continent in section 4.2. chapter 5 concludes the whole paper, and in it we evaluate if our objectives were achieved, while also proposing discussion on future work in the studied topic.

2 LITERATURE REVIEW

The equality of opportunity (EOP) concept has always been a widely discussed topic because, among other reasons, it relates two extremely important aspects: freedom of choice and inequality. By its nature, freedom of choice prevents a perfect natural equality of outcomes, since the effort realized by the individual during his lifetime will clearly impact his own future well-being. Because of this, the topic of Equality of Outcomes (EOE) is considered slightly controversial, on the other hand, the goal of equality of opportunity is much more popular and is pursued and studied by policy makers around the world. In this chapter we will explore and discuss the history and the future of research in Equality of Opportunity, trying to understand the the various types of analysis that can be conducted using this approach, and the advancement of new methodologies in the field.

Authors such as Dubet (2011) question the goal of equality of opportunity, and choose to instead prioritize equality of outcome. Dubet, in particular, criticizes the focus given to inequality of opportunity by policymakers, arguing that the model of equality of results ends up benefiting the most disadvantaged, thus decreasing, consequently, inequality of opportunity as well. Hence, even though there is a much greater focus in the literature on understanding and measure equal opportunity, there is a discussion about what is the best way to understand inequality.

The concept of equality of opportunity has been studied by philosophers and economists over the years, and among the most important contributors we highlight Rawls (1971), as his theory of social justice defines that the distribution of outcomes and social roles is limited by the necessity to take into account the individual's personal choices, since their responsibility and effort would influence their final outcome. Other authors, such as Dworkin (1981), focused on the distinction between preferences and resources, and suggesting that we should seek to correct inequalities caused by differences in resource quantity, not in inequalities caused by personal choices and preferences. The most influential concept, however, is that made by Roemer (1998), whose contribution formalized the definition of inequality of opportunity (IOP), and resulted in the birth and expansion of empirical literature concerning the topic, especially among economists. His definition bases itself on the division between exogenous circumstances over which an individual cannot have control over (such as his place of birth or the educational attainment of his parents) and the individual's personal choices and effort. According to Roemer, inequality between people belonging to people whose initial exogenous circumstances are similar is a morally irrelevant inequality, while inequality between people who have different circumstances is morally unjust. This two types of inequality are respectively defined as within type inequality and between type inequality. Hence, the main objective of public policies that seek to address the

inequality of opportunity problem should be, following Roemer's idea, to level the playing field, that is, reducing the differences caused by exogenous circumstances as much as possible.

Following Roemer's contribution, a considerable quantity of methods have been proposed to measure inequality of opportunity. There are methods such as that developed by Ferreira and Gignoux (2011), that seeks to quantify how much inequality is caused by differences in initial opportunities, while others, such as that of Checchi and Peragine (2010), seek to estimate indices that quantify the impact of inequality of opportunities in society, and there are also authors who conduct statistical tests to detect the existence of inequality caused by differences in opportunities, as was done in Kanbur and Snell (2019). There are two recurring problems faced by researchers who wish to tackle this issue: the need to define the circumstances that are beyond individual control and the need to assume how these circumstances interact with the effort variable to produce different individual outcomes. The outcomes in this case may be variables such as income distribution, years of schooling, and health status, among others.

The study of inequality of opportunity can take a few different directions. The field includes research on public policies, healthcare, education and income, among others. We can explore in many different can be explored in order to better understand our society and how the circumstances of an individual's birth affect their own future achievements.

The study of public policy aims to understand what is the best course of action the government can take to reduce this problem. Generally, there are two different approaches: ex ante and ex post. As analyzed by Fleurbaey and Peragine (2013), this translates into a dichotomy between "compensation" and "responsibility" principles. By the ex ante compensation view, opportunities are evaluated by circumstances and possibilities resulting from the various levels of effort that individuals can exert. In this case, there is equality of opportunity if the set of opportunities is the same for all individuals, regardless of their initial conditions. Thus, Inequality of opportunity decreases if the inequality between sets of opportunities falls. This methodology reduces the degree of personal responsibility, because in this case what really matters is that the opportunity set leaves individuals fully responsible within itself. The ex post compensation approach, however, seeks to use each individual's level of effort to evaluate unequal successes, that is, it seeks to compensate people based on other individuals with different opportunities but the same degree of effort. Clearly, this approach requires an identification of effort variables, and implies that inequality of opportunity falls in the case where inequality of outcomes decreases between individuals sharing the same level of effort. The compensation principle is thus generally formulated with the goal of reducing inequality between individuals sharing the same level of effort but with different

circumstances, while responsibility principles seek to target subgroups with identical circumstances with the goal of raising their opportunity set. These two approaches, as we can see, represent two different ways of addressing inequality of opportunity through public policy.

The educational system possesses an important role in improving opportunities and increasing economical growth in society. therefore, educational policies are also studied in an equality of opportunity viewpoint. Arrow, Bowles, and Durlauf (2018) argue that even a basic concept like equality of educational opportunity can share many meanings, and policy proposals can end up being both moderate and extremist. In general, the most basic objective is to ensure that every young citizen possesses similar educational resources, in order to have the chance to achieve the same goals. there is evidence suggesting that much of the variation in income across different countries can be attributed to variations in ability dispersion Nickell (2004), which further justifies our need to better understand the problem.

Concerning Brazil, research has been made regarding inequality of opportunity in access to higher education Carvalho and Waltenberg (2015) and on individual educational outcomes Ribeiro (2011). One can also study how access to higher quality educational institutions can impact n individual's income in other areas, such as income. It is also possible to study inequality of educational opportunity through intergenerational mobility. Two such papers are Jiménez and Jiménez (2019), in which Latinobarómetro data is used to estimate education inequality indices in Latin America, and Mahlmeister et al. (2019), in which intergenerational education mobility in Brazil is analyzed, because, as claimed by the authors, a parent's education affects the education that their child will eventually attain, and understanding the severity of this influence can aid in the formulation of public policy.

The study of inequality of opportunity in education, unfortunately, is bounded by structural limitations, due to the rarity of extensive and reliable databases where the researcher can conduct his analysis.

Social and economic conditions, some of which arise from the circumstances of birth of individuals, tend to influence their health status. Clearly, this effect is more intense in the poorest parcel of the population. Among the studies on inequality of opportunity in health we can highlight Sreenivasan (2007) and Donni, Peragine, and Pignataro (2014), whose work investigates the possible definitions on inequality in healthcare, and seeks to combine the equal opportunity approach with the methods commonly utilized in health economics that measure social and income inequalities in healthcare. Most of the literature, however, focuses on socioeconomic and healthcare system inequalities seeking to explain health inequalities through differences in factors such as living or working conditions, access to health, or growth environment.

Inequality of opportunity with a focus on income is also often analyzed in con-

junction with intergenerational mobility, which is a related way of studying inequality of opportunity, as it analyzes the income of the son in relation to that of the father, thus seeking to understand how the correlation between the two. This method is commonly used as a basis in most works on the subject. Generally these papers seek to study how inequality of opportunity affects income distribution in a specific country or region. Aaberge, Mogstad, and Peragine (2011), for example, estimate a welfare index based on the equal opportunity approach using Norwegian data.

It is also important to highlight the impact that a high inequality of opportunity index can have on a country's economic growth. There is some controversy in the discourse concerning the relation between inequality and growth. A literature that addresses the causal relationship that economic growth has on inequality of opportunity exists, according to which economic development and growth alone would end up reducing income inequality, such as Kuznets (2019) and Milanovic (1994), but there are also a number of studies that analyze the inverse causal relationship, that is, how a reduction in income inequality can cause economic growth. This debate on equality of opportunity allows us to look at this problem more closely, thus being able to distinguish inequality caused by differences in individuals' initial circumstances. One of the most common arguments about the negative effect that inequality of opportunity has on economic growth is that when exogenous circumstances such as gender, ethnicity, or place of birth strongly influence individual income and possible future employment there exists a suboptimal allocation of resources, because a citizen who, for example, had the potential and desire to become a doctor ends up being prevented from accessing university since his place of birth and his parents' income prevented him from following that career. Population groups being excluded from participating in economic activities are, as we can imagine, detrimental to economic growth.

As measuring inequality of opportunity is not a trivial task, researchers end up using different, creative techniques depending on what data they have at their disposal and what specific relationship they want to capture. A researcher, wanting to compute equality of opportunity, encounters several difficulties. First, he needs to decide in which outcomes he wants to focus on, and to understand which variables represent circumstances and which show effort. The problem with this division is that it depends a lot on the researcher's own understanding, and a choice that is not entirely fair can derail the results significantly. We also know that not all circumstances can be observed, and often there is not enough data to conduct a good analysis. Some problems can be solved using parametric or non-parametric techniques. According to Roemer (2002), if you have many observable variables, parametric estimation will marry the data better. Those methods seek to estimate the conditional expectation, while non-parametric methods seek to estimate the conditional distribution, hence they end up being more "ambitious", and have the advantage that they are not limited by data availability, and

are usually solved by regression analysis.

The measurement of equal opportunity indices can be done with methods such as measuring the Shapley index, as was done by Dill and GONÇALVES (2012) and Carvalho and Waltenberg (2015), or using multidimensional matrices that seek to relate the variables chosen as circumstances with the effort variables, so as to have as a result a single value that represents how unequal a certain region is in the Roemerian view.

To overcome the problem of choosing circumstances, there have recently been some attempts to use regression tree algorithms, or other machine learning methods, to estimate indices of inequality of opportunity. These methods are fairly new in empirical inequality research, so we do not yet know how successful they may be in the future, but they have, among the advantages of these methods, the ability to choose which circumstances are important to analyze and how to analyze them. However, the disadvantages include not being very explanatory, as the researcher must try to understand the results without any specific indication of the model, and also not being very robust yet, as these methods are still new in academic research. Works that have used these methods include Brunori, Hufe, and Mahler (2018) and Brunori and Neidhöfer (2021).

The study of inequality of opportunity is able to give us better perspectives on inequality, making us understand better the way mechanisms that generate those inequalities work in a more general sense. We also know that, as pointed out in Bank (2005), the degree of inequalities caused by circumstances can be related to economic growth and development. There are "inequality traps", which end up excluding entire socioeconomic groups of the population in the economic and social participation of society, which in turn ends up harming it, are a result of a high level of inequality of opportunities. One of the reasons for the inconclusiveness of the literature on inequality and growth is also explained by the nature of the concept of inequality, which ends up ignoring the difference between inequalities that are explained by conditions exogenous to the choices of individuals and those that are the result of their conscious choices.

3 METHODOLOGY

Our primary objective consists in understanding how the circumstances surrounding the birth and infancy of an individual, such as ethnicity, citizenship, or educational attainment of their parents, end up shaping his or her own personal educational attainment.

To achieve the goals stated, we will be adapting the method used in Brunori, Hufe, and Mahler (2018) by using conditional inference trees and random forests to estimate inequality of opportunity and, in addition to that, we will be estimating the conditional importance of each variable utilizing the method developed by Strobl et al. (2008). This will allow us to visualize the relations between circumstances and results in a new light. By predicting the results of different identifiable groups, we connect to Roemer's idea of equality of opportunity.

In this chapter we will start by discussing why we chose the Latinobarómetro Survey to be our data source, and what are the survey's limitation for the type of analysis we wish to conduct. In section 3.2 we will explain how tree-based methods work and how we can utilize conditional inference to improve the partitioning method. In section 3.3 we will introduce Random Forests, which will be utilized both to estimate conditional variable importance and inequality of opportunity estimates for each country. The methodology utilized to estimate both the conditional variable importance and the inequality of opportunity estimates will be discussed in subsection 3.3.1 and subsection 3.3.2.

We will utilize the *r* package partykit to develop the conditional inference trees and forests for each country and the package ineq to estimate the Gini index.

3.1 DATA

We utilize data from Latinobarómetro, a public opinion survey conducted quasi-annually in Latin American countries, covering eighteen countries in total. The survey is conducted since 1995, but only became representative of the surveyed population in 2003.

The main reason for the Latinobarómetro use is that, other than being a continent-wide survey, covering all countries equally, it also includes the retrospective information about the education of the parents for every individual surveyed, regardless of their place of residence. Unfortunately, the Latinobarómetro neither includes accurate data about income nor about other retrospective information regarding the interviewed individual, like the occupation of his father or the presence of both parents at home, whose presence would certainly improve our understanding of inequality of opportunity in Latin America, as those variables are utilized in similar works concerning other continents (BRUNORI; HUFE; MAHLER, 2018).

We utilize data from all the available surveys for the period 2011-2020[1]. In our analysis, we will consider only individuals from 25 to 64 years of age, as 25 is generally the year in which formal education is concluded.

3.1.1 Feature Selection and Data Cleaning

While Latinobarómetro is primarily a public opinion survey, we will utilize only socioeconomic data. The circumstances included are: sex, ethnicity, an indication whether the individual is a citizen of the country he resides in, the parents' years of education, the size of the city he lives in, and his religion. Our outcome of interest is the completed years of education of the respondent.

The data formatting for city size and religion utilized in this work is available in Table 1 and in Table 3.

Table 1 – City size coding in Latinobarómetro

Code	Inhabitants in Respondent's city
8	Capital City
7	More than 100.000 inhabitants
6	50.001 - 100.000
5	40.001 - 50.000
4	20.001 - 40.000
3	10.001 - 20.000
2	5.001 - 10.000
1	Less than 5.000 inhabitants

Source: Latinobarómetro Codebook.

Latinobarómetro only counts numeric years of education until the twelfth year; respondents with a higher education are labeled with a textual description of their education, such as "Incomplete university", 'High school/academies/Complete technical training". Because of that, we have to estimate the amount of years these labels account for. The changes made are described in Table 3.

These changes are applied to both the years of education of the respondent and their parents. We also change the ethnicity variable to be a binary variable to be an indication on whether the respondent is white or not. We provide the python code utilized for the data cleaning process in Appendix C.

[1] The survey was not conducted in 2012, 2014 and 2019.

Table 2 – Religion coding in Latinobarómetro

Code	Religion
1	Catholic
2	Evangelic without specifications
3	Evangelic Baptist
4	Evangelic Methodist
5	Evangelic Pentecostal
6	Adventist
7	Jehovah Witness
8	Mormon
9	Jewish
10	Protestant
11	Afro-American Cult, Umbanda, etc
12	Believer, not belonging to any church
13	Agnostic
14	Atheist
96	Others

Source: Latinobarómetro Codebook.

Table 3 – Years of education per category

Respondent Type of education	Estimated years of education
Incomplete University	14
Complete University	17
High school/academies/Incomplete technical training	14
High school/academies/Complete technical training	16

Source: Estimated by the author.

3.1.2 Sample Size

After removing missing data and respondents outside the analyzed age group, the sample size for each country is as follows in Table 4

While some Central American countries have smaller sample sizes than the rest, every country has a similar amount of observations.

3.2 TREE-BASED METHODS

Prediction trees are a supervised learning method that can be applied to both regression and classification problems. The process of building a prediction tree can

Table 4 – Sample Size for Each Country

Country	Nº of Observations
Argentina	4395
Bolivia	4862
Brazil	5132
Colombia	4904
Costa Rica	3887
Chile	4861
Ecuador	4947
El Salvador	3665
Guatemala	3866
Honduras	3773
Mexico	5234
Nicaragua	3415
Panama	3595
Paraguay	4291
Peru	4969
República Dominicana	3311
Uruguay	4461
Venezuela	4898

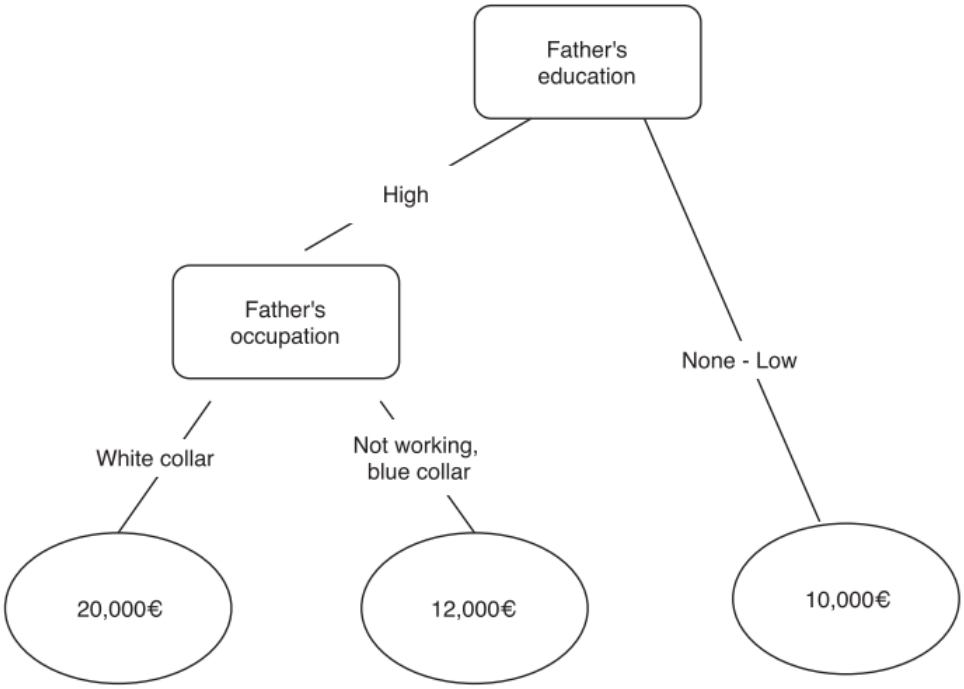
Source: Estimated by the author using Latinobarómetro Data.

roughly be divided in two steps: partitioning the data and predicting each partition's result. We will start by explaining the process of recursive partitioning.

Conceptually, given a set of variables $X : x_1, x_2, ..., x_i$ and an outcome y, we partition the predictor space into smaller non-overlapping regions. After that, we partition those regions again and continue doing so until we meet a predetermined stopping rule. The result of this process is a partition of the predictor space into N regions $R_1, R_2, ..., R_j$. This process can be represented visually, like the name implies, by an upside-down tree. Each terminal node, or leaf, represents a partition of the data, and each branch represents a data split. A point c is part of a terminal node if it falls in the corresponding cell of the partition. To understand how a specific point c gets assigned to a specific leaf, we have to start at the root node of the tree, and ask a certain sequence of questions about its features. Each interior node is labeled by a question, and each branch between two nodes is labeled by an answer. Using the simplified decision tree in figure 1, we can see that the first split of the data is made regarding the *Father's education* variable, after which the tree is split into two nodes, depending on the answer to the first question.

Once we have a complete tree, we predict the result of each terminal node.

Figure 1 – Simplified Example of a Decision Tree



Source: Brunori and Neidhöfer (2021). The graph is an example of a regression tree predicting an individual's income based on the education and occupation of their father.

Since we will have many relatively small data clusters in each leaf, local models are easy to find. In simple predictive trees, the mean of the response for each observation inside a terminal node can be enough. Note that this approach is very useful at making out-of-samples prediction of a dependent variable.

3.2.1 Recursive Binary Splitting

Dividing the predictor space in smaller regions is arguably the most important step in decision tree construction. The technique known as recursive binary splitting is a widely popular tool for this type of analysis. To understand how this method works, we have to describe more thoroughly the first step in creating a decision tree, that is, the partitioning of the entire predictor space.

