Telecommunications Policy 47 (2023) 102598

Contents lists available at ScienceDirect

Telecommunications Policy

journal homepage: www.elsevier.com/locate/telpol

The use of ICTs and income distribution in Brazil: A machine learning explanation using SHAP values

Gabriel Paes Herrera [a,*], Michel Constantino [b], Jen-Je Su [a], Athula Naranpanawa [a]

[a] *Department of Accounting, Finance and Economics, Griffith University, 170 Kessels Road, Nathan, Queensland, 4111, Australia*
[b] *Dom Bosco Catholic University - UCDB, Av. Tamandaré 6000, Campo Grande, Brazil*

ARTICLE INFO

Keywords:
Neural network
XGBoost
Inequality
Gender gap
Digital divide

ABSTRACT

This study explores the complex relationship between information and communication technologies (ICTs) and socioeconomic characteristics. We employ a cutting-edge explainable machine learning approach, known as SHAP values, to interpret an XGBoost and neural network model, as well as benchmark traditional econometric methods. The application of machine learning algorithms combined with the SHAP methodology reveals complex nonlinear relationships in the data and important insights to guide tailored policy-making. Our results suggest that there is an interaction between education and ICTs that contributes to income prediction. Furthermore, level of education and age are found to be positively associated with income, while gender presents a negative relationship; that is, women earn less than men on average. This study highlights the need for more efficient public policies to fight gender inequality in Brazil. It is also important to introduce policies that promote quality education and the teaching of skills related to technology and digitalization to prepare individuals for changes in the job market and avoid the digital divide and increasing social inequality.

1. Introduction

The rapid and constant development of information and communication technologies (ICTs) has accelerated globalization. There is increasing recognition among global leaders of the opportunities and benefits of digitalization (OECD, 2017). The expansion of the ICT industry has direct and indirect impacts on social, economic, and cultural development by promoting, for example, new jobs, innovation, and productivity increases (Matsubayashi et al., 2019). The relationship between ICTs and socioeconomic variables is complex and often exhibits nonlinear patterns as demonstrated by Albiman and Sulong (2017) and Richmond and Triplett (2018). Therefore, the main objective of this study is to uncover any existing nonlinear relationships between ICTs and income by employing machine learning (ML) models and a cutting-edge approach for ML explainability.

Brazil is the ninth largest economy in the world by nominal gross domestic product (GDP) which reached approximately US$2.06 trillion in 2022. In terms of purchasing power parity (PPP), the country is ranked sixth largest in the world (World Bank, 2022). The service sector is the largest contributor to the Brazilian economy, accounting for approximately 65% of its GDP followed by the industrial and agricultural sectors. The country is also a significant player in the global marketplace being a major exporter of

* Corresponding author. Department of Accounting, Finance & Economics Griffith University, Nathan campus 170 Kessels Road, Building N50, Room 0.35, QLD, 4111, Australia.

E-mail addresses: gabriel.paesherrera@griffithuni.edu.au (G.P. Herrera), michel@ucdb.br (M. Constantino), j.su@griffith.edu.au (J.-J. Su), a.naranpanawa@griffith.edu.au (A. Naranpanawa).

https://doi.org/10.1016/j.telpol.2023.102598
Received 3 October 2022; Received in revised form 20 June 2023; Accepted 21 June 2023
Available online 5 July 2023

commodities such as soybeans, iron ore, beef, and coffee (IBGE, 2021).

Further, Brazil is the seventh most populous country in the world. As an emerging economy with an upper-middle income population, the country has experienced a consistent increase in the utilization of ICTs. It is estimated that, by 2025, Brazil will be one of the five largest smartphone markets (Matsubayashi et al., 2019). According to the OECD (2020a), between 2008 and 2019, the number of broadband subscribers tripled in the country, reaching 32.9 million subscriptions. As highlighted by Silva et al. (2020), internet access is considered an essential service by families in Brazil. Across the country, a large part of the population accesses the internet exclusively via mobile phones, which restricts them from performing more sophisticated digital activities (OECD, 2020a). Currently, connectivity and digital inclusion are essential, and as stated by Silva et al. (2020), income inequality in Brazil might be associated with disparities in ICT access.

The absorption and consequent positive impact of technology on people's lives varies from place to place. This is closely related to the diffusion of innovations theory from Rogers (1962), which states that people's perception of aspects such as advantage, compatibility, and complexity of an innovation will determine the diffusion of such innovation among a given population. Further, according to the knowledge economy theory, the creation and dissemination of knowledge is one of the most important resources of the 21st century. In this sense, few researchers have demonstrated the role of ICTs in enhancing the dissemination and consequent impact of knowledge on the economy and, consequently, on people's income (Johannessen & Olsen, 2010).

As a proxy of ICT usage and diffusion, researchers usually employ variables such as the number of internet users, mobile phone subscriptions, and broadband subscriptions (see, e.g., Albiman & Sulong, 2017; Cheng et al., 2021). Although relevant, these variables do not fully cover ICTs and much information is not accounted for. The economic impact of ICTs has been explored before using traditional econometric methods. For example, Saidi and Mongi (2018) employed a vector error correction model to analyse the impact of ICTs on economic growth in high-income countries and found a bidirectional effect between the number of internet users and GDP. Chien et al. (2020) used the generalized-momentum method and found that telephone and internet diffusion positively influences financial development. The application of ML techniques in this area is still very limited.

The advantages of machine learning models are well documented in a wide range of research areas. These models are capable of uncovering complex nonlinear relationships among several variables, but the downside is their lack of interpretability. This is possibly the reason why ML methods are rarely applied in socioeconomic studies. Decision tree-based models are usually chosen when researchers want to employ ML models and at the same time still have some level of transparency – see, for example, Hidalgo et al. (2020) and Antulov-Fantulin et al. (2021) – while models such as artificial neural networks (ANNs) are left aside. Explainable artificial intelligence (AI) is a research field concerned with the interpretation and transparency of AI models. Recently, a few different approaches have been proposed to explain machine learning models, the most popular being SHAP (SHapley Additive exPlanation) introduced by Lundberg and Lee (2017). To the best of our knowledge, this methodology has never been applied to explore the association between ICTs and socioeconomic factors.

This paper expands the research frontier and makes several contributions. First, we contribute to the literature by exploring the role of ICTs on income distribution from a machine learning perspective, as ML is known to be able to capture complex nonlinear relationships and deliver more accurate results. In addition to tree-based ML models, we also employ an ANN algorithm in the framework. Furthermore, as ICT indicators, we use a broad set of variables that account for different related aspects, namely, technological devices, type of internet connection, and digital engagement.

Second, to interpret the results and uncover nonlinear patterns revealed by the ML models, we employ a cutting-edge methodology, i.e., SHAP values. The use of an explainable AI approach allows us to fully leverage ML. Thus, we explore this issue using models capable of finding complex characteristics intrinsic to the data, and then employ the SHAP values methodology to interpret and understand the analysis process and the patterns captured by the ML models. Third, we demonstrate that the so-called black-box models can be transparent and interpreted similarly to traditional econometric techniques such as linear regression and elastic net. We demonstrate how this approach can be used to support decision-making and policy formulation.

Regarding the research questions addressed in this study, we highlight the following: i) What are the complex nonlinear relationships between socioeconomic characteristics, such as education, age, and gender, and income distribution? ii) How does the application of SHAP values in the interpretation of machine learning models enhance our understanding of the complex relationship between information and communication technologies and socioeconomic characteristics? iii) Does the interaction between education and information and communication technologies affect income prediction? iv) How can the utilization of SHAP values combined with machine learning algorithms provide actionable guidance for tailored policy-making to address the digital divide and socioeconomic inequality, particularly in the context of education, age, and gender dynamics?

The remainder of this paper continues in five sections as follows. Section 2 presents a literature review of relevant studies. Section 3 describes the data and methods employed in this research. We then present and discuss the results in Section 4. Finally, Section 5 presents the conclusions and implications.

2. Literature review

2.1. The Brazilian case and the impact of ICTs

Brazil is a continent-sized country with a proportional population of approximately 214 million people, which is mainly concentrated in the Southeast region. The country is divided into 26 states and a federal district that come under five major regions. The nation is considered an emerging economy and is responsible for a large part of the international agricultural market supply, being a major exporter of, for example, coffee, soybeans, maize, beef, and sugar. The country experienced a long period of economic growth and social progress until 2014, when Brazil entered a period of recession due, among other factors, to a serious political corruption scheme.

Currently, the IMF (2021) projects only a modest 1.4% GDP growth for Brazil in 2022, while other countries such as Mexico and Argentina are expected to grow by approximately 3%. The economic situation of the largest country in South America is challenging; the annual GDP per capita in Brazil (US$ 6796.84) is the lowest among OECD countries, except Colombia, and the unemployment rate registered in 2020 was 13.5% (World Bank, 2022). Brazil has been fighting poverty for many years but without success. The government social program "Bolsa Familia" is one of the largest and oldest conditional cash transfer programs in the world and has been playing an important role in reducing poverty across the country (Fruttero et al., 2020). However, the sharp contrasts among the regions of the country, the inefficiency of public agents and corruption are still obstacles.