When we construct the regions $R_1, ..., R_j$, the goal of recursive binary splitting is to find regions that minimize the *residual sum of squares* (RSS), given by

$$RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{Rj})^2 \tag{1}$$

Where $\hat{y}$ is the mean response for the observations within the jth region. Since it is computationally impossible to consider every single possible split of the predictor space into J regions, recursive binary splitting is most commonly used (HASTIE et al., 2009).

This method begins at the top of the tree, that is to say, with the whole predictor space. Consider a splitting variable j, and a split point s. We then define a pair of half-planes such that

$$R_1(j,s) = \{X \mid X_j \leq s\} \text{ and } R_2(j,s) = \{X \mid X_j > s\} \tag{2}$$

after which we seek both the splitting variable j and the split point s that minimizes the following equation:

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R2})^2, \tag{3}$$

Essentially, we are splitting the predictor space into the two regions that lead to the greatest possible reduction in *RSS*. One advantage of this method is that it allows us to determine the most optimal pair (j,s) very quickly. After having partitioned the data once, we repeat the process on each of the new regions. The process is repeated iteratively until a certain predetermined stopping requirement is met. The specific size of the optimal tree depends largely on the available data.

While this is an effective and efficient method, it is not without its flaws. In particular, Hothorn, Hornik, and Zeileis (2006) cite overfitting and selection biased towards covariates have long been recognized as two fundamental problems in exhaustive search procedures such as recursive binary splitting. Because of this, they propose a framework using conditional inference to partition regression trees.

3.2.2 Conditional Inference Trees

In this subsection we will explain the method proposed by Hothorn, Hornik, and Zeileis (2006), which we will ultimately use in our tree construction process. One of the main problems with the standard method of recursive binary partitioning is the lack of a concept of statistical significance. To solve that, the conditional inference framework changes the way the partitioning of the predictor space is made.

Just like in the method described in the previous section, we need to select a certain variable j and a splitting point s from where we conduct the first split of the predictor space. We start by checking the density function independence between each covariate and the response variable, obtaining a p-value for each test. To do this, we test the null hypothesis of density function independence $H_0^j : D(\mathbf{Y} \mid X_j) = D(\mathbf{Y})$ for each X_j in the variable space, obtaining a p-value for each test conducted, p^{X_j}. We select the variable j with the lowest p-value, and, if that p-value is lower than a previously chosen significance level α, we select that as the splitting variable.

Similarly, to select the splitting point we test the null hypothesis of density function independence between the sub samples given by each binary splitting point s, also obtaining a p-value associated with each test. Then, we partition the data based on the

splitting point *s* which yields the lowest p-value. After that, just like the recursive binary splitting method, we repeat the process for each resulting sub-sample, stopping in case none of the possible splits result in a p-value lower than α.

Our chosen level of α will be 0.05.

3.3 RANDOM FORESTS

Decision trees, unfortunately, still suffer from some shortcomings. One of the most important ones is that trees alone can be very non robust, since small changes in the data can cause very large changes in the final estimated trees. This can happen, for example, in case two or more very important variables are highly correlated, as splitting the data using one of the two means that the other would likely not have enough information to cause another split (GARETH et al., 2013). Because of those problems, methods like boosting, bagging and random forests utilize trees as building blocks to construct powerful prediction models. The following part of this section will explain in greater detail the inner workings of random forests, as this will be the method utilized to achieve our stated objectives.

Forests loosely are, like the name implies, a collection of many decision trees. To construct a forest, we first need to build a number of decision trees and then average their resulting predictions. But the tree-building process in random forests is slightly different. Since the tree-building process is non-random, constructing two trees on the exact same set of data and utilizing the same variables will result in two exactly equal trees. To create trees that differ from each other, we make two specific changes: First, for each tree built, only a random subsample of the data is used. And second, only a random subset of variables is allowed to be used on each splitting point. Those two techniques increase the likelihood that every circumstance with informational content will be eventually utilized as a splitting variable, and improve the robustness of the model.

Thus, by forcing only a subset of the full data, the model guarantees that the estimated trees will not look similar to each other and the correlation between different trees will not be high. This also makes it so that that the average of the resulting trees tends to be more reliable.

3.3.1 Conditional Feature Importance

While supervised machine learning models can have strong predictive power, many algorithms construct complex models that are too opaque for humans to understand effectively. Because of this limitation, the field of Explainable Artificial Intelligence

(XAI) has emerged in recent years (BURKART; HUBER, 2021). In this work, we intend to use the method developed by Strobl et al. (2008) to estimate the importance score of each variable utilized in our random forest model.

Feature importance models attempt to estimate how much each variable in a random forest contributes in the out of sample prediction accuracy achieved by the model. The process developed by Strobl et al. (2008) estimates the importance of each variable by permuting it within a grid defined by the covariates whose correlation to the determined variable is higher than a certain predetermined threshold, and computing the out of sample model accuracy, then comparing it to the original out of sample prediction accuracy of the original model. after doing this operation for each tree in the forest, the importance score for our given variable X_j will be the mean difference between the original prediction accuracy and the post-permutation prediction accuracy over all trees.

3.3.2 Calculating Inequality of Opportunity using the Gini Index

One of our objectives consists on utilizing the models we developed to estimate a score of educational inequality of opportunity in each analyzed country. To achieve this, we follow the methodology utilized by Brunori, Hufe, and Mahler (2018). The process will consist in running the models on each individual country's data, and then applying the resulting prediction function $\hat{f}()$, that is, the random forests developed for each country to obtain an estimate of the conditional distribution of opportunities in the population $\hat{y}^C$.

Once we have our distribution estimate, we summarize it with the Gini index, thus obtaining an inequality of opportunity estimate for each analyzed country.

4 RESULTS

The educational attainment of a person is determined not only by his own effort, but also by circumstances that do not depend on the individual's deliberate choices. Inequality of opportunity in education exists if outside circumstances, like ethnicity, gender or citizenship, end up influencing the end result of an individual's educational attainment level.

Our objective is to capture how ethnicity, gender, parents' years of education, religion, citizenship and city size influence the years of education completed by individuals in working age over 18 countries in Latin America and the Caribbean.

In this chapter, we will discuss the results obtained by our modeling, by analyzing the specific findings related to five select countries; the three biggest economies on the survey (Brazil, Mexico and Argentina) and the ones with the highest IOP index (Nicaragua and Guatemala).

In section 4.1 we will analyze the conditional inference trees obtained following the method proposed by Hothorn, Hornik, and Zeileis (2006) for each of those 5 select countries and discuss the calculated importance for each variable utilized in the construction of our forests. In section 4.2 we will visualize the IOP index results for each of the 18 countries in the dataset.

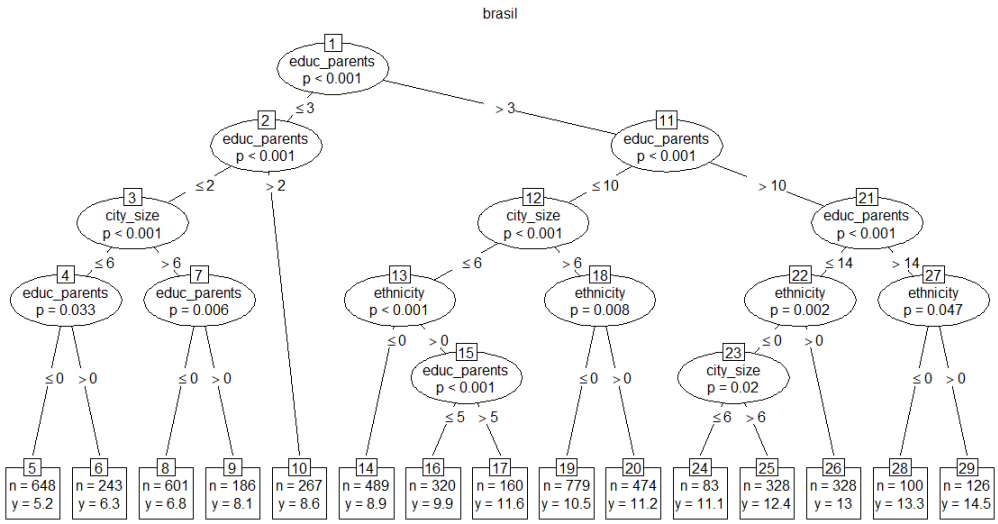
4.1 VISUALIZING OPPORTUNITY TREES AND FEATURE IMPORTANCE

Utilizing the methodology described in section 3.2.2, we estimated and plotted 18 conditional inference trees, one for each Latin American country in the survey. We will show five of those trees in this section. The rest will be available in the annex A.

While Brazil's GDP is one of the ten highest in the world, the country still has a large quantity of people living in absolute poverty (BARBOSA; SOUZA; SOARES, 2020) ; this is reflected both on income and education. Specifically, as of 2013, only 12 percent of people in university age were studying in higher education, and as explained by Carvalho and Waltenberg (2015), while in the previous decade the affirmative action policies improved the equality of educational opportunities in the country, there still is a long way to go, as the family income and the educational attainment of the chief income earner are the most important factors in determining the educational possibility of an individual.

Figure 2 shows the opportunity tree for Brazil. Notice that seven out of the 23 internal nodes was a binary split made by partitioning the data based on the parents' educational attainment. In this case, we can see how the completed years of education of an individual in Brazil depends on his parents' completed years of education more than any other computed variable. While our dataset doesn't include family income, the

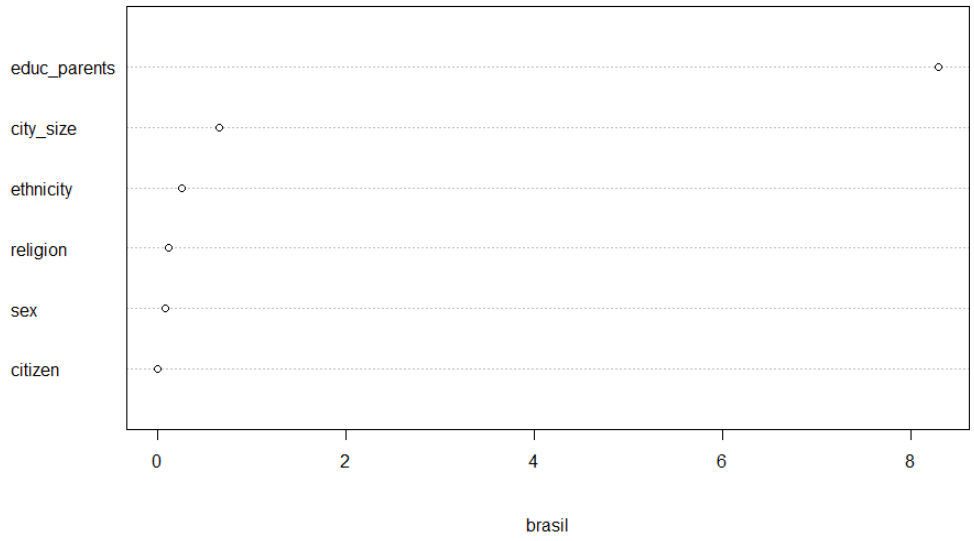
Figure 2 – Educational Opportunity Tree for Brazil.



Source: Elaborated by the author using Latinobarómetro data. Note: The reference codes for religion and city size are available in subsection 3.1.1. Sex: 1 = female, 0 = male. Citizen : 1 = Has the country's citizenship, 0 = does not have the country's citizenship.

importance on parents' years of education was in line to the research made by Carvalho and Waltenberg (2015). We can also see that both ethnicity and city size play a part, as nonwhite people and people living in small cities yield a smaller expected result.

Figure 3 – Conditional Feature Importance for Brazil.

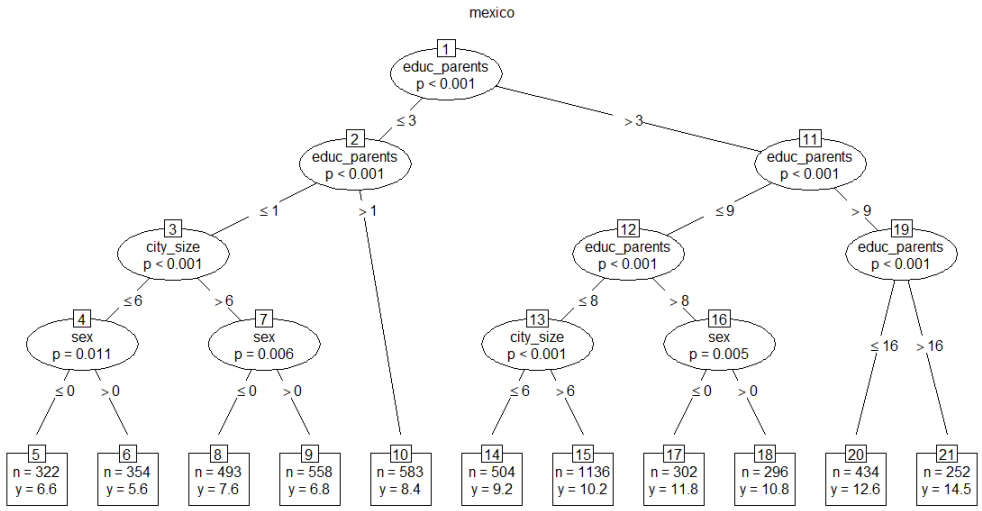


Source: Elaborated by the author using Latinobarómetro data

Looking at the feature importance in Figure 3, we can see that, as expected, the parents' educational level is the most important variable in determining the individual's

years of education. A distant second is city size, followed by ethnicity. Religion, sex and citizenship, on the contrary, are not good predictors of educational attainment in Brazil.

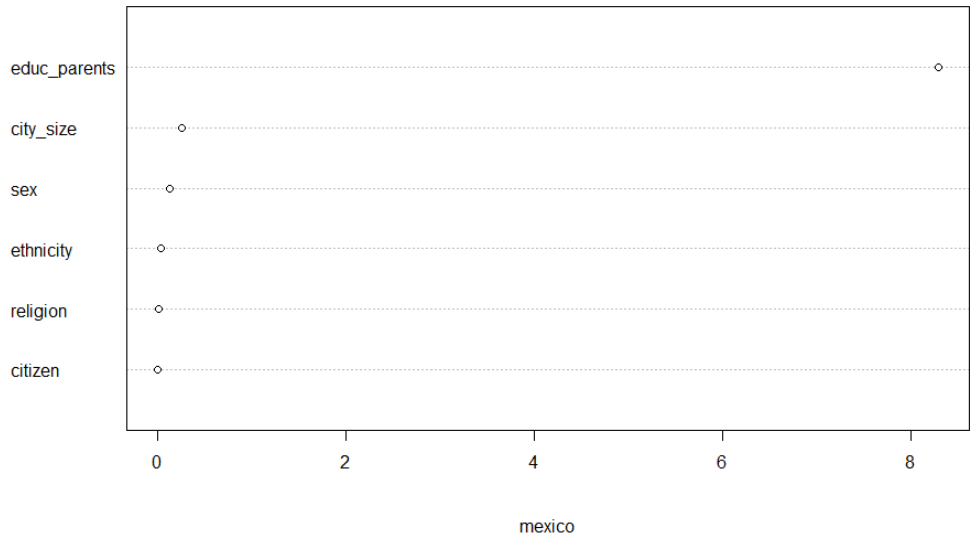
Figure 4 – Educational Opportunity Tree for Mexico.



Source: Elaborated by the author using Latinobarómetro data. Note: The reference codes for religion and city size are available in subsection 3.1.1. Sex: 1 = female, 0 = male. Citizen : 1 = Has the country's citizenship, 0 = does not have the country's citizenship.

Mexico is the second biggest economy in Latin America. Despite this, just like Brazil, the country has high levels of inequality, and the inter-generational social mobility rates are very low at the extremes of the wealth distribution (GRAJALES; MONROY-GÓMEZ-FRANCO; YALONETZKY, 2018).

Figure 5 – Conditional Feature Importance for Mexico.

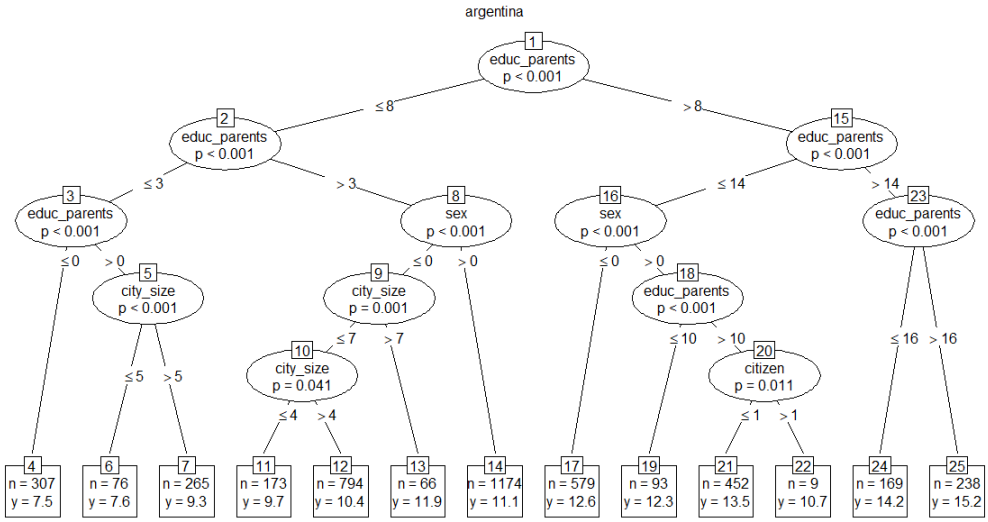


Source: Elaborated by the author using Latinobarómetro data

In Figure 4 we can see the opportunity tree constructed for Mexico. Unlike the Brazil tree, we can see that, when the parents' education is higher, males tend to spend more years in the school system, while when the parents educational level is low, the opposite happens, with females having an higher predicted result, as we can see in nodes 8 and 9.

Figure 5 shows that even in Mexico the number of years of education is explainable in large part by the parents' level of education. Unlike Brazil, ethnicity has very little importance, while sex is slightly more important.

Figure 6 – Educational Opportunity Tree for Argentina.



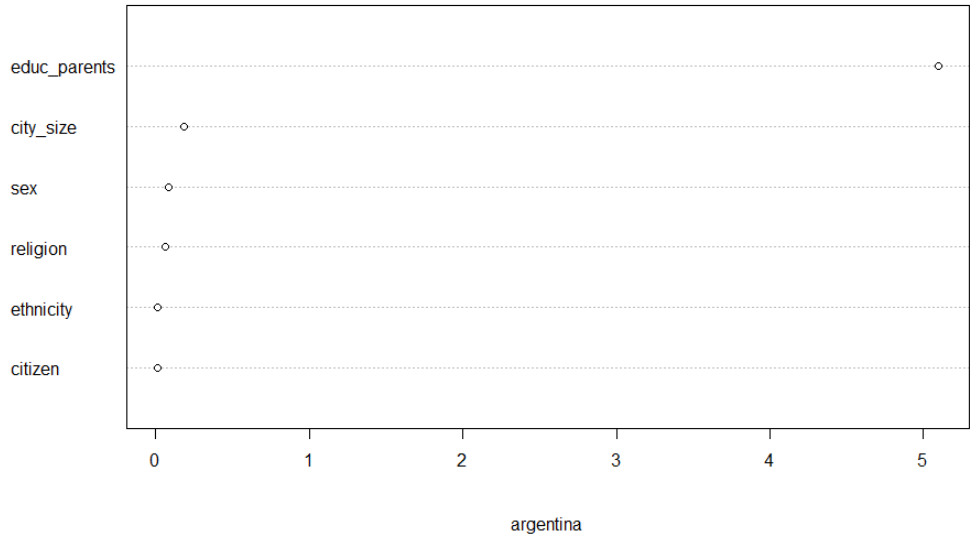
Source: Elaborated by the author using Latinobarómetro data. Note: The reference codes for religion and city size are available in subsection 3.1.1. Sex: 1 = female, 0 = male. Citizen : 1 = Has the country's citizenship, 0 = does not have the country's citizenship.