Inequality is a well-known problem in the country, where it is higher than in other OECD countries. According to the OECD (OECD, 2020a), 42% of the country's total income is concentrated in 10% of the population. This problem is exacerbated by inefficient public spending, corrupt practices, a complex tax scheme, and a precarious justice system. Providing high quality education is considered the most efficient and sustainable approach to promoting social and economic inclusion in Brazil. However, the country has one of the strongest correlations between socioeconomic status and educational outcomes, which reflects the unequal educational services provided in schools across the country (OECD, 2020b).

As highlighted by the OECD (OECD, 2020a), the completion rate of primary education in Brazil is significantly lower than the OECD average, and less than 50% of the population attends secondary education. According to Barakabitze et al. (2019), the teaching and learning experience in Africa has improved significantly following the adoption and popularization of ICT among teachers, schools, and students. While digital technologies and ICTs can effectively improve education and training programs, Brazil still faces a major problem, as in 2018 approximately 23% of adults reported that they had never accessed the internet. Furthermore, the country is developing its digital capacity and has seen the export of computer services increase by 4% in 2020, but it is still far behind most countries (WTO, 2021). More and better investments might be the key to the diffusion of technologies and digital literacy in Brazil. As shown in the study conducted in Italy and Spain by Llorent-Vaquero et al. (2020) local and national public policies are essential to guarantee access to technological devices and digital competence.

In Brazil, many socioeconomic aspects hinder the adoption of ICTs. As previously mentioned, the theory of Rogers (1962) on the diffusion of innovations highlights that the individual perception of an innovation is deeply related to its adoption. In the case of Brazil, different socioeconomic factors impact the adoption of digital technology innovations, for example, the low average income across the country means that only a small percentage of the population has access to these technologies. Additionally, a deficient educational system results in a large part of the population perceiving technological innovations as complex and difficult to use. These particular characteristics of Brazil act as obstacles to the diffusion of ICTs in the country and negatively impact development and growth. As demonstrated by Novak et al. (2018) the digitization of various industries through the use of ICTs has the potential to add €200 billion to Central and Eastern Europe's GDP by 2025. Further, the work of Gordon (2003) explores the knowledge economy theory and shows that a significant increase in ICT investments in the U.S. resulted in considerable productivity acceleration. Briglauer and Gugler (2017) analysed panel data of 27 EU members between 2003 and 2015 and found that a 1% increase in basic broadband adoption has the potential to increase GDP by approximately 0.015%.

The relationship between ICTs and socioeconomic variables has increasingly become an object of study due to constant technological development and its impact on society. One of the first studies to address this issue is that of Vu (2011), who analysed a sample of 102 countries from 1996 to 2005 using traditional econometric methods and revealed strong evidence of the positive and large impact of ICT on economic growth. In the last few years, researchers have been exploring this topic in different ways, but the results are inconsistent. For example, Cheng et al. (2021) explored the relationship between ICT diffusion and GDP per capita growth using the generalized method of moments and found that there is a positive association in high-income countries, while the effect is unclear in middle and low-income nations. Furthermore, Ashraf Ganjoei et al. (2021) employed a fuzzy nonlinear regression to analyse the effect of ICTs on income distribution and reported that income inequality decreases as ICT use increases.

On the other hand, a panel regression analysis conducted by Richmond and Triplett (2018) showed that the effect of ICT on income inequality is subject to the measure of the latter and the specific type of ICT, as well as economic and political characteristics. Njoh (2018) analysed more than 50 countries in Africa and found a positive relationship between ICTs and development, as measured by the human development index (HDI). Likewise, a study conducted in Brazil revealed that entrepreneur mothers, who balance motherhood and their business, consider that ICTs have a positive impact on the financial performance of their companies. Furthermore, the study reveals that, especially after COVID-19, social media and mobile devices are key tools for business success (de Oliveira Malaquias et al., 2021). There are no studies that employ ML interpretation methods to assess this issue and explore the complex relationships that exist in the data, and this might be one of the reasons why previous research has reported mixed results.

In many real world situations, the relationships between variables are complex and follow nonlinear patterns that can arise from interactions between variables and threshold effects, for example. Ignoring nonlinearities in a dataset can lead to inaccurate models and predictions, as well as missing opportunities to uncover important insights. While regression models such as polynomial regression are capable of modelling nonlinear relationships, machine learning models present a more robust and effective way of modelling complex nonlinear interactions and threshold effects (Chen & Guestrin, 2016).

In relation to socioeconomic problems, such as the one explored in this study, Richmond and Triplett (2018) state that a possible reason for the appearance of nonlinear relationships is that, contrary to the Kuznets theory, the process of economic development causes structural changes and labour market transformations that have complex implications for income inequality that are unique to each country. In this sense, we explore these nonlinear relationships and thresholds that are unique to Brazil.

2.2. Machine learning interpretation

Traditional regression-based approaches are known to have difficulty modelling nonlinear relationships and to be sensitive to multicollinearity (Hastie et al., 2009). On the other hand, machine learning methods are capable of uncovering complex patterns and interaction effects but are often criticized for their "black-box" nature (Carbo-Valverde et al., 2020). The explainable AI research area has been working to mitigate this problem and make ML models more transparent and easier to understand. According to Antwarg et al. (2021), ML interpretation builds trust with users and helps to identify possible model biases.

Recently, SHAP values have drawn the attention of researchers as an ML interpretability approach due to their robustness and efficiency. SHAP values were introduced by Lundberg and Lee (2017) and are based on the cooperative game theory concept proposed by Shapley (1953). As stated by Booth et al. (2021), the SHAP technique can empirically reveal the importance of each feature to the model output, allowing for combinations of features and nonlinear interactions. In addition, SHAP values are the only approach for feature importance interpretation that fulfils the mathematical properties of consistency and local accuracy (Li et al., 2020).

It is important to note that machine learning models, as well as the SHAP methodology, also present disadvantages that need to be considered, such as algorithmic bias, training data bias, overfitting, and increased training time. Nonetheless, the literature demonstrates that the advantages of these models often offset the disadvantages and both ML and SHAP values models have been effectively implemented in several studies (Athey & Imbens, 2019; Lundberg & Lee, 2017).

Researchers have successfully employed SHAP values for ML interpretation in the medical research area. For example, Li et al. (2020) used gradient-boosted tree models and SHAP values to improve the prediction of prostate cancer mortality risk. Booth et al. (2021) applied a support vector machine model and SHAP values to reveal important features related to COVID-19 mortality. A gradient tree boosting model and SHAP values were employed by Ballester et al. (2021) to analyse the problem of suicide risk among young adults.

In addition, this approach has been used for machine learning explanation in different areas and has shown promising results.

Table 1
Description of the variables selected from the National Household Sample Survey.

Variable	Description	Value
Panel A: Socioeconomic variables		
North	North region	0 – No, 1 – Yes
Northeast	Northeast region	0 – No, 1 – Yes
Midwest	Midwest region	0 – No, 1 – Yes
South	South region	0 – No, 1 – Yes
Southeast	Southeast region	0 – No, 1 – Yes
Location	Location	1 – Urban, 2 – Rural
Gender	Gender	1 – Male, 2 – Female
Age	Age	Continuous (Adults 18+)
Level_edu	Highest level of education achieved [1]	1 (lowest) to 7 (highest)
Workforce	Workforce status	0 – Unemployed, 1 – Employed
BPC_LOAS	Received income from BPC-LOAS benefit [2]	0 – No, 1 – Yes
Bolsa_Fa	Received income from Bolsa-Familia program [2]	0 – No, 1 – Yes
Other_social	Received income from other government social programs [2]	0 – No, 1 – Yes
Gov_pension	Received retirement income or pension from a government-pension institution [2]	0 – No, 1 – Yes
Work_insurance	Received income from unemployment insurance [2]	0 – No, 1 – Yes
Donation	Received income from child support, donation, or allowance [2]	0 – No, 1 – Yes
Rent_lease	Received rent or lease income [2]	0 – No, 1 – Yes
Other_income	Received other income (study scholarship, savings account, financial investment, etc.) [2]	0 – No, 1 – Yes
Total_income	Income from all sources (Target variable)	Continuous in BRL currency
Panel B: ICTs variables		
Internet access devices:		
Net_computer	Accessed internet through computer [3]	0 – No, 1 – Yes
Net_tablet	Accessed internet through tablet [3]	0 – No, 1 – Yes
Net_mobile	Accessed internet through mobile phone [3]	0 – No, 1 – Yes
Net_tv	Accessed internet through smart TV [3]	0 – No, 1 – Yes
Net_other	Accessed internet through other devices [3]	0 – No, 1 – Yes
Type of internet connection:		
3G_4G	Accessed internet via 3G/4G [3]	0 – No, 1 – Yes
Dial_up	Accessed internet via dial-up [3]	0 – No, 1 – Yes
Broadband	Accessed internet via broadband [3]	0 – No, 1 – Yes
Digital engagement:		
Email	Used internet to send/receive email [3]	0 – No, 1 – Yes
Messages	Used internet to send/receive text, voice, or image messages [3]	0 – No, 1 – Yes
Calls	Used internet for voice/video calls [3]	0 – No, 1 – Yes
Videos	Used internet to watch videos/movies/series [3]	0 – No, 1 – Yes
Mobile_net	Personal mobile phone has access to internet	0 – No, 1 – Yes

Note: This table presents the variables employed in the analyses, their description and the values attributed to them. [1] 1 - Less than one year of study; 2 – Incomplete primary education; 3 – Complete primary education; 4 – Incomplete high school; 5 – Complete high school; 6 – Incomplete degree; 7 – Complete degree. [2] The month before the interview. [3] In the last three months before the interview.

Rico-Juan and de La Paz (2021) applied the SHAP technique and the random forest method to identify important variables that predict house prices in Spain. Parsa et al. (2020) used an XGBoost machine learning model combined with SHAP values to reveal important features for real-time traffic accident detection. Cooray et al. (2021) applied the SHAP approach and an extreme gradient boosting model to identify the most important variables related to tooth loss among aged adults in Japan. Antwarg et al. (2021) demonstrated the advantages and how SHAP values can be used for anomaly detection using an autoencoder model in artificial and real-world datasets. The SHAP approach was used to quantify the contributing factors of air pollution in urban areas using linear regression and a random forest model by Gu et al. (2021). John-Mathews (2022) used logistic regression and the random forest method to demonstrate the application of the SHAP methodology in a popular bank credit database and discuss the construction of ethical AI models. To the best of our knowledge, SHAP values have not yet been applied to explore the impact of ICTs on socioeconomic variables.

3. Materials and methods

3.1. Data source

The analyses conducted in this research are based on data from the Brazilian Institute of Geography and Statistics (IBGE). The IBGE is a public organization and the main institution that performs data collection, analysis, and statistics in Brazil. The institute is responsible for all activities related to data analysis and statistics, such as gathering data related to key economic indicators, running the population census, and analysing industry performance. The Continuous National Household Sample Survey (Pesquisa Nacional por Amostra de Domicílios Contínua – PNADC) is a quarterly nationwide survey carried out by the IBGE to monitor the country's socioeconomic development. The survey collects data at the household level and on its individual residents based on a two-step stratified probability sampling and using the computer-assisted personal interviewing (CAPI) technique.

The sampling methodology is designed to produce results at multiple levels, such as the national level, major regions, states, and metropolitan areas. Furthermore, the ongoing nature of this survey makes it possible to monitor changes in the behaviour of indicators, which is achieved using a rotating panel format and interviewing each selected household in five consecutive quarters. The survey provides anonymized microdata information on thousands of Brazilians and includes socioeconomic variables and any relevant bespoke studies. We analyse data from the last quarter of the 2019 Continuous National Household Sample Survey, conducted by the institute, which includes a one-time complementary study on ICTs in Brazil. After selecting the variables relevant to this study, the data was cleaned to remove any missing values or outliers and the final database contained 216,883 individuals from all 26 states and the

Table 2
Summary statistics.

	Mean	Min.	Max.	Std. dev.	Kurtosis	Skewness
North	0.1096	0	1	0.3124	4.2478	2.4996
Northeast	0.2616	0	1	0.4395	−0.8225	1.0851
Midwest	0.1169	0	1	0.3213	3.6887	2.3851
South	0.2016	0	1	0.4012	0.2125	1.4874
Southeast	0.3104	0	1	0.4626	−1.3280	0.8198
Location	1.1577	1	2	0.3644	1.5296	1.8787
Gender	1.5033	1	2	0.5000	−1.9998	−0.0134
Age	41.5998	18	99	14.4994	−0.4708	0.4621
Level_edu	4.5677	1	7	1.8060	−1.1549	−0.2117
Workforce	0.8376	0	1	0.3689	1.3499	−1.8303
BPC_LOAS	0.0097	0	1	0.0980	98.0444	10.0022
Bolsa_Fa	0.0762	0	1	0.2654	8.2025	3.1941
Other_social	0.0028	0	1	0.0527	354.6843	18.8860
Gov_pension	0.1468	0	1	0.3539	1.9863	1.9966
Work_insurance	0.0095	0	1	0.0970	100.2949	10.1141
Donation	0.0284	0	1	0.1662	30.2040	5.6748
Rent_lease	0.0351	0	1	0.1839	23.5666	5.0563
Other_income	0.0173	0	1	0.1303	52.8852	7.4084
Total_income	2498.1845	1	100761	3618.3251	93.7596	7.0779
Net_computer	0.4432	0	1	0.4968	−1.9477	0.2288
Net_tablet	0.0985	0	1	0.2980	5.2631	2.6950
Net_mobile	0.9944	0	1	0.0746	173.6612	−13.2537
Net_tv	0.3067	0	1	0.4611	−1.2967	0.8386
Net_other	0.0085	0	1	0.0916	113.0749	10.7272
3G_4G	0.8362	0	1	0.3701	1.2992	−1.8164
Dial_up	0.0035	0	1	0.0594	277.0425	16.7045
Broadband	0.8432	0	1	0.3636	1.5654	−1.8882
Email	0.6239	0	1	0.4844	−1.7382	−0.5117
Messages	0.9684	0	1	0.1750	26.6719	−5.3546
Calls	0.9242	0	1	0.2646	8.2786	−3.2060
Videos	0.8710	0	1	0.3352	2.8982	−2.2132
Mobile_net	0.9833	0	1	0.1281	54.8976	−7.5430

Note: This table presents the descriptive statistics of the variables employed in the analyses.

federal district. Table 1 presents the variables and their descriptions.

The data accounts for socioeconomic variables such as age, gender, employment status, education, and financial social assistance. The information provided also includes the region where the individual lives, which can be one of the five main regions in Brazil, and whether it is an urban or rural area. Additionally, the survey data contains information about the use of ICTs in the country, including the electronic devices used to access the internet, the type of internet connection, and the digital activities normally carried out. Table 2 presents the summary statistics of the variables.

3.2. Linear regression and elastic net models

In this study, the output variable considered is the total income of each individual. We calculate a simple multiple regression model as a benchmark. Although this methodology has limited capacity to interpret nonlinear relationships, it is a fast and efficient approach commonly employed in econometrics. The model follows a simple ordinary least squares estimation and is estimated as:

$$\mathbf{Total\_income_i} = \mathbf{c} + \mathbf{X_i}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_\mathbf{i} \tag{1}$$

where the dependent variable considered is the total income of each individual and **X** is a vector containing the independent variables, i.e., the socioeconomic and ICT-related variables presented in Table 1.

We further employ an extension of linear regression models, i.e., the elastic net model, which helps to overcome problems related to variance and bias in the estimation of the linear regression model. According to Hastie et al. (2009), the elastic net method can be employed with any linear model, especially for regression and classification problems. The elastic net approach was introduced by Zou and Hastie (2005) as a regularized linear regression technique that improves model stability by changing the loss function. The method combines both penalties of the lasso (L1) and ridge (L2) techniques, which can be balanced according to the problem to improve performance. The linear regression and elastic model are implemented using Python (Van Rossum & Drake, 2009) and the Scikit-learn library (Pedregosa et al., 2011).

3.3. Machine learning models

Machine learning models are well known for their flexibility and ability to model complex large datasets. We exploit this advantage and employ two ML algorithms in our study. First, we train a feedforward artificial neural network, which is a powerful and popular ML method; see, for example, John-Mathews (2022) and Rico-Juan and de La Paz (2021). An ANN algorithm is characterized by having an input layer, one or more hidden layers, and an output layer. Each layer contains **n** neurons and each neuron in one layer connects to every neuron in the next layer with a certain weight, creating a fully connected neural network (Lantz, 2013).

The training process is an important part of the ML analysis that involves an iterative process of building and refining a predictive model through hyperparameter tuning. As there is no methodology to define the optimal structure and parameters of a neural network, this process is basically based on trial and error (Samarasinghe, 2016). We tested different settings and parameters and defined an ANN containing two hidden layers with 50 neurons in each, rectified linear unit activation function, and Adam as the network stochastic gradient descending (SGD) optimization algorithm. Adam is introduced by Kingma and Ba (2014) as an efficient approach that computes individual adaptive learning rates for each parameter, allowing for more efficient convergence on non-stationary objectives. The model was implemented using the Keras Library (Chollet, 2015) and Python.

Furthermore, we apply a tree-based machine learning method known as extreme gradient boosting (XGBoost). The algorithm was introduced by Chen and Guestrin (2016) as an alternative to improve the performance and speed of gradient boosted decision trees. The boosting technique consists of creating trees in a sequence such that the new tree learns from the previous tree to reduce the error. Therefore, the model is based on so-called weak learners that develop to create a strong final learner which also reduces bias and variance (Friedman, 2001; James et al., 2013). XGBoost is an ensemble learning technique that has become popular for its efficiency, overall performance, speed, and scalability. The model optimizes the gradient boosting algorithm using parallel processing and employs regularization to avoid overfitting. The model is tested with different sets of parameters and the final form selected for the analyses contains 50 gradient boosted trees, a maximum tree depth equal to six, and a learning rate of 0.3 to prevent model overfitting. We applied this model using the XGBoost library (Chen & Guestrin, 2016) and Python.