Figure 6 shows the opportunity tree in Argentina. Parents educational attainment is again the most important thing, as shown in Figure 7, followed by city size and sex. In general, Argentina's tree results and feature importance are very similar to Mexico's. In node 20, we can see how non-citizens with educated parents have a much smaller expected results than Argentinian citizens.

The three countries analyzed, being the biggest economies in Latin America, are not very different between themselves in terms of educational inequality structure. While we will discuss their estimated IOP level score in section 4.2, we will now analyze the two countries with the highest estimated inequality of opportunity: Guatemala and Nicaragua.

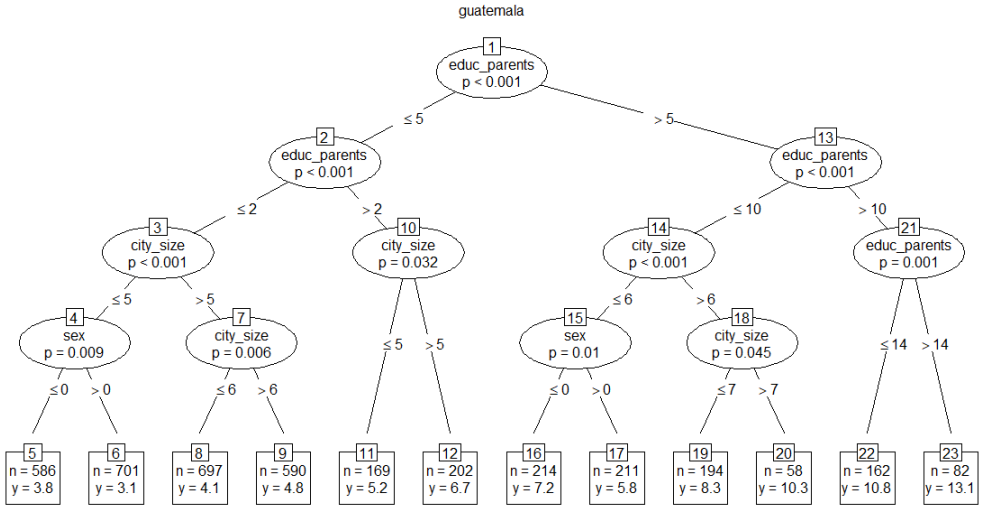
Guatemala is the Latin American country with the highest incidence of poverty, with 35.9 percent of the population living in extreme poverty and 52.6 percent living in poverty as of 2010 (CABRERA; LUSTIG; MORÁN, 2015). In Figure 8 we can see Guatemala's opportunity tree. Unlike Mexico, Argentina and Brazil, the distribution is

Figure 7 – Conditional Feature Importance for Argentina.



Source: Elaborated by the author using Latinobarómetro data

Figure 8 – Educational Opportunity Tree for Guatemala.



Source: Elaborated by the author using Latinobarómetro data. Note: The reference codes for religion and city size are available in subsection 3.1.1. Sex: 1 = female, 0 = male. Citizen : 1 = Has the country's citizenship, 0 = does not have the country's citizenship.

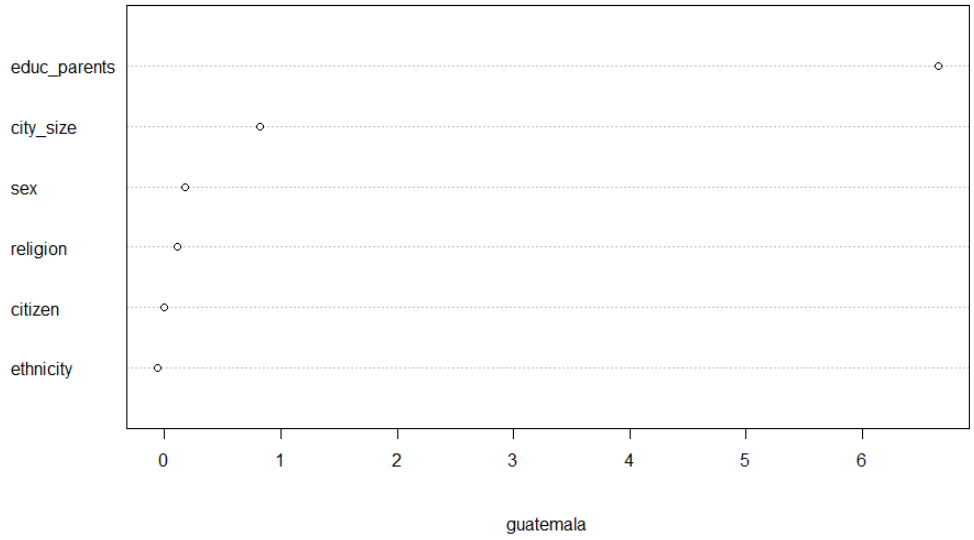
clearly very unequal between the educated and the non-educated people, as the model gives us very low predictions in the left side of the tree, like in node 5 and 6.

While Cabrera, Lustig, and Morán (2015) cites the extreme poverty of the indigenous population of Guatemala as being one of the main reasons for the country's high inequality, ethnicity appears in Figure 9 as being the least important factor in determining the individual's education level. One of the possible reasons for this oversight is the

way that ethnicity has been labeled: since it is only differentiated between whites and nonwhites, the inequality specific to the indigenous population is not noticeable, as the indigenous population blends with the hispanic population in the data, and only 26 percent of the population in the survey identifies as white.

The educational attainment of the parents, as usual, is by far the feature with the highest importance for Guatemala, but city size is unusually high. This makes us understand that the educated population is more concentrated in the larger cities in Guatemala than in other Latin American countries.

Figure 9 – Conditional Feature Importance for Guatemala.



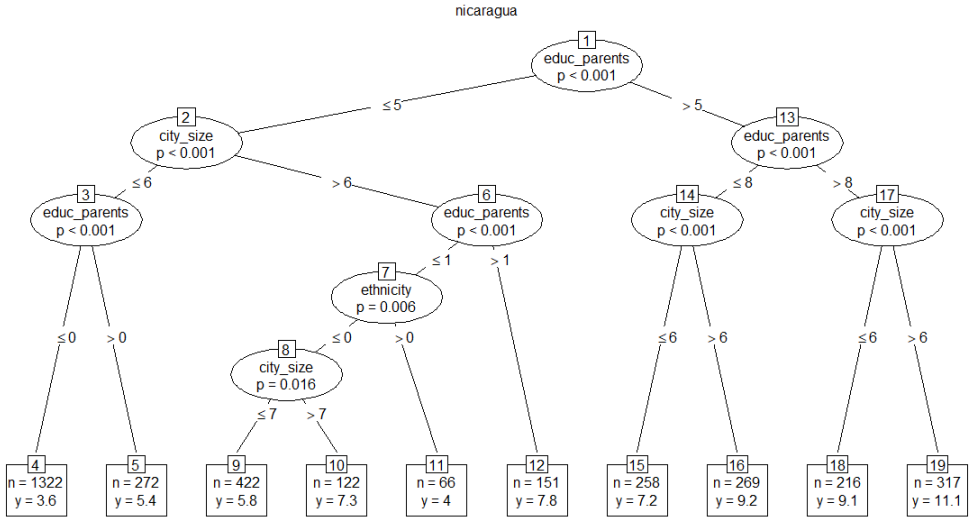
Source: Elaborated by the author using Latinobarómetro data

Nicaragua is the country with the highest estimated inequality of opportunity by our model. In literature, there has been a focus explicitly on the spatial side of inequality, as we can see in Wall (1993), the Nicaraguan economy has been viewed as excessively concentrated in Managua. This is made clear by looking at Figure 10, as living outside the capital always results in a lower expected education.

In Figure 11 we can see that the feature importance of city size is higher than 2, a number much higher than Guatemala and all the other observed countries. We can infer that, according to our modeling, other than the ever-present parents' educational attainment, the educational inequalities in the most inequal countries of Latin America are mostly spatial.

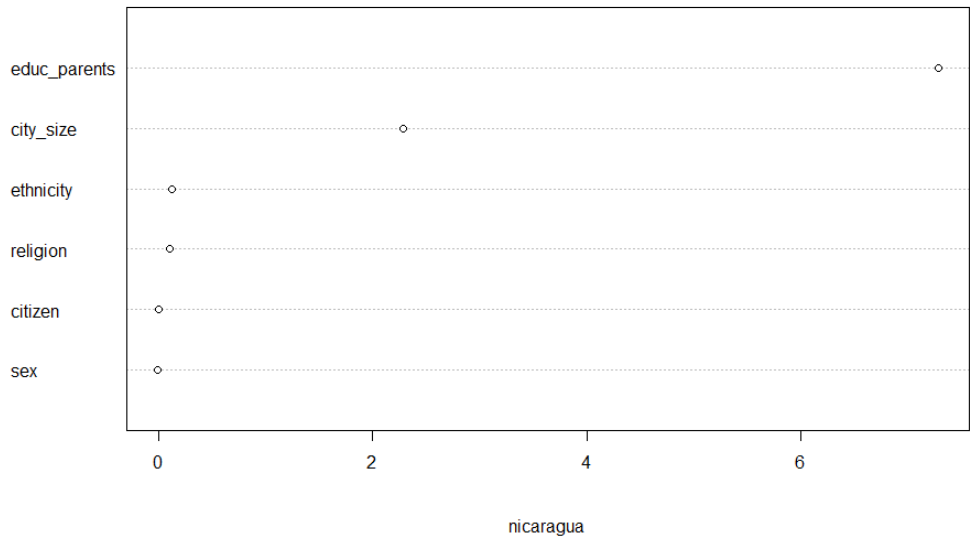
One important thing to underline is that conditional inference regression trees should be interpreted with caution, as they tend to be very sensitive to sample size. Since we are also working with imperfect data, looking just at the trees without proper context can cause mistaken conclusions to arise. Nevertheless, the visualization and

Figure 10 – Educational Opportunity Tree for Nicaragua.



Source: Elaborated by the author using Latinobarómetro data. Note: The reference codes for religion and city size are available in subsection 3.1.1. Sex: 1 = female, 0 = male. Citizen : 1 = Has the country's citizenship, 0 = does not have the country's citizenship.

Figure 11 – Conditional Feature Importance for Nicaragua.



Source: Elaborated by the author using Latinobarómetro data

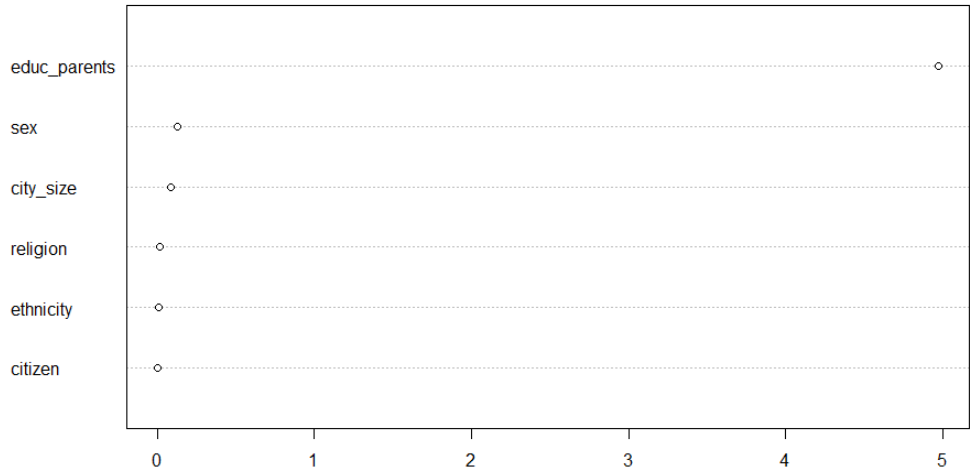
the general characteristics of the tree can certainly help us understand how inequality of opportunity works in the analyzed countries.

Parents' educational attainment, as we can see, is clearly the most important predictor when looking at an individual's level of education. This is expected, since there are works focusing solely on this variable to estimate inequality of educational opportunity (JIMÉNEZ; JIMÉNEZ, 2019).

4.2 THE BIG PICTURE

In section 4.1 we analyzed the results in five select countries. We now focus on the whole continent of Latin America. Figure 12 shows how parents' educational attainment is the most important variable not only across the countries previously analyzed, but in the whole Latin American continent. Citizenship, on the other hand, has almost no use in predicting years of education in our dataset.

Figure 12 – Conditional feature Importance for Latin America.

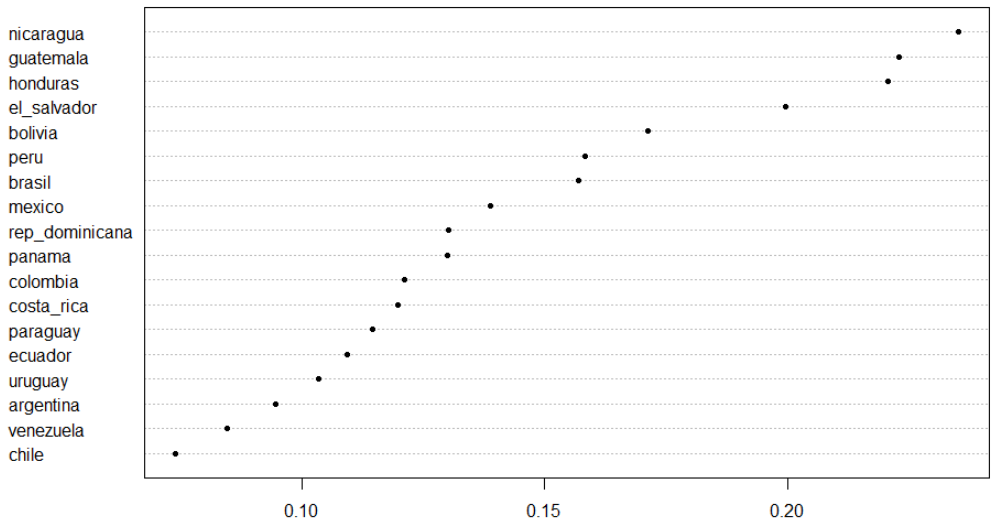


Source: Elaborated by the author using Latinobarómetro data

Figure 13 presents the IOP index calculated for every Latin American country in the Latinobarómetro survey. These results were computed utilizing the method specified in section 3.3.2. A score between 0 and 1 is calculated for each country, with 0 representing complete equality and 1 complete inequality. In our case, a country with a low score is a country where the circumstances reflecting a person's birth have little importance in determining the educational attainment of an individual, while a country with a high score is the opposite, that is, one in which birth circumstances play a big part in the educational opportunities of its inhabitants.

We plot the countries in a way to rank them from highest IOP to lowest. As stated in section 4.1, the two countries with the highest estimated educational IOP are Nicaragua (0.235) and Guatemala (0.222). We can also see that the four highest scores are all from Central American countries. Despite this, the other two countries from the region, Panama (0.129) and Costa Rica (0.119) both have scores lower than the mean value (0.143). The countries with the lowest score are Chile (0.074) and Venezuela (0.085). The countries whose score was closest to the mean were Brazil (0.157) and

Figure 13 – Estimated IOP scores for Latin American countries.

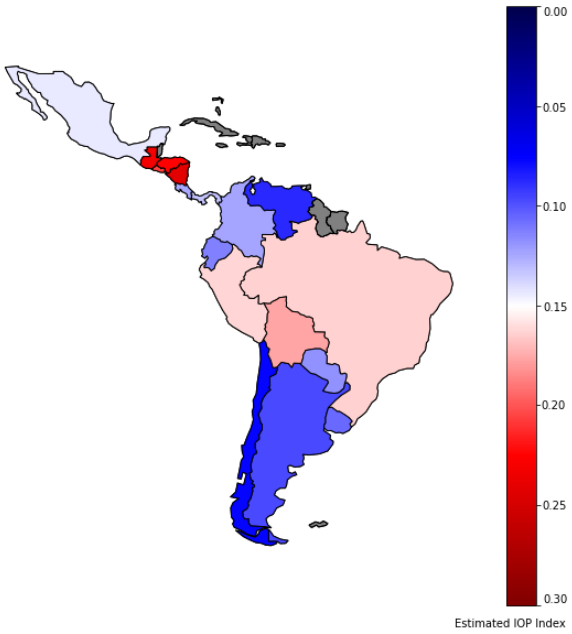


Source: Elaborated by the author using Latinobarómetro data and the methodology proposed in subsection 3.3.2.

Mexico (0.139).

We also visualize the estimated inequality of opportunity scores using a map plot in Figure 14. We can see how the countries with the highest IOP are all located in Central America.

Figure 14 – Estimated IOP scores for Latin American countries.



Source: Elaborated by the author using Latinobarómetro data and the methodology proposed in subsection 3.3.2.

The results obtained show that we have great regional heterogeneity in inequality of educational opportunity in the continent, as the scores obtained differ substantially between countries, with some being greatly unequal, while others less so.

5 CONCLUSION

In this work, we analyze inequality of educational opportunity in Latin America utilizing data from ten years of Latinobarómetro surveys. By applying conditional inference trees, we visualize an estimate of opportunity trees for each of the eighteen countries analyzed, gaining insights about the roots of educational inequality. Computing feature importance for each variable in each country allows us to determine how much these factors end up influencing the final educational result of individuals surveyed. We also estimate an inequality of opportunity score for each country in the dataset. This lets us compare the educational inequality structure between the eighteen analyzed countries.

Our results show that in every analyzed Latin American country the parents' years of education is by far the most important factor in determining an individual's educational level. While the biggest economies in the continent present similar tree structures and feature importance scores, some smaller and more unequal countries possess slight differences caused by specific factors intrinsic to their own socioeconomic organization, such as Nicaragua's economy being focused almost solely on the capital, causing individuals born and living in other cities to have a worse expected educational outcome.

The estimated inequality of opportunity scores exhibit Central America as the region with the highest level of educational inequality, since all four countries with the highest inequality score all come from the same aforementioned region. The countries with the lowest estimated educational IOP score are Chile and Venezuela.

While our methodology provided us with coherent results, the scarcity of more retrospective information about interviewed individuals limits the possible scope of future works in the field. More data concerning the individual's infancy is extremely important in inequality of opportunity research, as variables such as the father's occupational status, the presence of both parents at home or the tenancy status in the house the respondent was living when they were a child are widely used in IOP research (BRUNORI; HUFE; MAHLER, 2018), (MARRERO; RODRIGUEZ, 2013). Because of those limitations, our results may not be as accurate as we would like, as circumstances like family income can be extremely useful when trying to understand inequality of educational opportunities.