3.4. SHAP values

Machine learning models are often criticized for their lack of interpretability and transparency. While these models generally achieve better performance and accuracy, it is also important that users have the option to examine how an ML model processed a specific dataset. Lundberg and Lee (2017) introduced the SHAP values approach as an alternative to interpreting machine learning models. According to Giudici and Raffinetti (2021), the advantage of this technique is that it allows us to reveal the importance of each explanatory feature for any ML prediction model.

In a machine learning model, each input feature has a certain impact on the output prediction. SHAP values provide a measure of the importance of each feature to the model's prediction for a given data point. They can be used to understand how a model is making decisions, to identify which features are most important, and to debug the model when it is not performing as expected. Additionally, Antwarg et al. (2021) highlight that SHAP values are a user-friendly method that is more consistent and robust than other explainable AI models, such as local interpretable model-agnostic explanations (LIME).

SHAP values provide a comprehensive and fair way to assign feature importance in machine learning models. The calculation of SHAP values involves a combinatorial approach based on cooperative game theory and the concept of Shapley values. To calculate SHAP values, the algorithm considers all possible coalitions of features. A coalition represents a subset of features that work together to make predictions. The algorithm iterates through all the possible combinations of features and evaluates their importance by comparing the prediction when a feature is included versus when it is absent.

The computation process initializes with an empty coalition and sets the initial SHAP value for each feature to zero. The order in which the features are added to the coalition is determined randomly. For each feature, the algorithm computes its contribution to the coalition by comparing the predictions when the feature is included and when it is absent. This is done by considering all possible permutations of features orders within the coalition. Finally, the average contribution of each feature is calculated by dividing the accumulated contribution of a particular feature by the total number of coalitions.

The resulting SHAP values represent the importance or influence of each feature on the model's predictions. Higher positive SHAP values indicate features that increase the predictions, while higher negative values indicate features that decrease the predictions. The resulting SHAP values provide a comprehensive and fair explanation of feature importance. They satisfy the properties of local accuracy, where the sum of SHAP values for all features equals the difference between the prediction for a specific instance and the average prediction for the entire dataset. This ensures that the feature weights are consistent at the individual level.

Further, SHAP values is a model-agnostic approach and can be applied to any model, regardless of its complexity or type. Whether it is a tree-based model, linear model, deep neural network, or an ensemble, SHAP values can be computed for individual predictions or aggregated across the entire dataset.

It is important to note that the SHAP values methodology includes simplified assumptions used to untangle the complexity of problems and improve performance, but these assumptions may not hold in real-world scenarios. Some assumptions adopted in the model include independence among features, homogeneity of feature effects across different instances or observations, and consistent underlying data distribution throughout the computation.

Finally, Li et al. (2020) emphasize that SHAP values create a visual representation of the interaction between two variables without requiring any arbitrary selection or knowledge from researchers. The methodology has been successfully employed in different research areas; see, e.g., Ballester et al. (2021) and Parsa et al. (2020). We implemented SHAP values using the SHAP library from Lundberg and Lee (2017) and Python.

4. Results and discussion

In this study, we evaluate the relationship between socioeconomic variables and ICT-related features using traditional econometric models and machine learning techniques. First, we examine the accuracy of each selected model in predicting the target variable, i.e., each individual's total income. To test the robustness of the models' predictions, we employ a 5-fold cross-validation technique as in Gu et al. (2021) and Parsa et al. (2020). The k-fold approach is a popular cross-validation technique used by researchers to verify the performance and stability of a model. The approach consists of randomly dividing the data into k folds, in this case 5, and using one of the folds as a test set and the remaining ones for training. The process is repeated until all folds have been treated as a test set once, and the final accuracy is an average of all. We measured accuracy using the root mean square error (RMSE) and mean absolute error (MAE) loss functions. The results are shown in Table 3.

The analysis shows that the ANN model achieved the best results in predicting the total income of individuals. There was a significant error reduction when employing machine learning models compared to using traditional methods, which is consistent with the literature (see, e.g., Carbo-Valverde et al., 2020; Cooray et al., 2021). The superior performance of machine learning models suggests that the relationships and patterns that exist in our dataset are likely complex and nonlinear.

Performance comparison between models is usually assessed using loss functions such as RMSE and MAE, and there is no agreement among researchers regarding the best measure of accuracy (Hyndman & Athanasopoulos, 2018). Therefore, we employ two widely used error measures, namely RMSE and MAE, which have been successfully implemented in similar studies recently, see for example Chen et al. (2019) and Thongpeth et al. (2021), along with a cross-validation procedure to ensure robustness in the results. We further assess the robustness of the results using the modified Diebold-Mariano test (M-DM) (Harvey et al., 1997) which tests whether the difference between the results of two models is statistically significant (Table 4). The null hypothesis of same accuracy level between the models can be rejected according to the test p-value and a positive test coefficient demonstrates the superior performance of model B over model A.

Table 3
Predictive accuracy of models using k-fold cross-validation.

	RMSE	MAE
Linear regression	3034.8095	1526.8962
Elastic net	3035.0126	1525.1063
XGBoost	2820.3488	1248.4862
Neural network	2803.4013	1221.0620

Note: This table presents the accuracy performance of all four models according to the RMSE and MAE loss functions in forecasting the income variable. We estimate each model using a k-fold cross-validation procedure.

Table 4
Modified Diebold-Mariano test results.

Model B \ Model A	Linear regression	Elastic net	XGBoost
Neural network	27.3335***	27.1598***	3.4780***
XGBoost	24.3498***	24.2014***	–
Elastic net	−1.6753*	–	–

Note: This table presents the results of the modified Diebold-Mariano robustness test for the pairwise comparison between models. ***p-value<0.01; **p-value<0.05; *p-value<0.1.

According to the M-DM test results, the differences between the models are all statistically significant and the neural network model presents the best performance. The results demonstrate that ML algorithms are better at capturing complex relationships and revealing information about the association between ICT variables and the socioeconomic characteristics of individuals in Brazil. Therefore, we proceed and explore the interpretation of these models and the variables that were considered the most important.

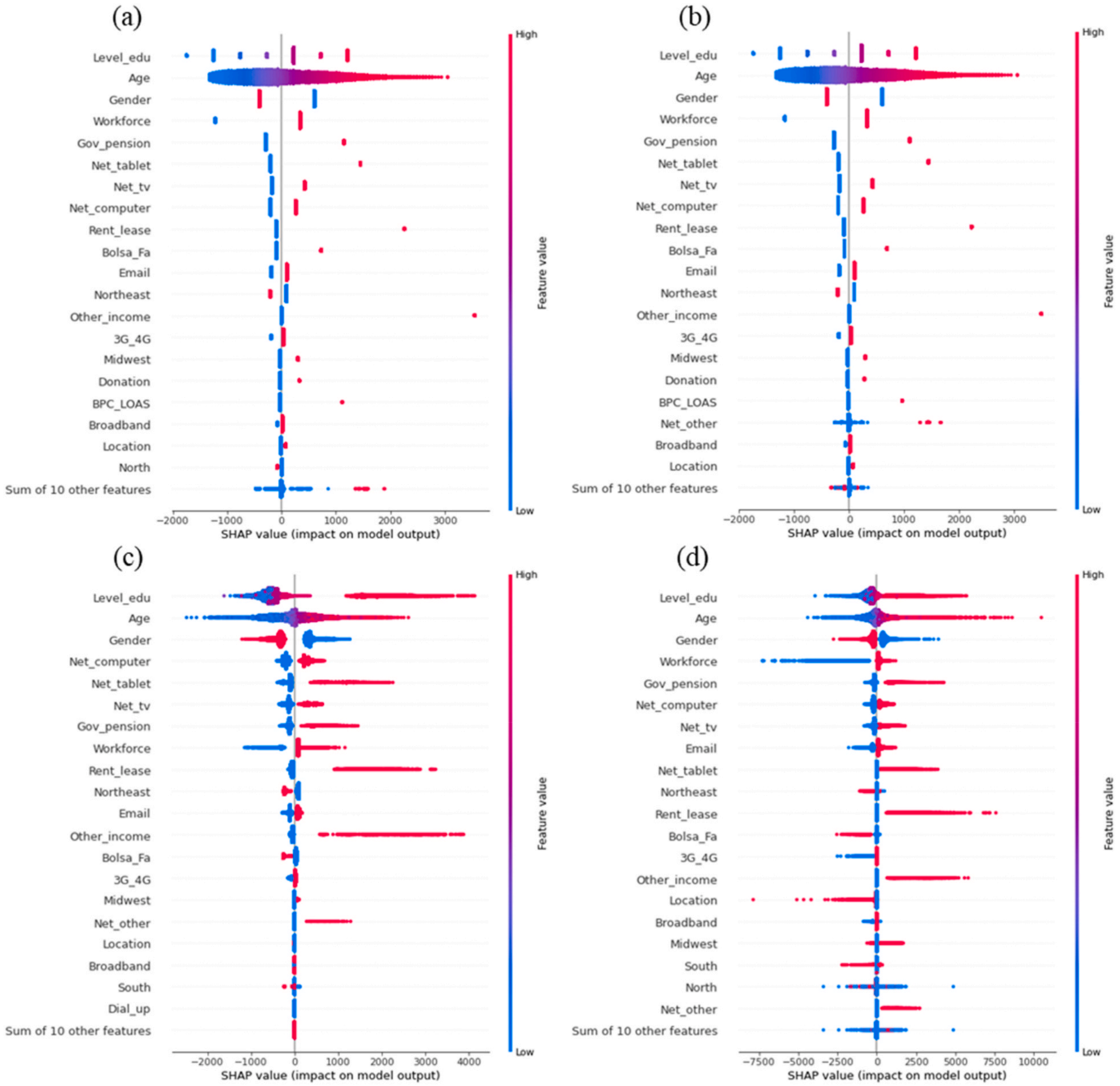

Fig. 1. SHAP summary plot of the most important variables considered by each model. (a) Linear regression. (b) Elastic net. (c) XGBoost. (d) Neural network.