Another limitation caused by our general methodological work is visible in section 4.1. As the data was uniform for the whole continent, specific patterns present in some countries can end up not being recognized by our modeling as they should, as it is the case of Guatemala. As explained by Cabrera, Lustig, and Morán (2015), *"an indigenous individual in Guatemala is more than twice as likely to be in poverty than a non-indigenous one"*, and *"the large poverty gap between indigenous and non indigenous individuals is highly correlated with the disparities in educational attainment by ethnicity"*. But, as shown in Figure 9, ethnicity is not considered a relevant feature

in our modeling. This happens because ethnicity is computed as a binary variable, distinguishing between white and non-white individuals. This distinction, however, is not really useful for the Guatemalan case, as only a very small portion of the population identifies as white. We can imagine such shortcomings are present in more than one case.

For future work, we can look at each country with more detail, utilizing data that makes the most sense given the countries' socioeconomic background, and not with a generic approach used in our work. Another possibility is studying the change in the opportunity tree within a country over the years, to study the evolution of inequality within the country, as was done in Brunori and Neidhöfer (2021) for the case of Germany. There is also space for improving the estimation of IOP scores, as the Gini index does not fully captures the relation between inequality of opportunity and inequality of returns to effort.

REFERENCES

AABERGE, Rolf; MOGSTAD, Magne; PERAGINE, Vito. Measuring long-term inequality of opportunity. **Journal of Public Economics**, Elsevier, v. 95, n. 3-4, p. 193–204, 2011.

ALESINA, Alberto; STANTCHEVA, Stefanie; TESO, Edoardo. Intergenerational mobility and preferences for redistribution. **American Economic Review**, v. 108, n. 2, p. 521–54, 2018.

ARROW, Kenneth; BOWLES, Samuel; DURLAUF, Steven N. **Meritocracy and economic inequality**. Princeton University Press, 2018.

BANK, World. **World development report 2006: Equity and development**. The World Bank, 2005.

BARBOSA, Rogério J; SOUZA, Pedro HG Ferreira de; SOARES, Sergei. **Distribuição de Renda nos Anos 2010: uma década perdida para desigualdade e pobreza**. 2020.

BAUM, Sandy. Higher education earnings premium: Value, variation, and trends. **Urban Institute**, ERIC, 2014.

BRUNORI, Paolo; HUFE, Paul; MAHLER, Daniel Gerszon. The roots of inequality: estimating inequality of opportunity from regression trees. **World Bank Policy Research Working Paper**, n. 8349, 2018.

BRUNORI, Paolo; NEIDHÖFER, Guido. The evolution of inequality of opportunity in Germany: A machine learning approach. **Review of Income and Wealth**, Wiley Online Library, v. 67, n. 4, p. 900–927, 2021.

BURKART, Nadia; HUBER, Marco F. A survey on the explainability of supervised machine learning. **Journal of Artificial Intelligence Research**, v. 70, p. 245–317, 2021.

CABRERA, Maynor; LUSTIG, Nora; MORÁN, Hilcıas E. Fiscal policy, inequality, and the ethnic divide in Guatemala. **World Development**, Elsevier, v. 76, p. 263–279, 2015.

CARVALHO, Márcia; WALTENBERG, Fábio D. Desigualdade de oportunidades no acesso ao ensino superior no Brasil: uma comparação entre 2003 e 2013. **Economia Aplicada**, SciELO Brasil, v. 19, n. 2, p. 369–396, 2015.

CHECCHI, Daniele; PERAGINE, Vito. Inequality of opportunity in Italy. **The Journal of Economic Inequality**, Springer, v. 8, n. 4, p. 429–450, 2010.

CRUCES, Guillermo; DOMENCH, C Garcıa; GASPARINI, Leonardo. Inequality in education: evidence for Latin America. **Falling inequality in Latin America. Policy changes and lessons**, p. 318–339, 2014.

DILL, Helena Cristina; GONÇALVES, Flávio de O. Igualdade de oportunidade no Brasil entre 1999 e 2009: estimação e decomposição através do valor de Shapley. **Pesquisa e Planejamento Econômico**, v. 42, n. 2, p. 185–210, 2012.

DONNI, Paolo Li; PERAGINE, Vito; PIGNATARO, Giuseppe. Ex-ante and ex-post measurement of equality of opportunity in health: a normative decomposition. **Health economics**, Wiley Online Library, v. 23, n. 2, p. 182–198, 2014.

DUBET, François. Égalité des places, égalité des chances. **Études**, SER, v. 414, n. 1, p. 31–41, 2011.

DWORKIN, Ronald. Part 1: Equality of Welfare. **Philosophy and Public Affairs**, v. 10, n. 3, p. 185–246, 1981.

FERREIRA, Francisco HG; GIGNOUX, Jérémie. The measurement of inequality of opportunity: Theory and an application to Latin America. **Review of income and wealth**, Wiley Online Library, v. 57, n. 4, p. 622–657, 2011.

FLEURBAEY, Marc; PERAGINE, Vito. Ex ante versus ex post equality of opportunity. **Economica**, Wiley Online Library, v. 80, n. 317, p. 118–130, 2013.

GARETH, James et al. **An introduction to statistical learning: with applications in R**. Spinger, 2013.

GRAJALES, Roberto Vélez; MONROY-GÓMEZ-FRANCO, Luis A; YALONETZKY, Gastón. **Inequality of opportunity in Mexico**. 2018.

HASTIE, Trevor et al. **The elements of statistical learning: data mining, inference, and prediction**. Springer, 2009. v. 2.

HOFFMAN, Kelly; CENTENO, Miguel Angel. The lopsided continent: inequality in Latin America. **Annual Review of Sociology**, Annual Reviews 4139 El Camino Way, PO Box 10139, Palo Alto, CA 94303-0139, USA, v. 29, n. 1, p. 363–390, 2003.

HOTHORN, Torsten; HORNIK, Kurt; ZEILEIS, Achim. Unbiased recursive partitioning: A conditional inference framework. **Journal of Computational and Graphical statistics**, Taylor & Francis, v. 15, n. 3, p. 651–674, 2006.

IBGE. **Normas de apresentação tabular**. 3. ed. Rio de Janeiro: Centro de Documentação e Disseminação de Informações. Fundação Intituto Brasileiro de Geografia e Estatística, 1993. Visited on: 21 Aug. 2013.

JIMÉNEZ, Maribel; JIMÉNEZ, Mónica. Intergenerational educational mobility in Latin America. An analysis from the equal opportunity approach. **Cuadernos de Economıa**, Cuadernos de Economıa, Facultad de Ciencias Económicas, Universidad Nacional . . ., v. 38, n. 76, p. 289–329, 2019.

KANBUR, Ravi; SNELL, Andy. Inequality indices as tests of fairness. **The Economic Journal**, Oxford University Press, v. 129, n. 621, p. 2216–2239, 2019.

KUZNETS, Simon. Economic growth and income inequality. In: THE gap between rich and poor. Routledge, 2019. P. 25–37.

LATINOBARÓMETRO. **Latin American Public Opinion**. Available from: https://www.latinobarometro.org/latContents.jsp. Accessed: 01/05/2022.

MAHLMEISTER, Rodrigo et al. Revisitando a mobilidade intergeracional de educação no Brasil. **Revista Brasileira de Economia**, SciELO Brasil, v. 73, p. 159–180, 2019.

MARRERO, Gustavo A; RODRIGUEZ, Juan G. Inequality of opportunity and growth. **Journal of development Economics**, Elsevier, v. 104, p. 107–122, 2013.

MILANOVIC, Branco. A cost of transition: 50 million new poor and growing inequality. **Transition**, v. 5, n. 8, p. 1–4, 1994.

NICKELL, Stephen. Poverty and worklessness in Britain. **The Economic Journal**, Oxford University Press Oxford, UK, v. 114, n. 494, p. c1–c25, 2004.

RAWLS, A. **Theories of social justice**. Harvard University Press Boston, 1971.

RIBEIRO, Carlos Antônio Costa. Desigualdade de oportunidades e resultados educacionais no Brasil. **Dados**, SciELO Brasil, v. 54, p. 41–87, 2011.

ROEMER, John E. A pragmatic theory of responsibility for the egalitarian planner. **Philosophy & Public Affairs**, JSTOR, p. 146–166, 1993.

ROEMER, John E. Equality of opportunity. In: EQUALITY of Opportunity. Harvard University Press, 1998.

ROEMER, John E. Equality of opportunity: A progress report. **Social Choice and Welfare**, JSTOR, p. 455–471, 2002.

SREENIVASAN, Gopal. Health care and equality of opportunity. **Hastings Center Report**, Wiley Online Library, v. 37, n. 2, p. 21–31, 2007.

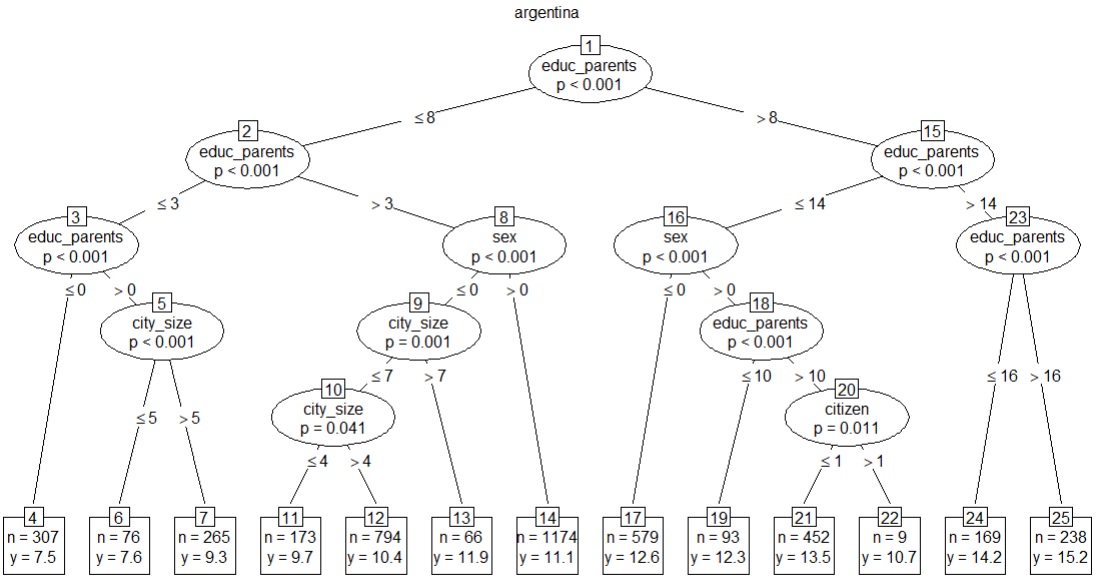
STROBL, Carolin et al. Conditional variable importance for random forests. **BMC bioinformatics**, Springer, v. 9, n. 1, p. 1–11, 2008.

VAN GIGCH, John P.; PIPINO, Leo L. In search for a paradigm for the discipline of information systems. **Future Computing Systems**, v. 1, n. 1, p. 71–97, 1986.

WALL, David L. Spatial inequalities in sandinista Nicaragua. **Geographical Review**, JSTOR, p. 1–13, 1993.

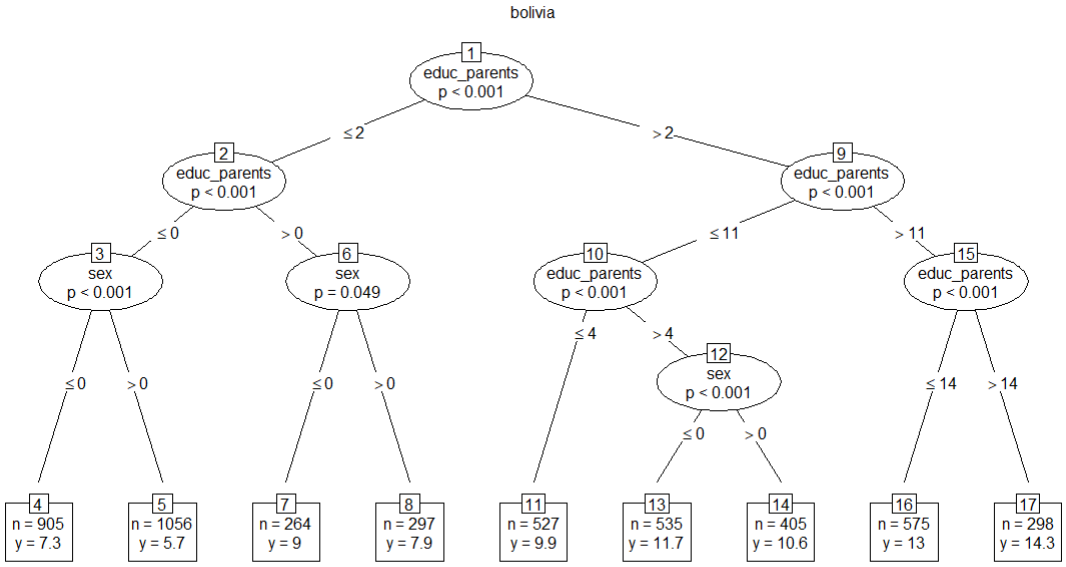
ANNEX A – OPPORTUNITY TREES

Figure 15 – Educational Opportunity Tree for Argentina.



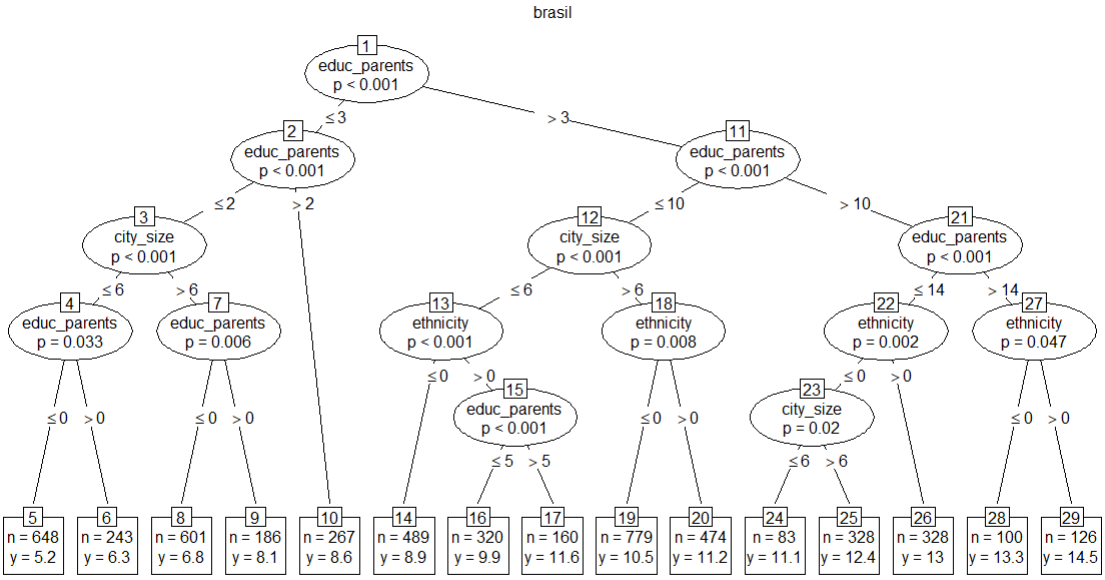
Source: Elaborated by the author using Latinobarómetro data

Figure 16 – Educational Opportunity Tree for for Bolivia.



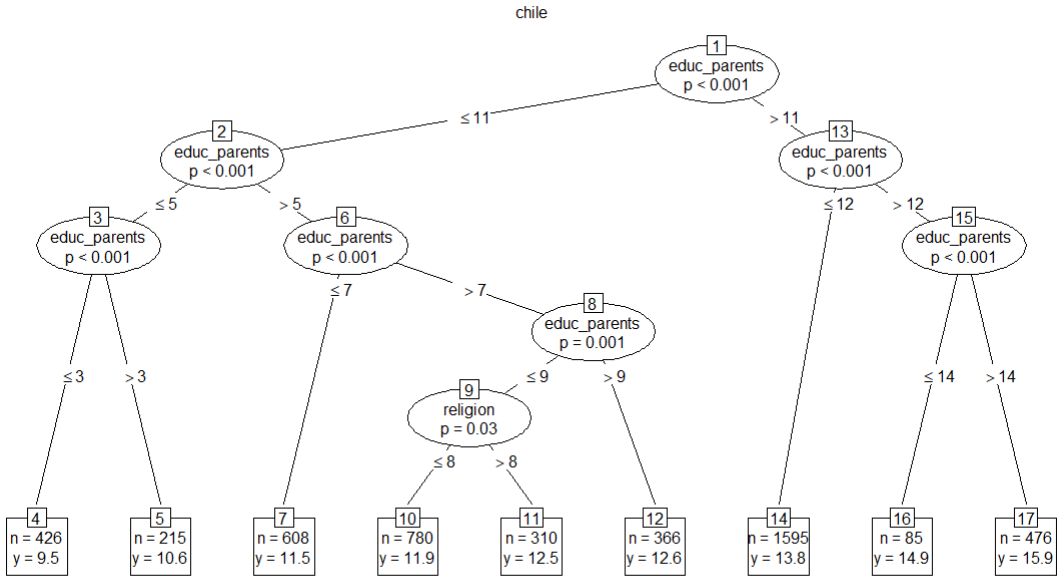
Source: Elaborated by the author using Latinobarómetro data

Figure 17 – Educational Opportunity Tree for Brazil.



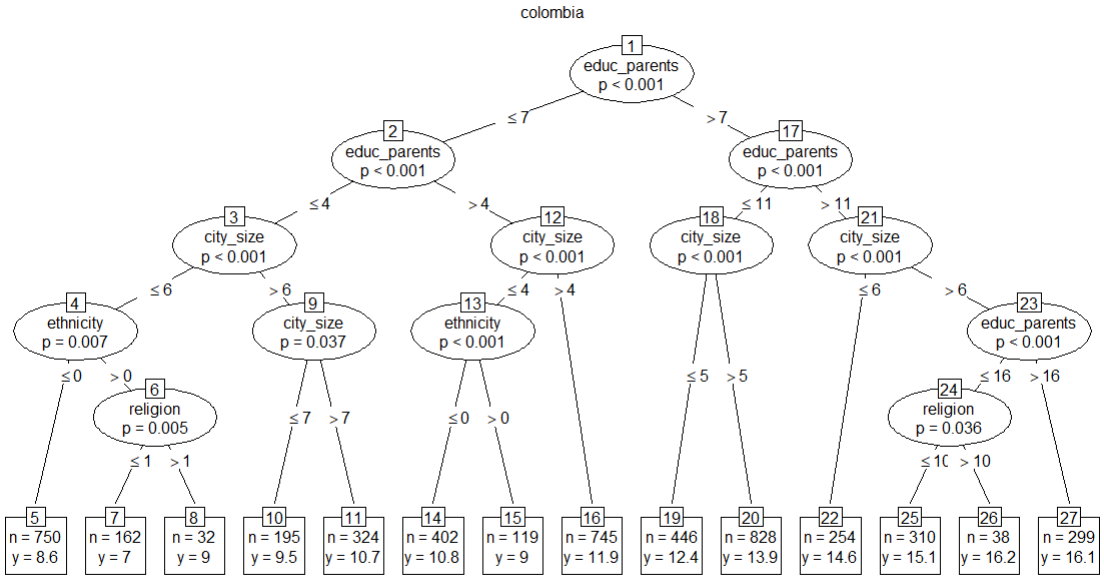
Source: Elaborated by the author using Latinobarómetro data

Figure 18 – Educational Opportunity Tree for Chile.