As discussed in Section 2, one of the main drawbacks of ML methods is their lack of transparency and interpretability. To overcome this challenge, we use the SHAP values approach to gain insights revealed during the training process of machine learning algorithms. For each model, we estimate the importance of each predictor to the model output. Fig. 1 presents the average importance of the 20 most important variables considered by the models in order of significance, with the most important feature at the top. Each dot represents an observation and stacks along each variable row to show density, with the x position of the dot determined by the SHAP value. Additionally, dots are coloured according to their original value for that specific feature on a scale from red (high) to blue (low). For example, for the neural network model, Fig. 1 (d), shows that level of education is the most important variable considered by the model. The position of most red dots (high level of education) on the right side of the x axis reveals that high educational levels positively impact the model output.

The SHAP values charts are a result of the prediction analysis detailed previously, which employs a 5-fold cross-validation technique. Thus, the charts display the last test set of 5 used during the training process, i.e., 20% of the data or 43,377 points. As we can see, the linear regression and elastic net models present very similar results in relation to the variables considered important and their order of significance. On the other hand, the neural network and XGBoost models differ considerably.

The SHAP values method is a robust interpretation approach that fulfils the mathematical properties of consistency and local accuracy (Li et al., 2020). A detailed explanation of the theory and the desirable properties implemented in it is provided by Lundberg and Lee (2017). Considering this, we note that traditional statistics and machine learning differ in many ways, both being equally important. The conventional definition of "proper statistical inference" has been constantly debated recently (see, e.g., Athey & Imbens, 2019; Breiman, 2001; Hastie et al., 2009) as the non-parametric nature of machine learning algorithms often steps outside of conventional parameters. However, that does not mean that those models are less significant. In light of this, Table A1 in the appendix presents the coefficients for the linear regression and elastic net models.

According to the SHAP results, some interesting insights can be easily detected in the model's summary plots. First, the top three most important variables are the same for all models, namely, level of education, age, and gender. Second, eight of the ten most influential variables are the same across models, although their order of significance differs between ML algorithms and traditional econometric approaches. Third, the ten most important variables include three ICT-related features. In general, all four models seem to consider almost the same set of variables. The main difference appears to be the ability of each method to model the relationships between variables and capture a greater amount of information hidden in the dataset, which affects the order of feature importance considered by each model.

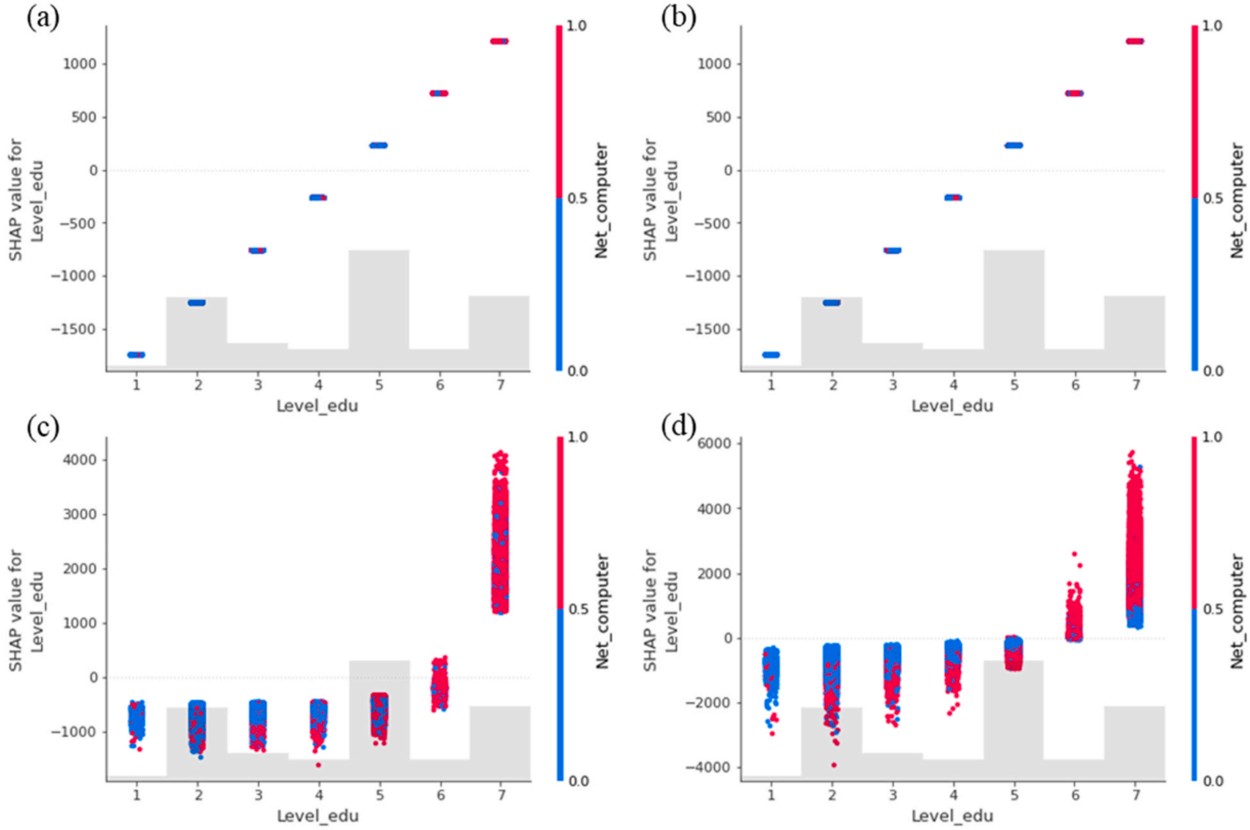

Fig. 2. SHAP dependence plot of level of education and the interaction with internet access via computer. (a) Linear regression. (b) Elastic net. (c) XGBoost. (d) Neural network.

Three of the most important variables concern the devices that people use to access the internet, i.e., whether the person accessed the internet using a computer, a tablet, or a smart TV. All models revealed a positive relationship between the use of these devices and income. As highlighted by Silva et al. (2020), this association is expected, as the increase in income is likely to lead to an increase in spending on less essential goods, such as smart TVs, tablets, and electronic devices. At the same time, this demonstrated relationship may also indicate that people with more exposure to ICTs and more digital literacy are expected to receive higher incomes.

Moreover, ML models seem to capture a more complex relationship. Fig. 2 shows how the model output (income prediction), represented on the left y axis, changes according to the level of education variable on the x axis. In addition, the interaction with the internet access via computer variable is demonstrated by the colour of each observation point as per the right Y axis. The linear regression and elastic net models capture a strict linear association and attribute an equal increase in the model output for each additional level of education.

However, the machine learning algorithms present a much more detailed relationship and show that level of education has a slight impact on income up to a certain threshold, when education starts to have a significant impact on the model output. The results also reveal the positive association between internet access via computer and the model output. This demonstrates one of the complex existing relationships and enhances our understanding around the issue analysed in this study, which relates to question (ii) mentioned previously.

Despite showing less influence, the variable related to whether the person accessed the internet via a 3G/4G connection presents a positive correlation, which means that lower incomes are associated with the absence of a 3G/4G network. Cheng et al. (2021) also report a positive impact of smartphone internet access, measured by the number of mobile subscriptions, on economic growth and on reducing income inequality. According to the results, the broadband variable presents a very low contribution. A possible reason for this effect is a peculiar practice that occurs in Brazil, where a considerable part of the population shares broadband subscriptions with neighbours to reduce costs (OECD, 2020a).

Additionally, there seems to be a positive relationship between the level of education and the digital engagement of individuals. As shown in Fig. 3, the number and difficulty of activities performed online differ significantly across educational levels. While most people only use the internet for messages and calls, individuals with higher levels of education take more advantage of the benefits of the digital world, for example, using email. The variable regarding whether the person used the internet to send or receive email is selected as the eighth most significant variable by the neural network algorithm, our best performing model, and seems to be positively correlated with educational level.

The OECD (2020a) stresses that education is the most influential factor impacting internet use in Brazil, and the severe inequality in the country might increase social exclusion. A recent government program implemented in Brazil, the “Brazil More Digital”, aims to provide IT training for young people aged 16 to 25 to increase the population’s computing skills and meet the demand for professionals in fields requiring these skills. However, the program has experienced high dropout rates, suggesting there is room for improvement in content, structure, and overall quality. The OECD (2020a) also notes that affordability is the main reason for Brazilians not having internet connection at home, and the government has been trying to promote digital inclusion by expanding free public internet access services. However, at the same time, it is also necessary to increase the digital literacy of the population.

It is also important to highlight the negative relationship between the Northeast region variable and income, which is considered one of the top ten most influential variables by the two ML models. This result is expected since the states in the Northeast region of Brazil have the lowest GDP per capita and account for 51% of the country’s population living in poverty (Beghin, 2008; Herrera et al., 2017). However, it seems that the neural network and XGBoost methods manage to model this relationship more easily and assign greater significance to this variable compared to the linear regression and elastic net models.

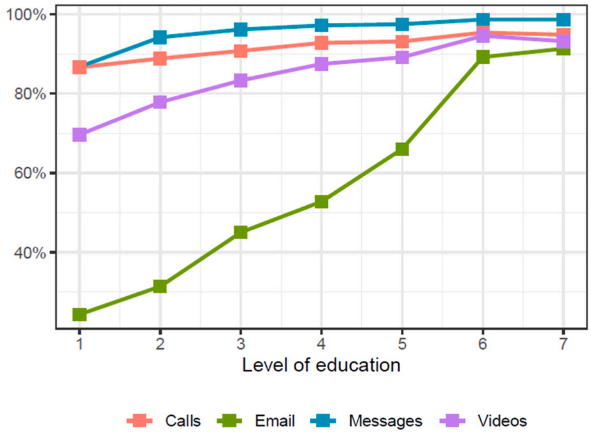

Fig. 3. Percentage of individuals performing digital activities for each level of education.

4.1. The role of education level, age, and gender

The analysis conducted using the four different methods presented a unanimous result regarding the three most important variables and their order of significance. The impacts of education, age, and gender on income are well documented in the literature. For example, Signor et al. (2019) and Silva et al. (2020) report a positive association between education and higher income, while economic theory holds that age tends to follow an inverted U-shaped impact on income, as described by Aldy and Viscusi (2008).

Our analysis suggests that the relationships among these variables are complex and exhibit nonlinear patterns that are not accounted for by the linear models. As shown in Fig. 4, the machine learning models are capable of capturing a much more detailed association and responding to the question (i) raised at the beginning of the study concerning the relationships between socioeconomic characteristics and income distribution. According to the results, up to a certain threshold, approximately 40 years old, individuals with lower education levels tend to earn more. However, this trend starts to change past the threshold, and people with higher levels of education are more likely to receive a higher income. A possible reason for this is the fact that teenagers and young adults seeking higher education tend to commit most of their time to studies in the early stages of life and less time to work. On the other hand, others prefer to spend less time studying and are more dedicated to work, but their jobs are usually less knowledge intensive and at the lower-middle salary level. After approximately 40 years of age, when most adults have already completed their degrees, those with higher levels of education are more likely to experience a faster income increase and earn the highest salaries. Those with low levels of education tend to slowly increase their income along with their experience, but are far behind individuals with a high level of education.

As stated by the OECD (2020b), the development of new technologies, the increased use of ICTs, and the structural changes imposed by the COVID-19 pandemic will profoundly impact the job market and require workers to adapt and develop new skills. In this sense, our results highlight the importance of policies such as upskilling and reskilling programs that provide opportunities for adults to learn ICT skills, acquire digital competencies, and stay updated with emerging technologies in order to facilitate employability, career advancement and prevent the digital divide. Matsubayashi et al. (2019) stress that ICTs have been playing a vital role in promoting education in Brazil, especially distance education. The authors highlight that it is necessary to include a new set of skills in the curriculum to prepare the new generation for the jobs of the future.

Although Brazil spends approximately 6% of GDP on education, which is more than the OECD average, the country often performs poorly on the OECD's Programme for International Student Assessment (PISA) tests. Other Latin American countries that spend a

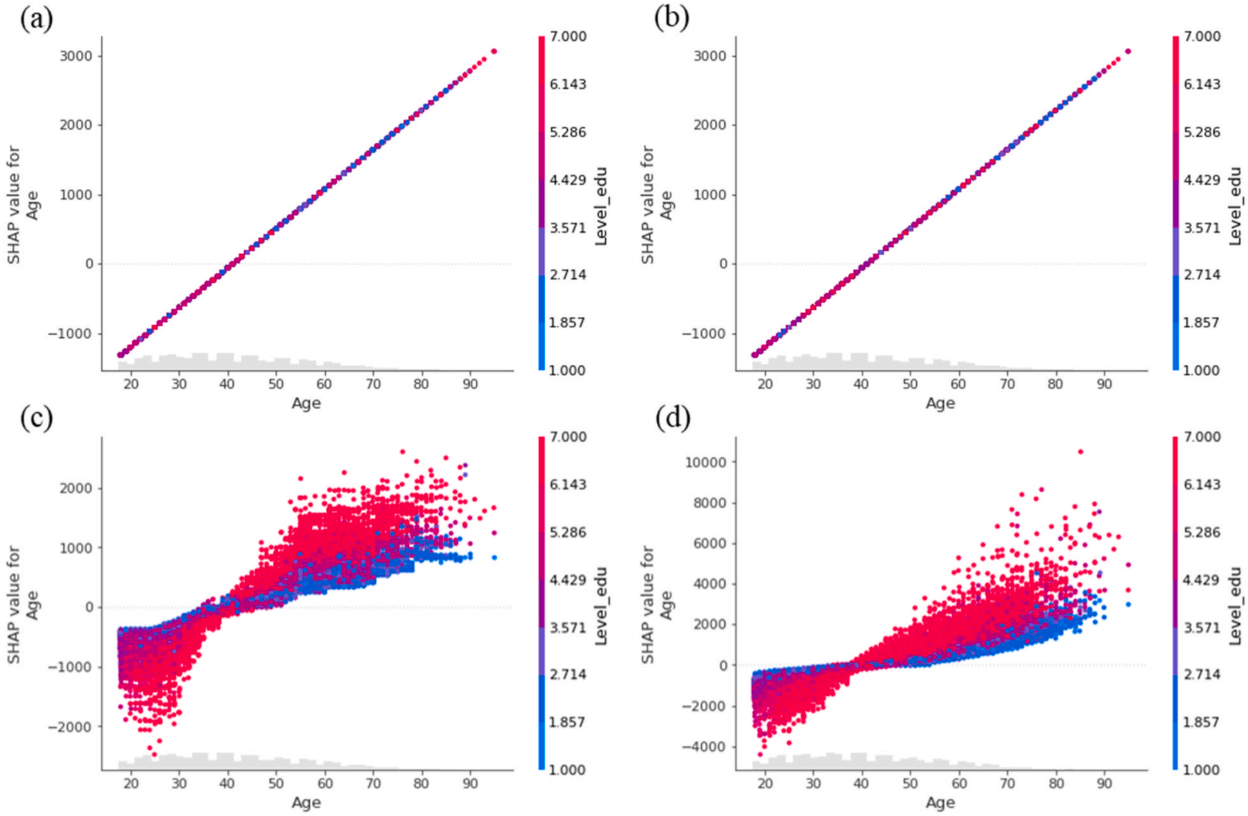

Fig. 4. SHAP dependence plot of age and the interaction with the level of education. (a) Linear regression. (b) Elastic net. (c) XGBoost. (d) Neural network.

smaller fraction of their GDP, such as Mexico and Uruguay, have scored higher in the past. The challenge in Brazil seems to be related to the more efficient use of financial resources and better public management.

The improvement of education and the development of ICTs are intertwined and work together to promote social and economic progress. According to Hidalgo et al. (2020), age and education are among the main socioeconomic factors that impact the population digital skills level. Further, it is clear that the development and implementation of ICTs have a profound impact on the job market. Digitalization and the introduction of new technologies require workers to develop a new set of skills in order not to risk losing their jobs. At the same time, the automation of processes and the mechanization of tasks reduce the number of vacancies available. Nevertheless, Venturini (2022) argues that, simultaneously, new non-routinised tasks are created and the negative impact of automation on the job market only exists where there is a gap between the level of skills required by new technologies and the skills of the workforce.

In Brazil, where the digital divide and lack of digital skills are prevalent, it is essential that public policies focus on bridging these gaps and empowering individuals to adapt and secure employment in the digital era. Education plays a critical role in addressing the challenges posed by the rapid development of ICTs in the job market and the consequent impact on individual income. The integration of ICTs into education has the potential to significantly increase individuals' income levels (as will be further demonstrated in section 4.2). This can be accomplished by introducing public policies that seek to integrate ICT-related subjects into the curriculum at all levels of education. At the same time, to maximize effectiveness and inclusive socioeconomic development, policymakers need to ensure quality and equal access across all social classes. In relation to question (iv) presented previously, the combined analysis using SHAP values and machine learning presents a valuable alternative for tailored policy-making in the context of socioeconomic inequality and ICT usage.