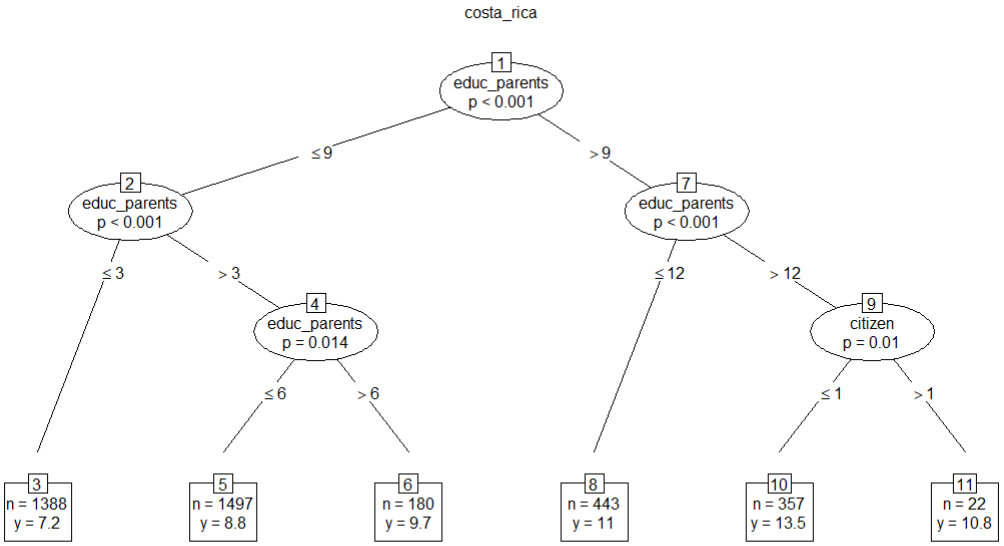
Source: Elaborated by the author using Latinobarómetro data

Figure 19 – Educational Opportunity Tree for Colombia.



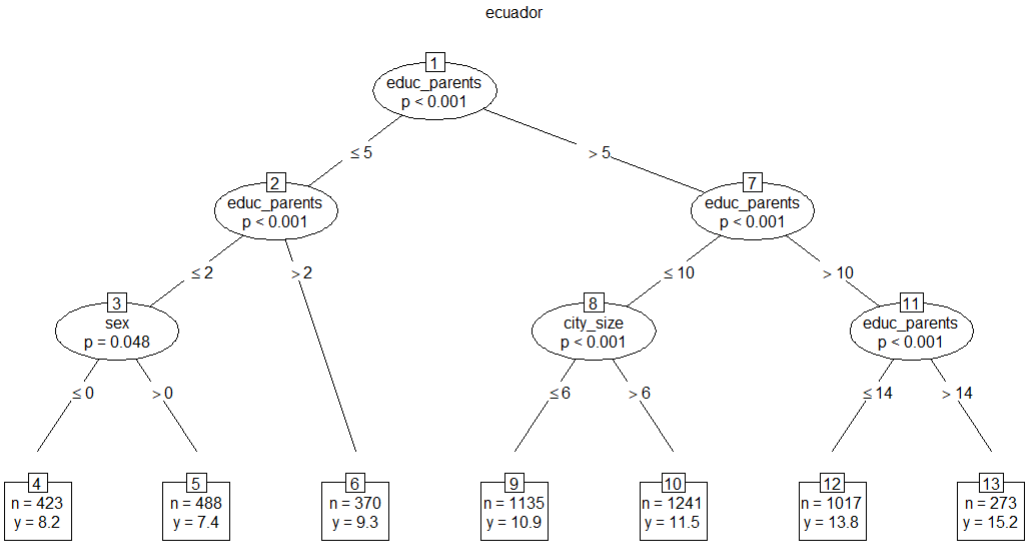
Source: Elaborated by the author using Latinobarómetro data

Figure 20 – Educational Opportunity Tree for Costa Rica.



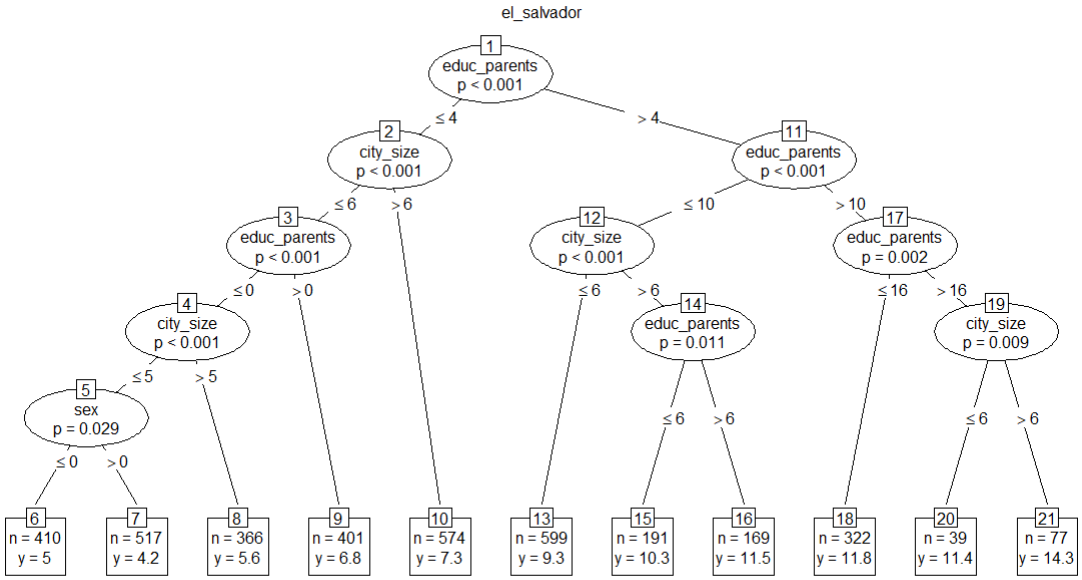
Source: Elaborated by the author using Latinobarómetro data

Figure 21 – Educational Opportunity Tree for Ecuador.



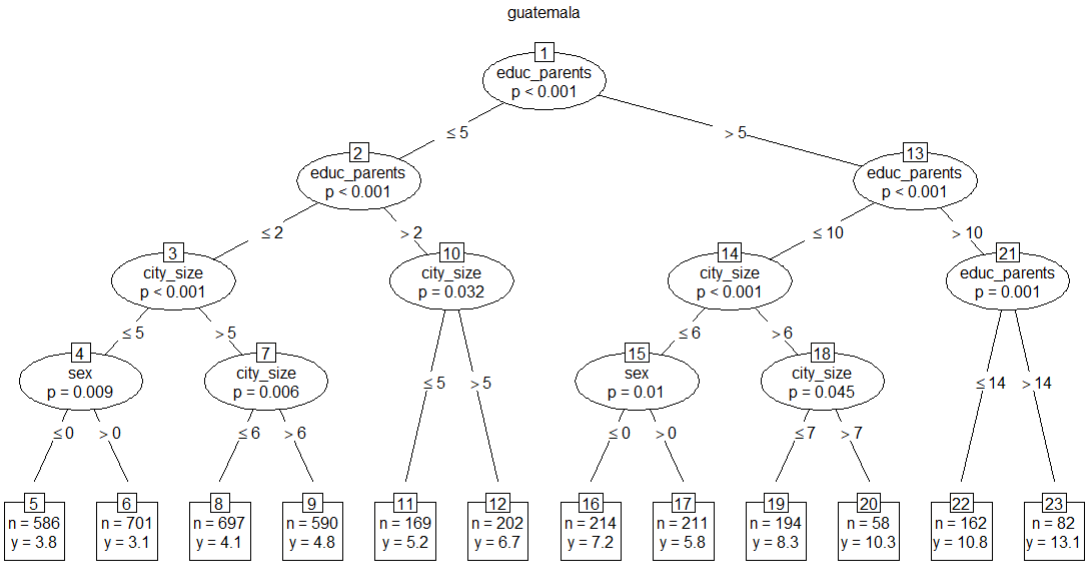
Source: Elaborated by the author using Latinobarómetro data

Figure 22 – Educational Opportunity Tree for El Salvador.



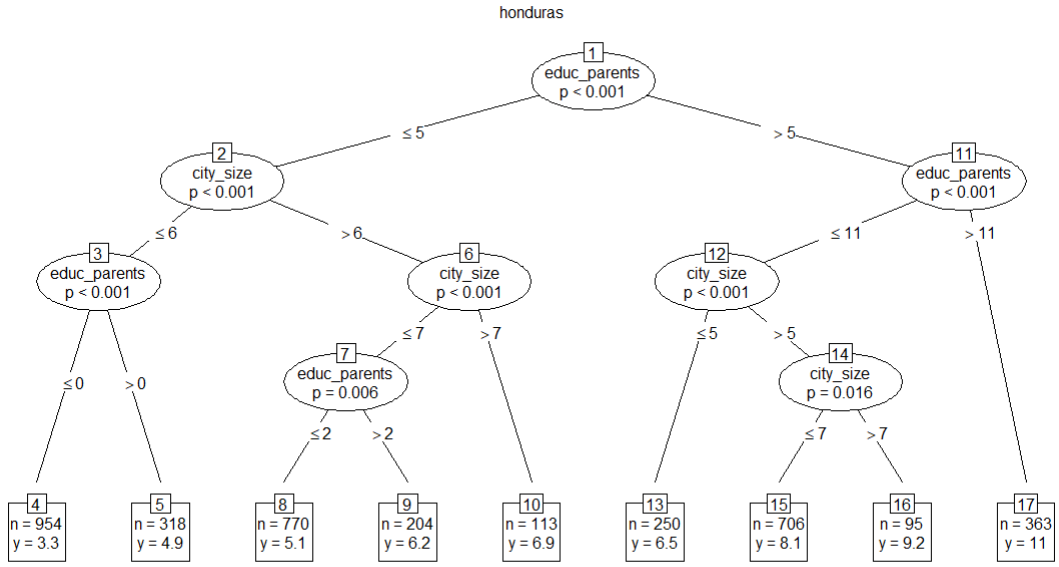
Source: Elaborated by the author using Latinobarómetro data

Figure 23 – Educational Opportunity Tree for Guatemala.



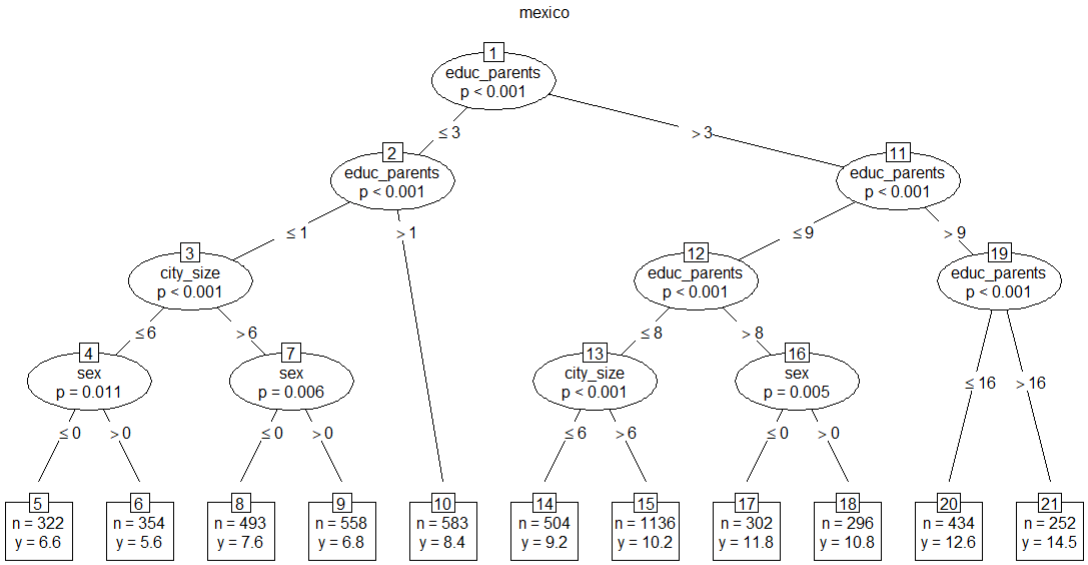
Source: Elaborated by the author using Latinobarómetro data

Figure 24 – Educational Opportunity Tree for Honduras.



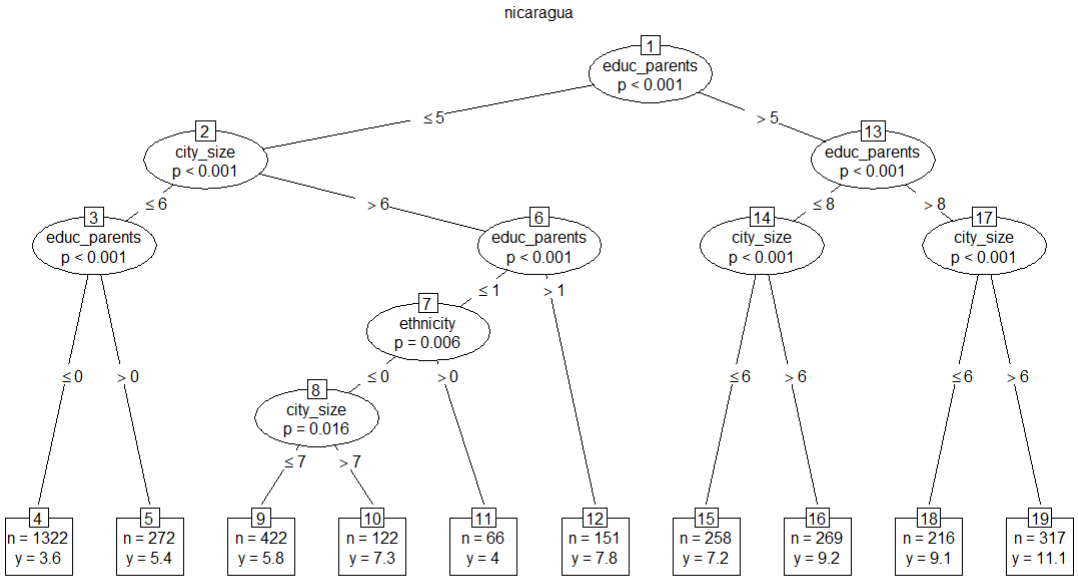
Source: Elaborated by the author using Latinobarómetro data

Figure 25 – Educational Opportunity Tree for Mexico.



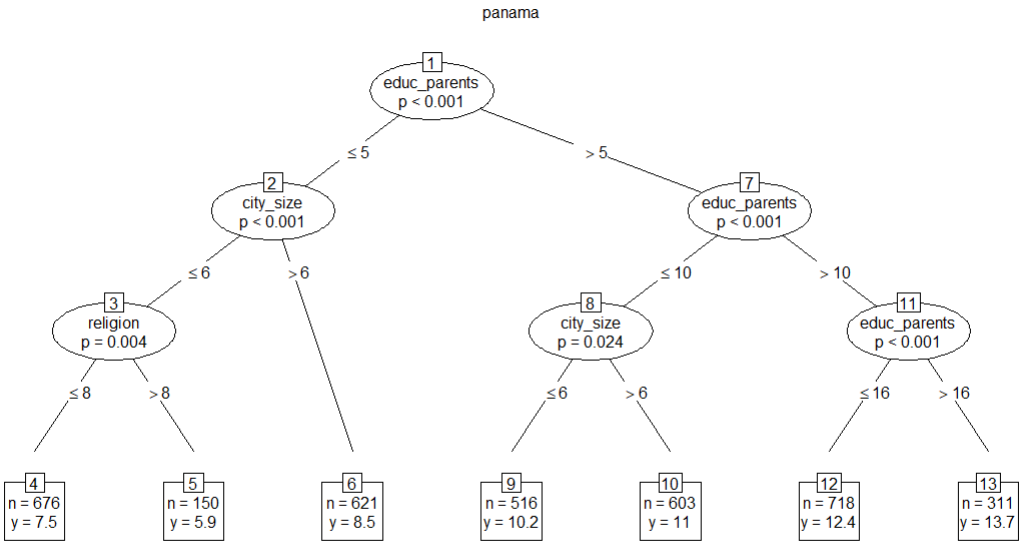
Source: Elaborated by the author using Latinobarómetro data

Figure 26 – Educational Opportunity Tree for Nicaragua.



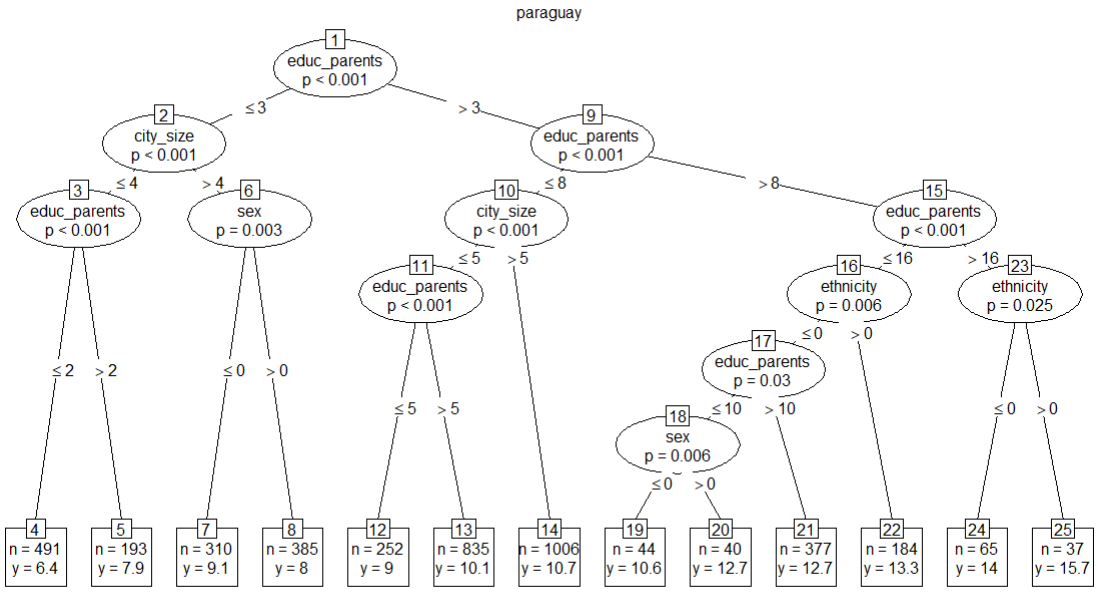
Source: Elaborated by the author using Latinobarómetro data

Figure 27 – Educational Opportunity Tree for Panama.