Finally, our analysis shows a well-known but extremely concerning problem: the income gap between male and female individuals. As highlighted by the OECD (2020b), in Brazil, female employees receive, on average, 20% less pay than men. The gender gap is demonstrated in Fig. 5. For individuals with the same high levels of education (6 – degree incomplete and 7 – degree complete) the model learns to predict a higher income for males and a lower income for females. This relationship is much clearer through the ML models, which again manage to capture this detailed information, contrary to the other two approaches.

The fourth industrial revolution, characterized by the automation and digitalization of processes, has the potential to significantly increase economic output, which is augmented by globalization and greater financial connectivity between markets. However, gender inequality is a persistent social problem documented in several countries. Danquah et al. (2021) highlight the need for public policies

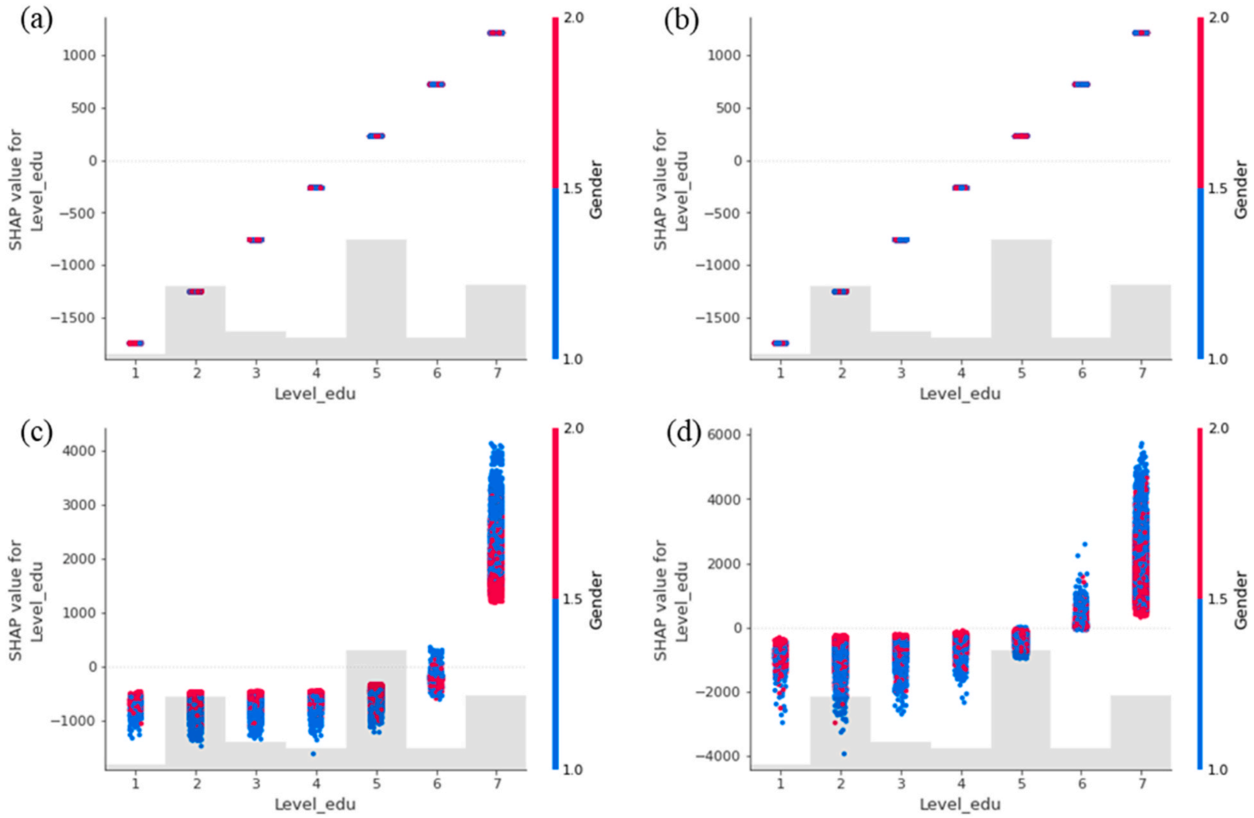

Fig. 5. SHAP dependence plot of level of education and the interaction with gender. (a) Linear regression. (b) Elastic net. (c) XGBoost. (d) Neural network.

that ensure preschools and early childhood education services to empower women and improve household welfare. In Brazil, the number of preschool venues has increased significantly in recent years, but it has had a small social impact, as most of these venues are in high-income neighbourhoods (OECD, 2020b).

Further, gender inequality also has a negative impact on the creation and dissemination of knowledge, as women are underrepresented in many important research fields globally. Women are central to the development of knowledge economy theory and gender bias, as well as the pay gap and lack of support and opportunities for women, hinder economic development and societal growth. The gender gap is a social problem that requires proper attention and action. Our results demonstrate that, thus far, public policies in Brazil have been ineffective and inadequate to solve the gender inequality problem in the country. In this regard, digitalization and ICTs present a valuable alternative that can be incorporated into public policies to help fight this matter. For instance, the government could offer training programs specifically targeting women using online learning platforms and digital resources, enabling women to acquire knowledge and skills regardless of their geographical location or personal circumstances.

4.2. *Variable interaction robustness test*

As shown in the previous section, many of the relationships that exist in the dataset are complex and follow a nonlinear pattern that is difficult to capture using traditional linear models. In particular, there seems to be a relationship between the ICT variables and the level of education that contributes to predicting people's income. As highlighted by Bauer (2018), ICTs usually interact with other social, economic, and political variables that define the impact it has on income. We further test this association and interactions between predictor variables. First, we predict the target variable, i.e., total income, using the ICT-related variables presented in Panel B of Table 1. Second, we perform the analyses using only the level of education variable to predict the dependent variable. Finally, we compute the results for all models employing the ICT-related and the level of education variables to predict the total income of individuals. Performance is measured following the same cross-validation process described at the beginning of Section 4 above, and the results are presented in Table 5.

The results respond to our question (iii) and confirm that there is an interaction between education and ICT variables that helps to predict the total income of individuals and that this interaction presents complex patterns such as nonlinearities that are better modelled by the two machine learning methods. As demonstrated throughout this study, the two traditional regression models are only capable of capturing simplistic relationships, and these models do not fully account for such more complex interactions between variables. The linear regression and elastic net models present the lowest RMSE in Model (3), while the lowest MAE values are found for Model (1). On the other hand, the XGBoost and neural network ML algorithms achieved the best performance according to both loss functions using Model (3). These results further support the application of ML models as tools to uncover complex nonlinearities and high-order interactions among covariates.

Lastly, inevitably this work has some limitations. SHAP values provide insights on associations between variables, comparable to linear regression models, and should be viewed as correlations rather than causal relationships. Further, due to the confidentiality of microdata from IBGE, most of the variables employed are binary. Additionally, the four methods employed in this study only represent a subset of available analytical techniques and further research is necessary to extend the demonstrated analysis to a wider range of models. We also highlight limitations related to simplified assumptions in the SHAP methodology and algorithmic bias and training data bias associated with machine learning models as discussed previously.

5. Conclusions

Machine learning algorithms are powerful data-driven models known for their improved prediction performance and ability to handle complex data. Nonetheless, these models have always been criticized for their lack of transparency and interpretability. Explainable AI is a novel research area dedicated to developing interpretable ML models, building trust with users, and reducing bias. In this study, we demonstrate how ML methods can be interpreted and applied to reveal insights about a socioeconomic issue. We apply two popular machine learning algorithms, a tree-based approach, i.e., XGBoost, and a feedforward neural network, as well as benchmark models.

We show how ML methods can be interpreted, similar to traditional econometric approaches, using the SHAP values methodology. This allows us to reveal robust insights, as ML models are able to uncover nonlinearities and complex interactions in the data, thus providing a more detailed picture and comprehensive information about the relationship between variables. Our analyses suggest that

Table 5
Interaction between ICT-related variables and level of education to predict the target variable.

	Linear regression		Elastic net		XGBoost		Neural network	
	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
(1) ICTs	3352.34	1647.27	3352.32	1646.25	3334.86	1618.58	3331.65	1603.64
(2) Education	3419.95	1748.85	3419.95	1748.68	3309.96	1595.91	3355.54	1654.91
(3) ICTs + Education	3296.69	1655.42	3296.71	1654.79	3177.87	1503.17	3174.82	1491.11

Note: This table presents the forecasting performance of each model using only the ICTs related variables, then only the level of education variable, and a final model with both sets of variables. We estimate the models using a k-fold cross-validation procedure.

there is a positive association between internet access devices and income, which is expected, as higher income leads to more expenditure on less essential items such as electronic devices. However, the observed correlation may also suggest that people with greater exposure to ICTs, and thus higher levels of digital literacy, are more likely to earn higher incomes.