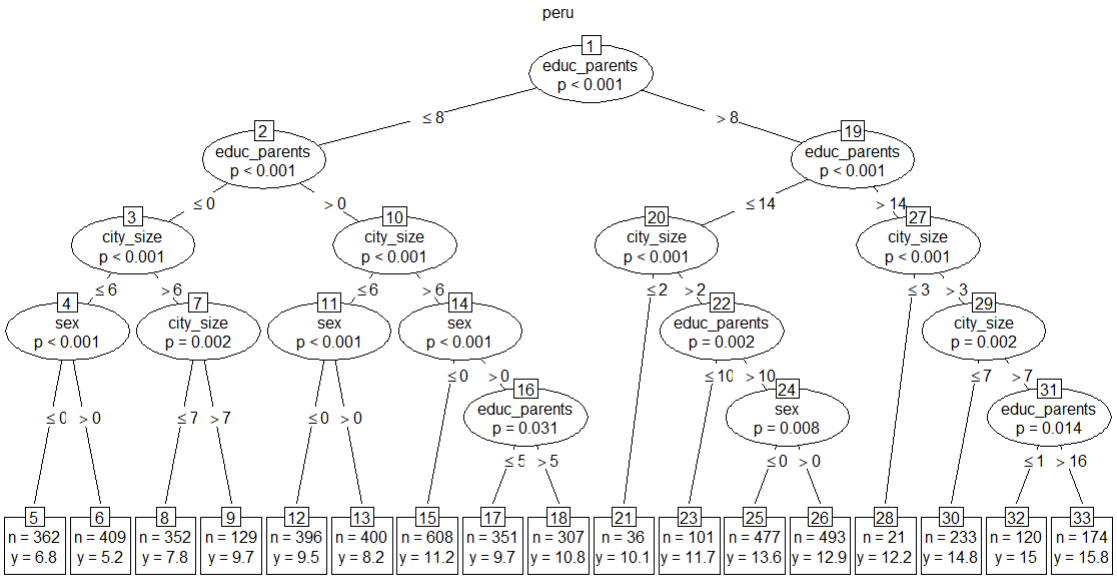
Source: Elaborated by the author using Latinobarómetro data

Figure 28 – Educational Opportunity Tree for Paraguay.



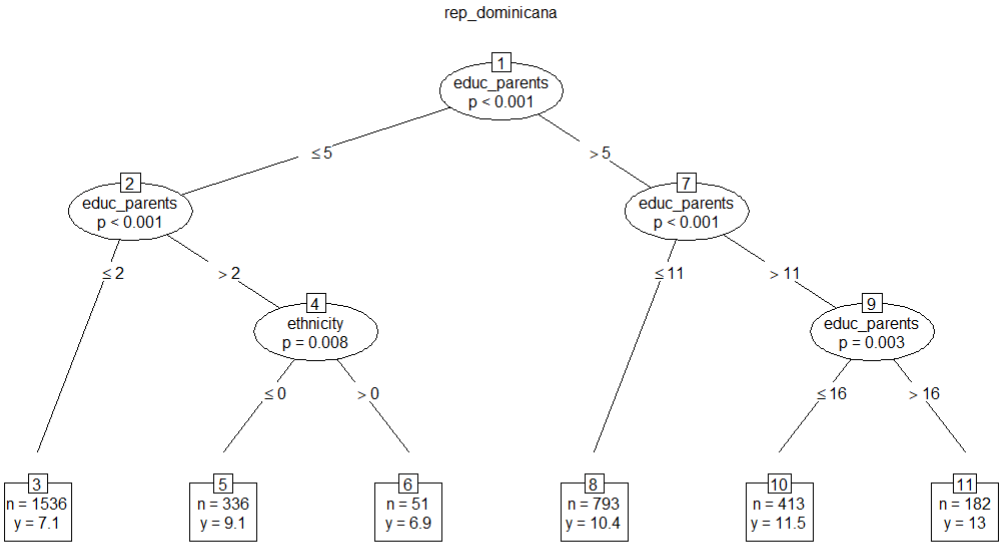
Source: Elaborated by the author using Latinobarómetro data

Figure 29 – Educational Opportunity Tree for Peru.



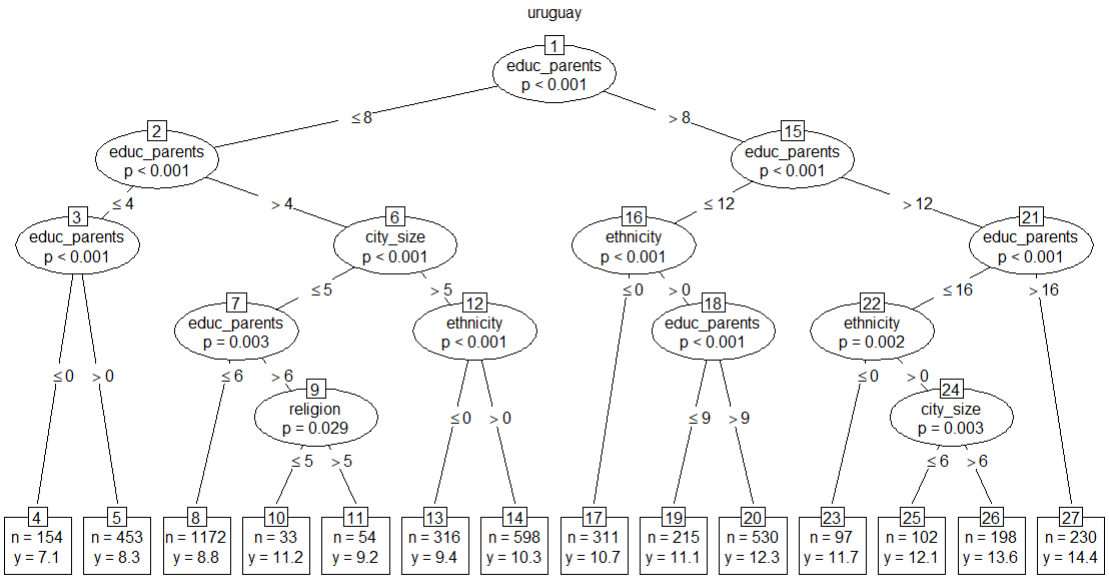
Source: Elaborated by the author using Latinobarómetro data

Figure 30 – Educational Opportunity Tree for Dominican Rep..



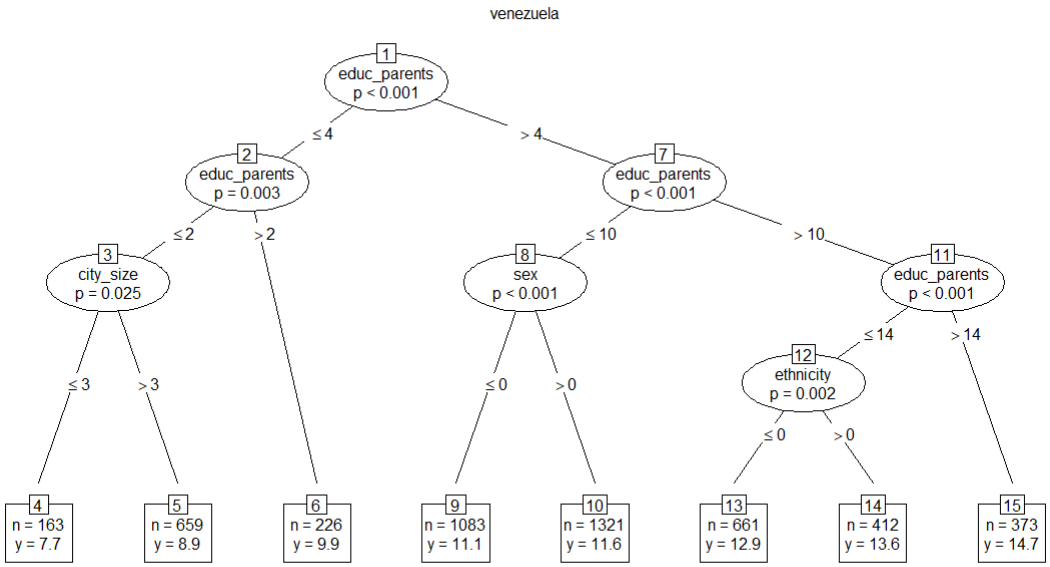
Source: Elaborated by the author using Latinobarómetro data

Figure 31 – Educational Opportunity Tree for Uruguay.



Source: Elaborated by the author using Latinobarómetro data

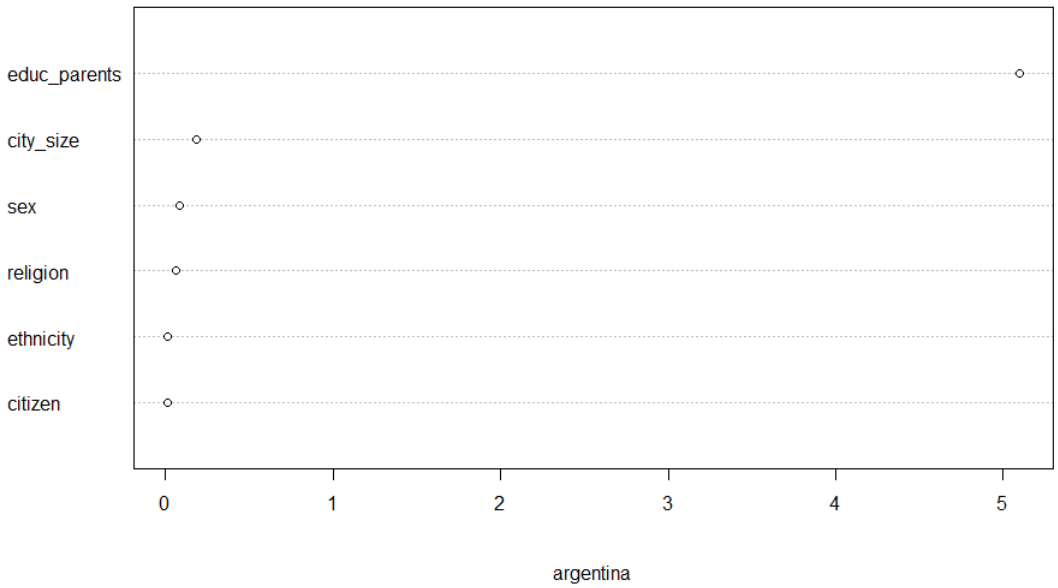
Figure 32 – Educational Opportunity Tree for Venezuela.



Source: Elaborated by the author using Latinobarómetro data

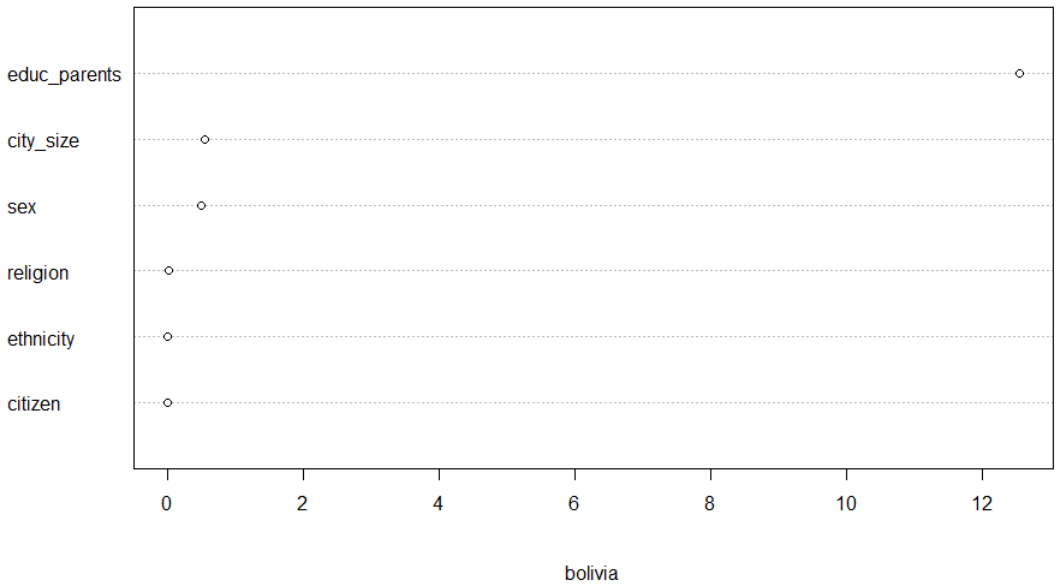
ANNEX B – FEATURE IMPORTANCE

Figure 33 – Conditional Feature Importance for Argentina.



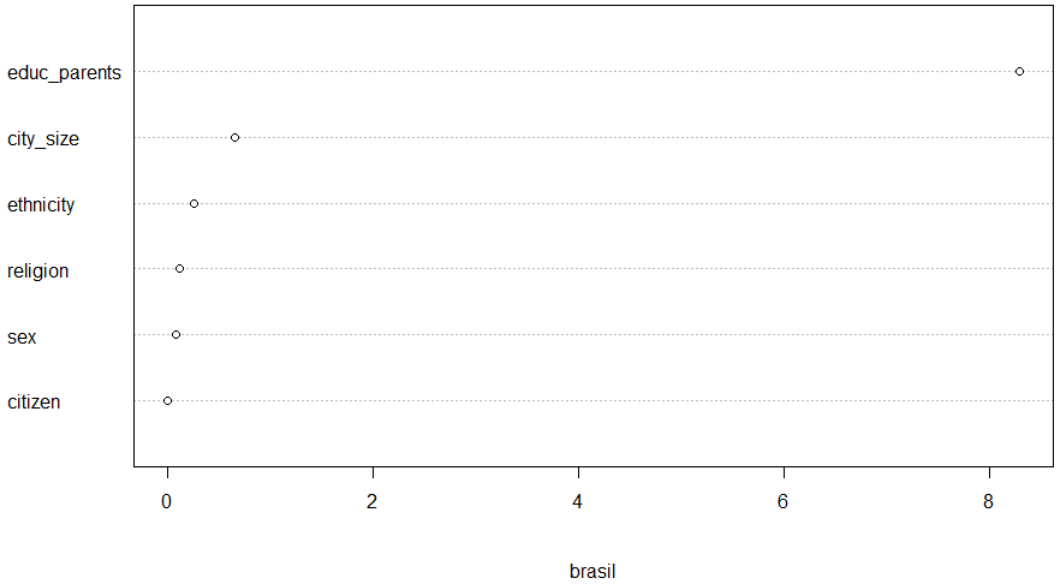
Source: Elaborated by the author using Latinobarómetro data

Figure 34 – Conditional Feature Importance for Bolivia.



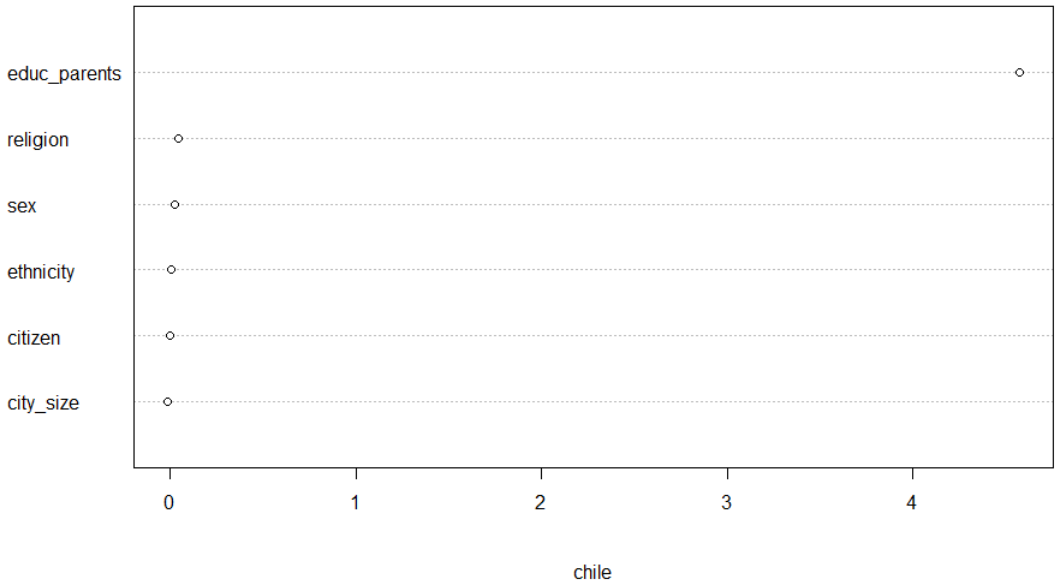
Source: Elaborated by the author using Latinobarómetro data

Figure 35 – Conditional Feature Importance for Brazil.



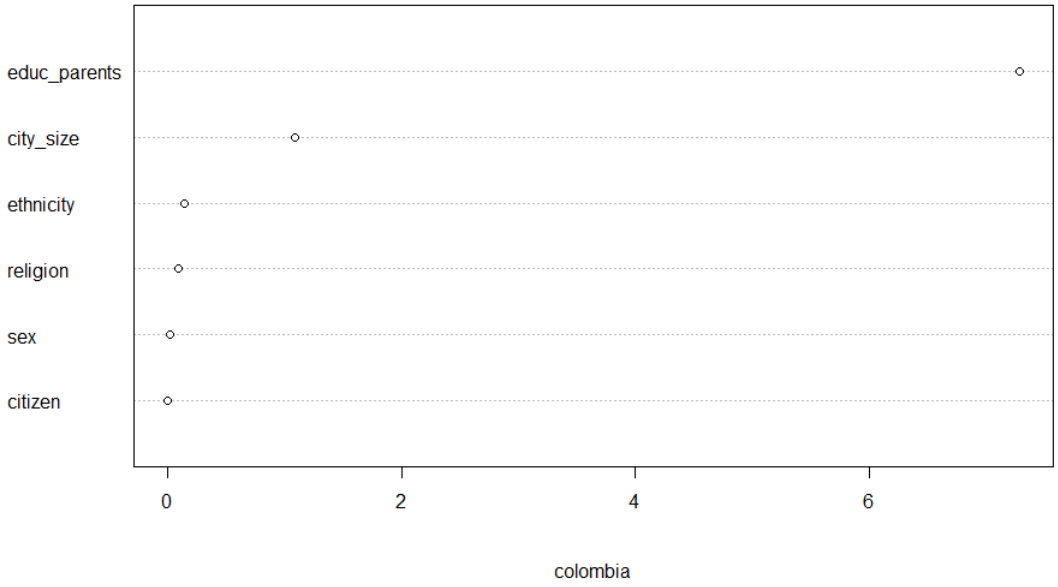
Source: Elaborated by the author using Latinobarómetro data

Figure 36 – Conditional Feature Importance for Chile.



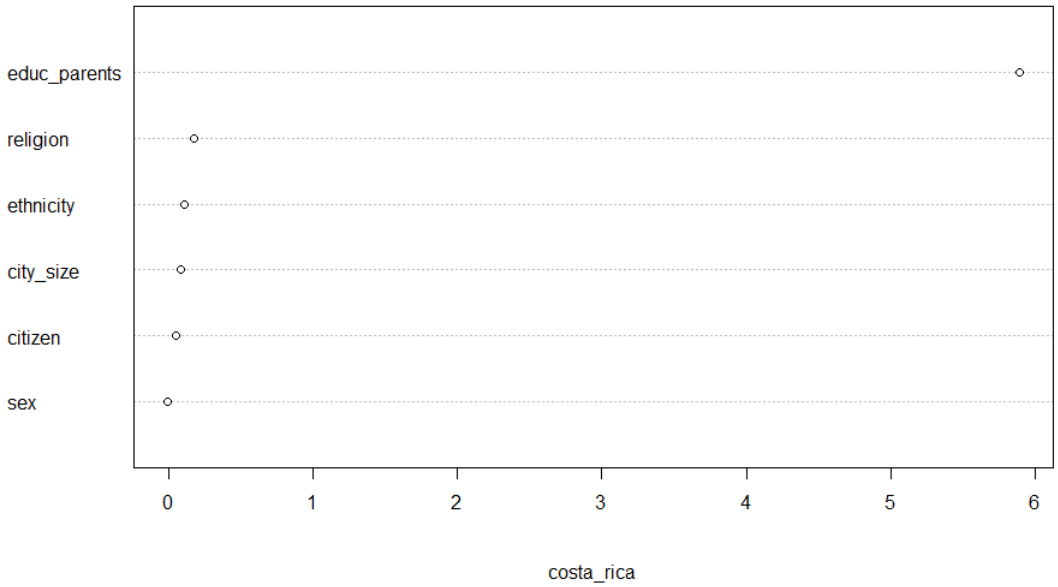
Source: Elaborated by the author using Latinobarómetro data

Figure 37 – Conditional Feature Importance for Colombia.



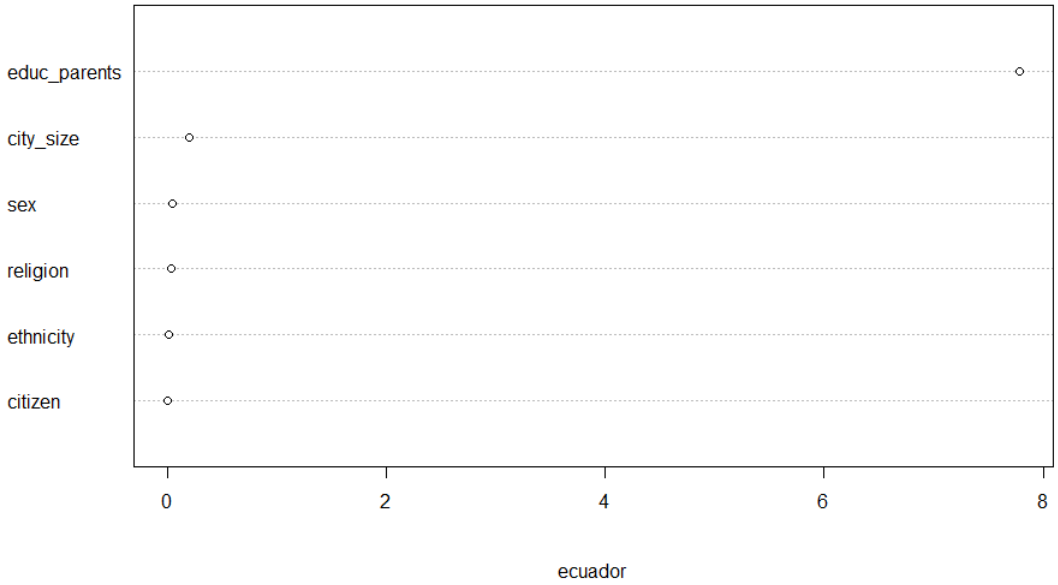
Source: Elaborated by the author using Latinobarómetro data

Figure 38 – Conditional Feature Importance for Costa Rica.



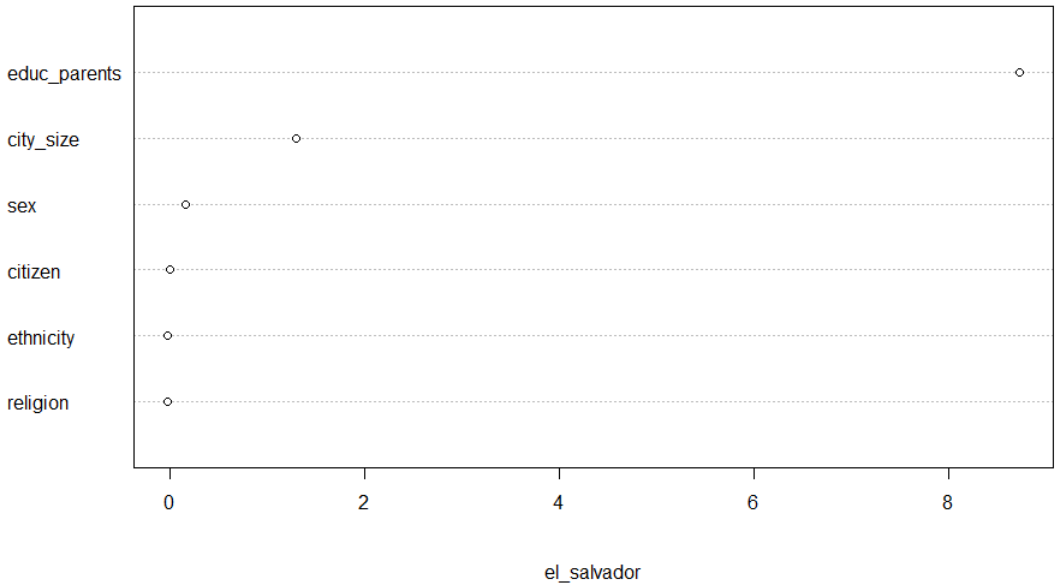
Source: Elaborated by the author using Latinobarómetro data

Figure 39 – Conditional Feature Importance for Ecuador.



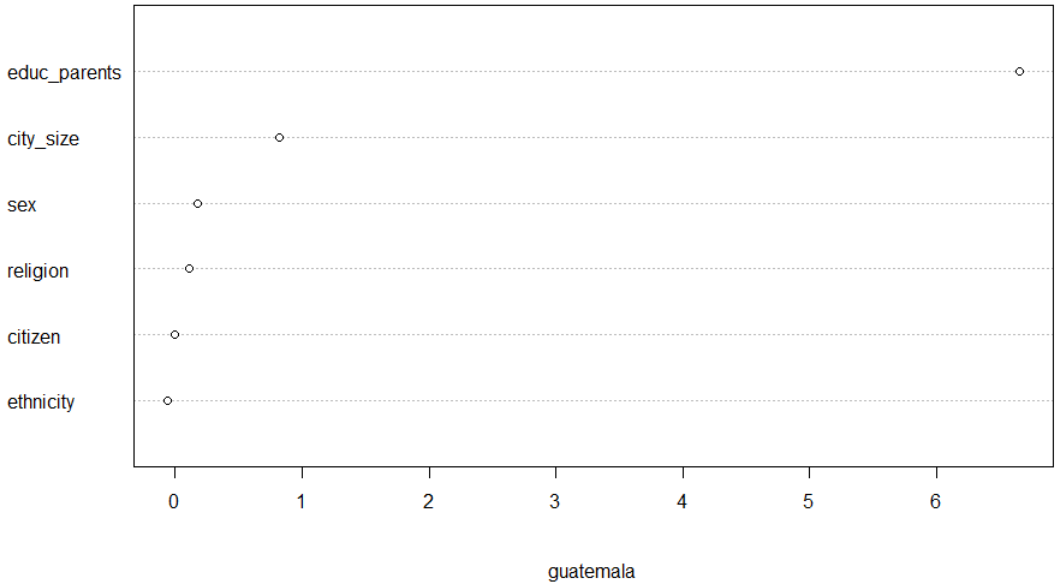
Source: Elaborated by the author using Latinobarómetro data

Figure 40 – Conditional Feature Importance for El Salvador.



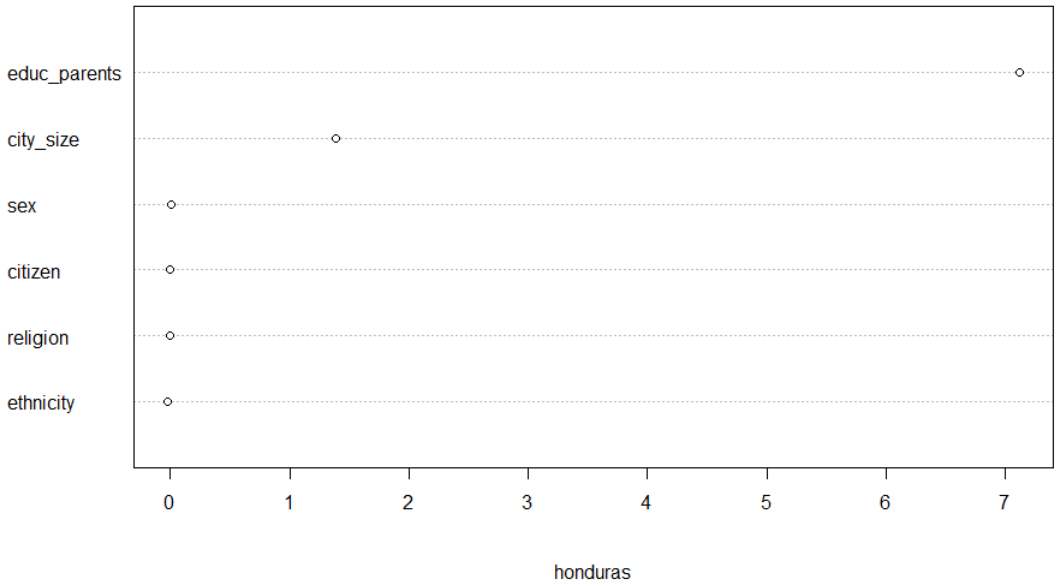
Source: Elaborated by the author using Latinobarómetro data

Figure 41 – Conditional Feature Importance for Guatemala.



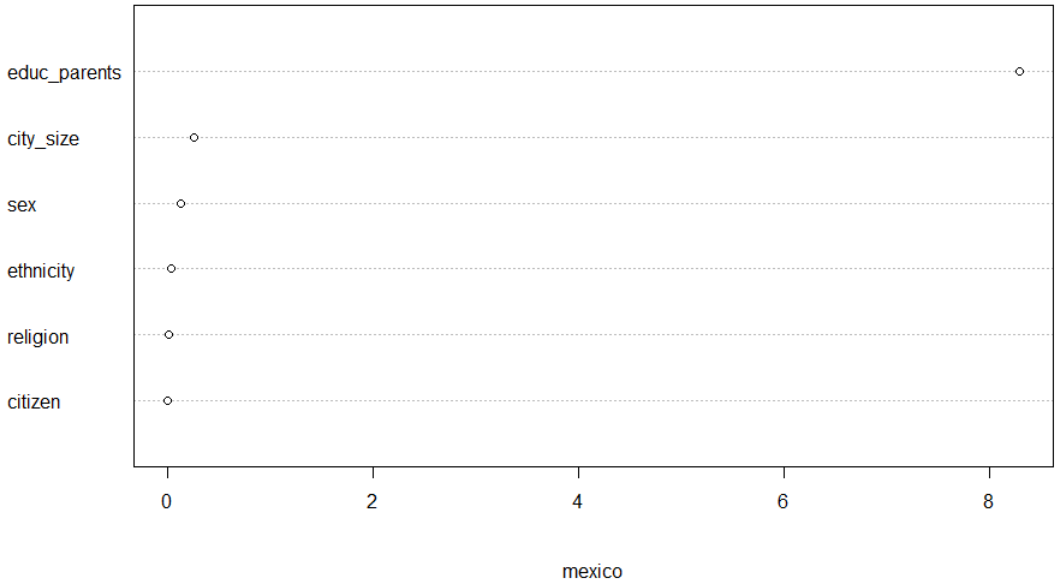
Source: Elaborated by the author using Latinobarómetro data

Figure 42 – Conditional Feature Importance for Honduras.



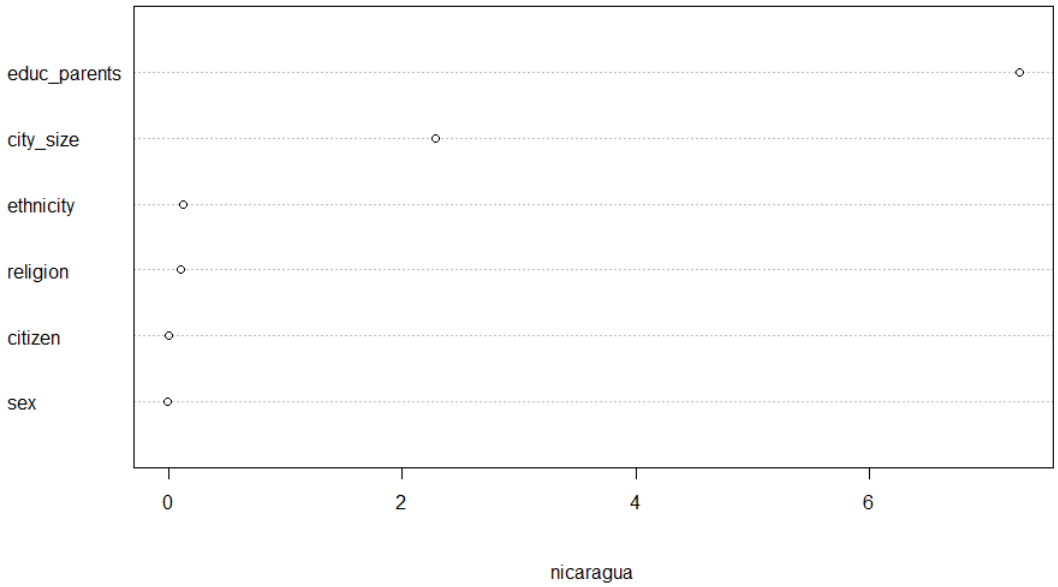
Source: Elaborated by the author using Latinobarómetro data

Figure 43 – Conditional Feature Importance for Mexico.



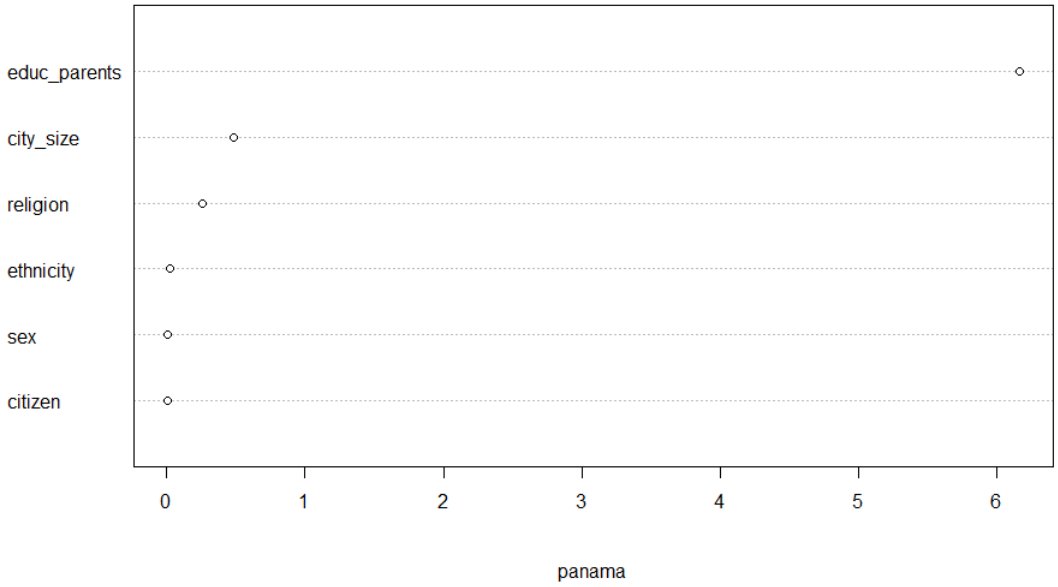
Source: Elaborated by the author using Latinobarómetro data

Figure 44 – Conditional Feature Importance for Nicaragua.



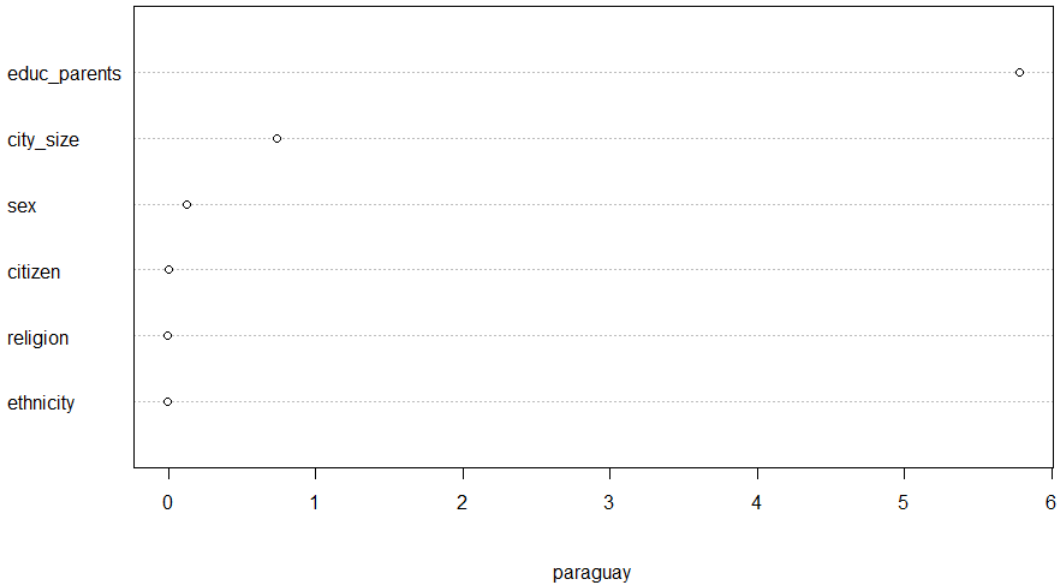
Source: Elaborated by the author using Latinobarómetro data

Figure 45 – Conditional Feature Importance for Panama.



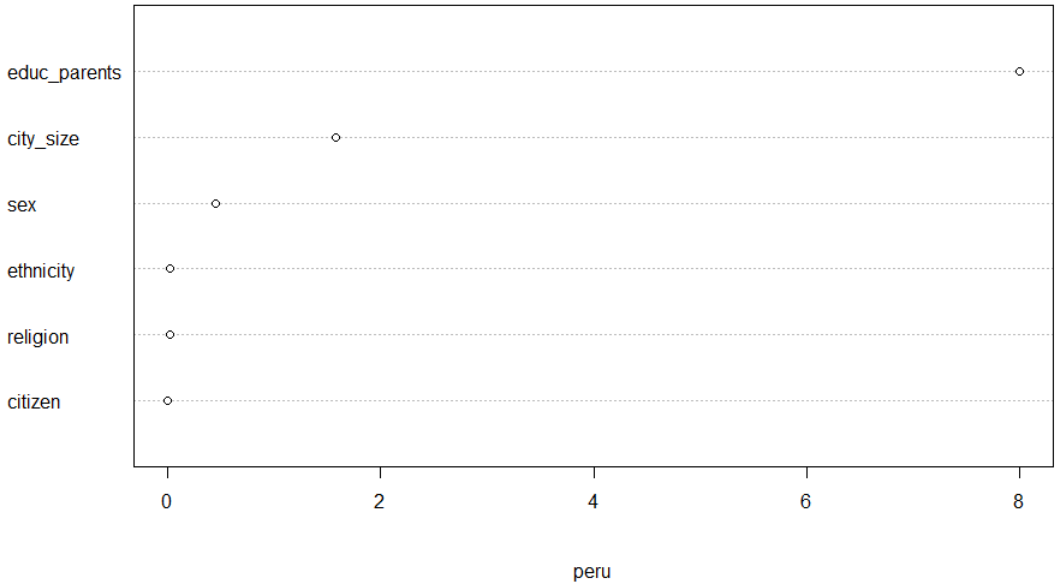
Source: Elaborated by the author using Latinobarómetro data

Figure 46 – Conditional Feature Importance for Paraguay.



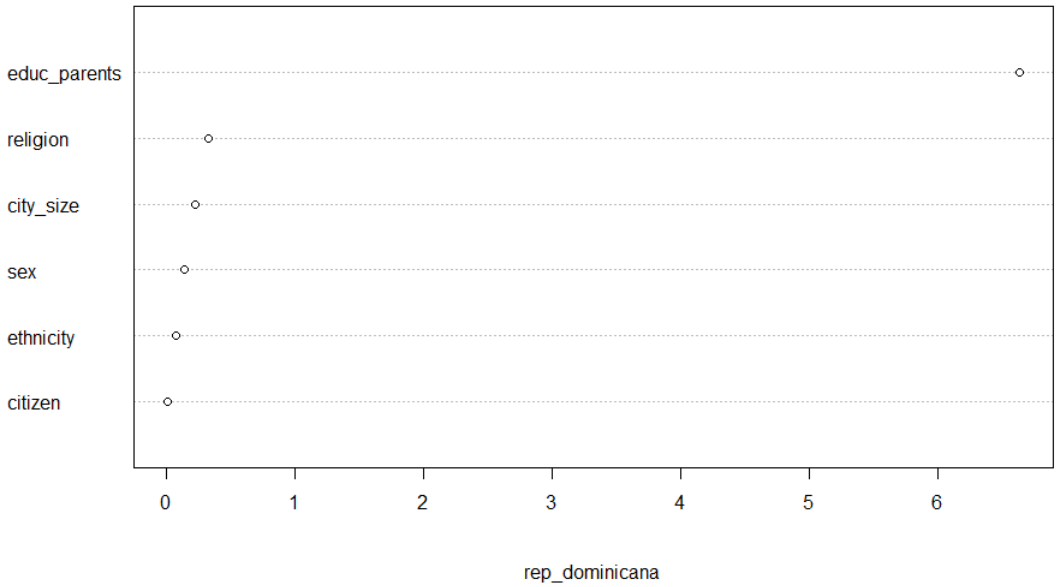
Source: Elaborated by the author using Latinobarómetro data

Figure 47 – Conditional Feature Importance for Peru.



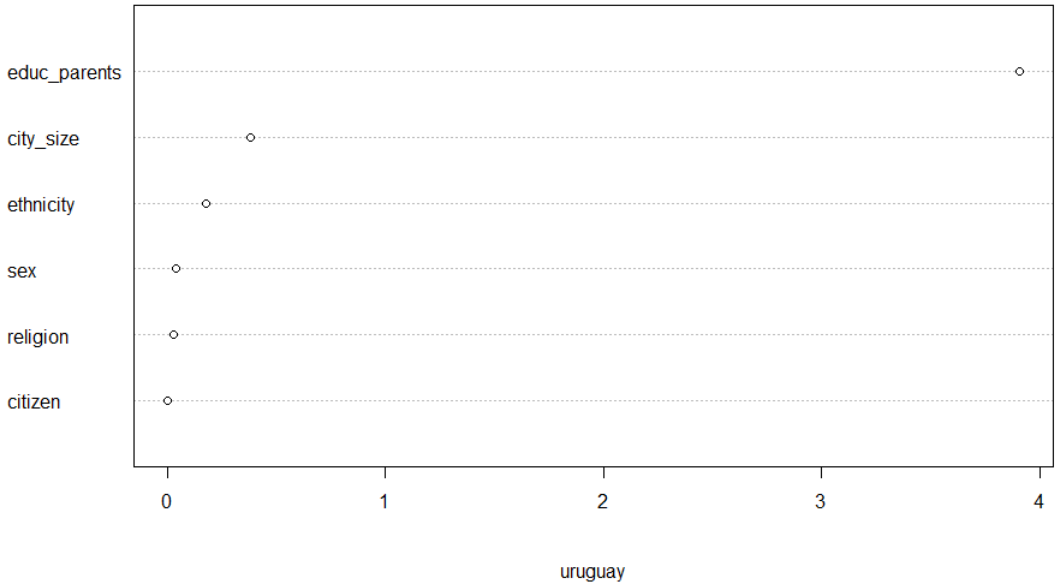
Source: Elaborated by the author using Latinobarómetro data

Figure 48 – Conditional Feature Importance for Dominican Rep.



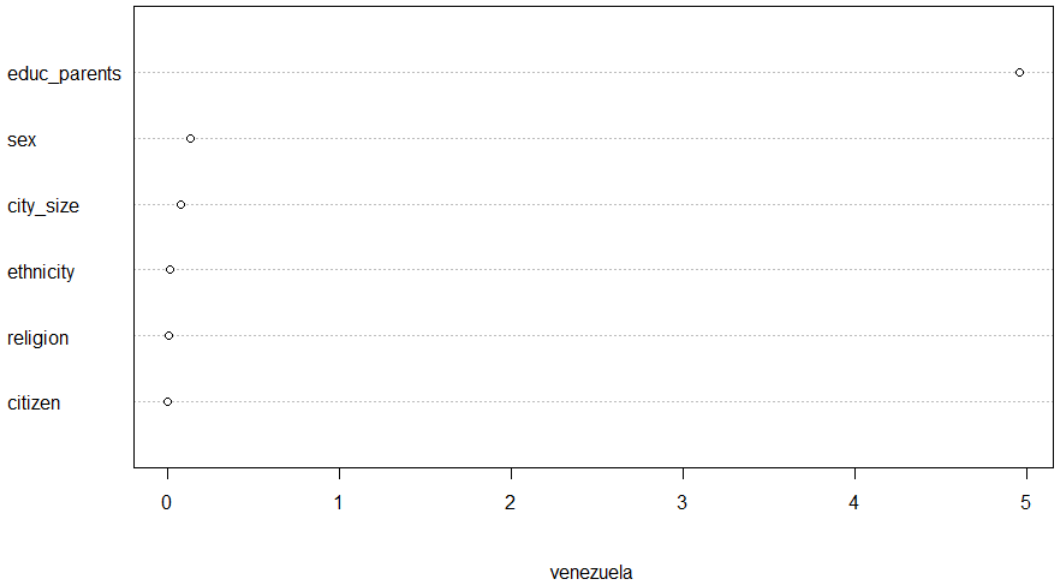
Source: Elaborated by the author using Latinobarómetro data

Figure 49 – Conditional Feature Importance for Uruguay.



Source: Elaborated by the author using Latinobarómetro data

Figure 50 – Conditional Feature Importance for Venezuela.



Source: Elaborated by the author using Latinobarómetro data

ANNEX C – DATA CLEANING

Code 1 – Python Code

```
import numpy as np
import pandas as pd
#grabbing data
#Since the dictionary changes between years, we have to do this manually

cols2011 = ['numinves', 'idenpa', 'reedad', 'ciudad', 'tamciud',
            'S16', 'S18', 'S25', 'S27', 'S21', 'S22']
dic2011 = {'S16': 'sex', 'S18': 'religion','S25': 'citizen', 'S27':'ethnicity',
           'S21': 'anos_educ', 'S22':'anos_educ_pai', 'numinves': 'year', 'idenpa': 'country', 'reedad': 'age',
           'ciudad': 'city', 'tamciud': 'city_size'}

cols2013 = ['numinves', 'idenpa', 'reedad', 'ciudad', 'tamciud',
           'S10', 'S12', 'S14', 'S17', 'S18', 'S21']
dic2013 = {'S10' : 'sex', 'S12': 'citizen', 'S14':'religion', 'S17': 'anos_educ', 'S18': 'anos_educ_pai',
          'S21': 'ethnicity', 'numinves' : 'year', 'idenpa': 'country', 'reedad': 'age',
           'ciudad': 'city', 'tamciud': 'city_size'}

cols2015 = ['numinves', 'idenpa', 'reedad', 'ciudad', 'tamciud',
           'S12', 'S14', 'S16', 'S19', 'S20', 'S23']
dic2015 = {'S12':'sex', 'S14': 'citizen', 'S16': 'religion', 'S19':'anos_educ',
           'S20': 'anos_educ_pai', 'S23':'ethnicity', 'SEXO': 'sex', 'numinves' : 'year', 'idenpa': 'country', 'reedad': 'age',
           'ciudad': 'city', 'tamciud': 'city_size'}

cols2016 = ['numinves', 'idenpa', 'reedad', 'ciudad', 'tamciud',
           'S7', 'S8', 'S9', 'S13', 'S14', 'sexo']
dic2016 = {'S7':'citizen', 'S8':'religion', 'S9':'ethnicity', 'S13': 'anos_educ',
           'S14':'anos_educ_pai', 'sexo': 'sex', 'numinves' : 'year', 'idenpa': 'country', 'reedad': 'age',
```

```
28            'ciudad': 'city', 'tamciud': 'city_size'}

30 cols2017 = ['numinves', 'idenpa', 'reedad', 'ciudad', 'tamciud',
31            'sexo', 'S8', 'S9', 'S10', 'S14', 'S15']
32 dic2017 = {'S8':'citizen', 'S9':'religion', 'S10':'ethnicity', '
      S14': 'anos_educ',
33            'S15':'anos_educ_pai', 'sexo': 'sex', 'numinves' : '
                 year', 'idenpa': 'country', 'reedad': 'age',
34            'ciudad': 'city', 'tamciud': 'city_size'}

36 cols2018 = ['NUMINVES', 'IDENPA', 'REEDAD', 'CIUDAD', 'TAMCIUD',
      'SEXO',
37            'S5', 'S6', 'S10', 'S11', 'S16']
38 dic2018 = {'S16':'citizen', 'S5':'religion', 'S6':'ethnicity', '
      S10': 'anos_educ',
39            'S11':'anos_educ_pai', 'SEXO': 'sex', 'NUMINVES' : '
                 year', 'IDENPA': 'country', 'REEDAD': 'age',
40            'CIUDAD': 'city', 'TAMCIUD': 'city_size'}

42 cols2020 = ['numinves', 'idenpa', 'ciudad', 'tamciud', 'sexo',
43           's10','s12', 's16', 's17', 's21', 'reedad']
44 dic2020 = {'s21':'citizen', 's10':'religion', 's12':'ethnicity',
      's16': 'anos_educ',
45            's17':'anos_educ_pai', 'sexo': 'sex', 'numinves' : '
                 year', 'idenpa': 'country', 'reedad': 'age',
46            'ciudad': 'city', 'tamciud': 'city_size'}

48 lat_2011 = pd.read_stata(r'python/Lat2011.dta',
      convert_categoricals = False, columns = cols2011)
49 lat_2013 = pd.read_stata(r'python/Lat2013.dta',
      convert_categoricals = False, columns = cols2013)
50 lat_2015 = pd.read_stata(r'python/Lat2015.dta',
      convert_categoricals = False, columns = cols2015)
51 lat_2016 = pd.read_stata(r'python/Lat2016.dta',
      convert_categoricals = False, columns = cols2016)
52 lat_2017 = pd.read_stata(r'python/Lat2017.dta',
      convert_categoricals = False, columns = cols2017)
53 lat_2018 = pd.read_stata(r'python/Lat2018.dta',
      convert_categoricals = False, columns = cols2018)
54 lat_2020 = pd.read_stata(r'python/Lat2020.dta',
      convert_categoricals = False, columns = cols2020)
```

```
lat_2011.rename(columns = dic2011, inplace = True)
lat_2013.rename(columns = dic2013, inplace = True)
lat_2015.rename(columns = dic2015, inplace = True)
lat_2016.rename(columns = dic2016, inplace = True)
lat_2017.rename(columns = dic2017, inplace = True)
lat_2018.rename(columns = dic2018, inplace = True)
lat_2020.rename(columns = dic2020, inplace = True)

#making the correct year appear in the "year" field

lat_2011.loc[lat_2011['year'] == 16, 'year'] = '2011'
lat_2015.loc[lat_2015['year'] == 18, 'year'] = '2015'

#the city list changed for Argentina in 2020 and city size is not
    avilable. We will only consider cities with city size in the
   dataset

lat_2018_ar = lat_2018[lat_2018['country'] == 32]
lat_2018_ar = lat_2018_ar[['city','city_size']].drop_duplicates()

city_dict = dict([(city, city_size) for city, city_size in zip(
   lat_2018_ar['city'].values, lat_2018_ar['city_size'].values)])

for key, value in city_dict.items():
     lat_2020.loc[lat_2020['city'] == key, 'city_size'] = value

lat_2020 = lat_2020[lat_2020['city_size'] > 0]

#joining dfs

dfs = [lat_2011, lat_2013, lat_2015, lat_2016, lat_2017, lat_2018
   , lat_2020]
lat_completo = pd.concat(dfs)

lat_completo.drop(['city'], axis = 1, inplace = True)

#removing people outside the analyzed age range 25-60

lat_completo = lat_completo.loc[lat_completo['age'] > 1]
lat_completo = lat_completo.loc[lat_completo['age'] < 4]

#ethnicity: 1 white, 0 non-white
```

```
94 etnia_dict = {6: 1, 1: 0, 2: 0, 3: 0, 4: 0, 5:0, 7:0, -1: 0, -2:
      0}
95 lat_completo['ethnicity'] = lat_completo['ethnicity'].map(
      etnia_dict)

97 country_dict = {32: 'argentina',
98 68: 'bolivia',
99 76: 'brasil',
100 152: 'chile',
101 170: 'colombia',
102 188: 'costa_rica',
103 214: 'rep_dominicana',
104 218: 'ecuador',
105 222: 'el_salvador',
106 320: 'guatemala',
107 340: 'honduras',
108 484: 'mexico',
109 558: 'nicaragua',
110 591: 'panama',
111 600: 'paraguay',
112 604: 'peru',
113 724: 'espana',
114 858: 'uruguay',
115 862: 'venezuela'}

117 lat_completo['country'] = lat_completo['country'].map(
      country_dict)

119 #drop NAs in years of education and turn into integers
120 lat_completo = lat_completo[lat_completo['anos_educ'] != 0]
121 lat_completo = lat_completo[lat_completo['anos_educ_pai'] != 0]
122 lat_completo['anos_educ'] = lat_completo['anos_educ'] - 1
123 lat_completo['anos_educ_pai'] = lat_completo['anos_educ_pai'] - 1
124 dict_educ = {13: 14, 14:17, 15: 14, 16:16}
125 lat_completo['anos_educ'] = lat_completo['anos_educ'].map(
      dict_educ).fillna(lat_completo['anos_educ'])
126 lat_completo['anos_educ_pai'] = lat_completo['anos_educ_pai'].map
      (dict_educ).fillna(lat_completo['anos_educ_pai'])
127 lat_completo = lat_completo[lat_completo['anos_educ'] > -1]
128 lat_completo = lat_completo[lat_completo['anos_educ_pai'] > -1]

130 #sex: 1 = female, 0 = male
```

```
131 lat_completo['sex'] = lat_completo['sex'] - 1

133 lat_completo = lat_completo.rename(columns = {'anos_educ': '
      educ_years', 'anos_educ_pai':'educ_parents'})

135 #get csv to run the model
136 lat_completo.to_csv('lat_completo.csv')
```

ANNEX D – MODELING

Code 2 – R code

```
library(tidyverse)
library(partykit)
library(party)
library(ineq)




df <- read.csv('lat_completo.csv', header = T)
df <- df[,-1]

data <- subset(df, !is.na(educ_years))



control = ctree_control(teststat = c("quad", "max"),
                        testtype = c("Bonferroni", "MonteCarlo",
                                     "Univariate", "Teststatistic
                                        "),
                        mincriterion = 0.95, minsplit = 20,
                           minbucket = 7,
                        stump = FALSE, nresample = 9999,
                           maxsurrogate = 0,
                        mtry = 0, savesplitstats = TRUE, maxdepth
                            = 0, remove_weights = FALSE)



countries <- list("argentina",      "bolivia" ,       "brasil",
           "colombia",       "costa_rica",     "chile"
                  ,"ecuador",        "el_salvador" ,   "guatemala
                     ",      "honduras",       "mexico",         "
                     nicaragua"
                  ,"panama"          ,"paraguay"       ,"peru"
                                ,"rep_dominicana" ,"uruguay"
                     , "venezuela" )

#computing trees

for (i in countries) {
```

```
29    coun = data[data$country == i,]
30    tree <- ctree(educ_years ~ city_size + sex + religion + citizen
          + ethnicity + educ_parents,
31                  data =  coun,
32                  control = control)
33    plot(tree,main = i, pop = T,
34         terminal_panel = node_terminal(tree,digits=1,id=T,fill=c('
             white','white')))
35    print(i)
36  }

38  #Forest scores and feature importance for each country

40  score <- numeric(18)
41  for (i in countries) {
42    coun = data[data$country == i,]
43    forest_output <- cforest(educ_years ~ city_size + sex +
        religion + citizen + ethnicity + educ_parents,
44                             data =  coun,
45                             controls = cforest_unbiased(ntree=
                                 100, mtry = 3), set.seed(10))
46    forest.importance <- varimp(forest_output, conditional = T)
47    round(forest.importance, 3)
48    dotchart(sort(forest.importance),
49             sub =i)
50    print(i)
51    country_score  <- ineq(predict(forest_output), type = 'Gini')
52    score <- append(score, country_score)
53    print(ineq(predict(forest_output), type = 'Gini'))
54  }

56  j <- 1

58  for (i in countries) {
59    coun = data[data$country == i,]
60    forest_output <- cforest(anos_educ ~ city_size + sex + religion
          + citizen + ethnicity + anos_educ_pai,
61                             data =  coun,
62                             controls = cforest_unbiased(ntree=
                                 100, mtry = 3), set.seed(10))
63    print(i)
64    country_score  <- ineq(predict(forest_output), type = 'Gini')
```

```
65    score[j] <- country_score
66    j <- j+1
67    print(ineq(predict(forest_output), type = 'Gini'))
68  }

70  #forest score and feature importance for the whole continent

72  forest_output <- cforest(anos_educ ~ city_size + sex + religion +
        citizen + ethnicity + anos_educ_pai,
73                           data =  df,
74                           controls = cforest_unbiased(ntree= 100,
                               mtry = 3), set.seed(10))
75  forest.importance <- varimp(forest_output, conditional = T)
76  dotchart(sort(forest.importance))

78  country<- unlist(countries)

80  scores_df <- data.frame(country, score)

82  scores_df <- scores_df[order(scores_df$score),]

84  dotchart(scores_df$score, labels = scores_df$country, pch = 20,
       cex = 1)
```