Additionally, there seems to be a positive influence of education on the number and complexity of tasks people perform online. Our robustness analysis confirms that there is an interaction between individuals' educational level and the ICT-related variables that contribute to predicting income. Thus, this study provides important insights to support bespoke public policy formulation in Brazil. Our empirical results suggest that combining ICTs and education can potentialize a positive impact on individual's income. This can be achieved by, for example, incorporating ICT-related subjects into the curriculum at all levels of education, from primary schools to higher education institutions. Policymakers need to ensure the quality of this training and equal access to resources by expanding broadband infrastructure, providing subsidies to low-income households, and establishing community centres with digital equipment.

Education plays a critical role in the diffusion of innovations theory, as it enables individuals to gain the knowledge and skills necessary to adopt and use new technologies. Through education, people can learn about the benefits of innovations and how they can be applied in their personal and professional lives.

Our results also highlight the income gap and education disparity between men and women, which is a persistent problem in Brazil. Past public policies have not been successful in solving this issue and new policies could benefit from insights presented in our analyses. Alternatives include public policies targeting partnerships with educational institutions, private sector organizations, and start-ups to provide scholarships and mentorship exclusively for women, as well as flexible online training programs specifically targeting this group.

Public policies that aim to strengthen and develop digital technologies are key to ensuring that the diffusion of ICTs will have a positive impact on society. Failure to do so might result in increasing social inequality and exclusion and deepening poverty. Our study has some limitations. We emphasize that SHAP values provide an interpretation approach based on associations, comparable to a linear regression model, and do not infer causality. Furthermore, the analyses presented are not immune to omitted variable bias and are limited by the large number of binary variables provided by the IBGE due to confidentiality of microdata. Finally, this study only employs four analysis methods and additional research is necessary to extend the demonstrated analysis to a wider range of models, such as nonlinear regression models, Support Vector Machines, and hybrid ML models. This study represents a step forward in the development of interpretable machine learning models to explore complex relationships, allowing the formulation of tailored public policies.

Declaration of competing interest

None.

Appendix

Table A1
Variables coefficients for the linear regression and elastic net models

Variables	Linear regression	Elastic net
Const	−3889.650***	–
North	−66.347***	−61.257
Northeast	−305.717***	−303.413
Midwest	313.552***	302.368
South	52.096***	45.795
Location	102.562***	92.044
Gender	−1015.293***	−1007.089
Age	56.405***	56.496
Net_computer	456.348***	453.283
Net_tablet	1653.394***	1647.076
Net_mobile	310.590***	155.265
Net_tv	612.115***	610.285
Net_other	1404.122***	1296.326
3G_4G	225.399***	219.953
Dial_up	−176.231	0.000
Broadband	90.860***	83.053
Email	278.977***	275.959
Messages	30.110	13.906
Calls	17.705	7.184
Videos	−27.998	−11.544
Mobile_net	157.523***	121.714
BPC_LOAS	1133.234***	974.414
Bolsa_Fa	818.315***	770.263
Other_social	572.781***	191.334

(continued on next page)

Table A1 (*continued*)

Variables	Linear regression	Elastic net
Gov_pension	1443.591***	1385.705
Work_insurance	339.857***	216.937
Donation	343.180***	291.931
Rent_lease	2346.723***	2312.752
Other_income	3553.671***	3496.953
Level_edu	494.557***	494.718
Workforce	1550.422***	1492.707

Note: ***p-value<0.01; **p-value<0.05; *p-value<0.1.

References

Albiman, M. M., & Sulong, Z. (2017). The linear and non-linear impacts of ICT on economic growth, of disaggregate income groups within SSA region. *Telecommunications Policy, 41*(7–8), 555–572. https://doi.org/10.1016/j.telpol.2017.07.007

Aldy, J. E., & Viscusi, W. K. (2008). Adjusting the value of a statistical life for age and cohort effects. *The Review of Economics and Statistics, 90*(3), 573–581. https://doi.org/10.1162/rest.90.3.573

Antulov-Fantulin, N., Lagravinese, R., & Resce, G. (2021). Predicting bankruptcy of local government: A machine learning approach. *Journal of Economic Behavior & Organization, 183*, 681–699. https://doi.org/10.1016/j.jebo.2021.01.014

Antwarg, L., Miller, R. M., Shapira, B., & Rokach, L. (2021). Explaining anomalies detected by autoencoders using Shapley Additive Explanations. *Expert Systems with Applications, 186*, Article 115736. https://doi.org/10.1016/j.eswa.2021.115736

Ashraf Ganjoei, R., Akbarifard, H., Mashinchi, M., & Jalaee Esfandabadi, S. A. M. (2021). Applying of fuzzy nonlinear regression to investigate the effect of information and communication technology (ICT) on income distribution. *Mathematical Problems in Engineering, 2021*. https://doi.org/10.1155/2021/5545213

Athey, S., & Imbens, G. W. (2019). Machine learning methods that economists should know about. *Annual Review of Economics, 11*. https://doi.org/10.1146/annurev-economics-080217-053433

Ballester, P. L., Cardoso, T. D. A., Moreira, F. P., da Silva, R. A., Mondin, T. C., Araujo, R. M., … de Mattos Souza, L. D. (2021). 5-year incidence of suicide-risk in youth: A gradient tree boosting and SHAP study. *Journal of Affective Disorders, 295*, 1049–1056. https://doi.org/10.1016/j.jad.2021.08.033

Barakabitze, A. A., William-Andey Lazaro, A., Ainea, N., Mkwizu, M. H., Maziku, H., Matofali, A. X., Iddi, A., & Sanga, C. (2019). *Transforming african education systems in science, technology, engineering, and mathematics (STEM) using ICTs: Challenges and opportunities* (Vol. 2019). Education Research International. https://doi.org/10.1155/2019/6946809

Bauer, J. M. (2018). The Internet and income inequality: Socio-economic challenges in a hyperconnected society. *Telecommunications Policy, 42*(4), 333–343. https://doi.org/10.1016/j.telpol.2017.05.009

Beghin, N. (2008). Notes on inequality and poverty in Brazil: Current situation and challenges. *From Poverty to Power: How Active Citizens and Effective States Can Change the World*.

Booth, A. L., Abels, E., & McCaffrey, P. (2021). Development of a prognostic model for mortality in COVID-19 infection using machine learning. *Modern Pathology, 34* (3), 522–531. https://doi.org/10.1038/s41379-020-00700-x

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science, 16*(3), 199–231.

Briglauer, W., & Gugler, K. P. (2017). Go for gigabit? First evidence on economic benefits of (ultra-) fast broadband technologies in Europe. *First Evidence on Economic Benefits of (Ultra-) Fast Broadband Technologies in Europe*. https://doi.org/10.2139/ssrn.3006513. July 21, 2017).

Carbo-Valverde, S., Cuadros-Solas, P., & Rodríguez-Fernández, F. (2020). A machine learning approach to the digitalization of bank customers: Evidence from random and causal forests. *PLoS One, 15*(10), Article e0240362. https://doi.org/10.1371/journal.pone.0240362

Chen, J., de Hoogh, K., Gulliver, J., Hoffmann, B., Hertel, O., Ketzel, M., … Hoek, G. (2019). A comparison of linear regression, regularization, and machine learning algorithms to develop Europe-wide spatial models of fine particles and nitrogen dioxide. *Environment International, 130*, Article 104934. https://doi.org/10.1016/j.envint.2019.104934

Cheng, C. Y., Chien, M. S., & Lee, C. C. (2021). ICT diffusion, financial development, and economic growth: An international cross-country analysis. *Economic Modelling, 94*, 662–671. https://doi.org/10.1016/j.econmod.2020.02.008

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. August. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794). https://doi.org/10.1145/2939672.2939785.

Chien, M. S., Cheng, C. Y., & Kurniawati, M. A. (2020). The non-linear relationship between ICT diffusion and financial development. *Telecommunications Policy, 44*(9), Article 102023. https://doi.org/10.1016/j.telpol.2020.102023

Chollet, F. (2015). *Keras*. https://keras.io.

Cooray, U., Watt, R. G., Tsakos, G., Heilmann, A., Hariyama, M., Yamamoto, T., … Aida, J. (2021). Importance of socioeconomic factors in predicting tooth loss among older adults in Japan: Evidence from a machine learning analysis. *Social Science & Medicine, 291*, Article 114486. https://doi.org/10.1016/j.socscimed.2021.114486

Danquah, M., Iddrisu, A. M., Boakye, E. O., & Owusu, S. (2021). Do gender wage differences within households influence women's empowerment and welfare? Evidence from Ghana. *Journal of Economic Behavior & Organization, 188*, 916–932. https://doi.org/10.1016/j.jebo.2021.06.014

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

Fruttero, A., Leichsenring, A. R., & Paiva, L. H. (2020). *Social programs and formal employment: Evidence from the Brazilian bolsa família program*. International Monetary Fund.

Giudici, P., & Raffinetti, E. (2021). Shapley-Lorenz eXplainable artificial intelligence. *Expert Systems with Applications, 167*, Article 114104. https://doi.org/10.1016/j.eswa.2020.114104

Gordon, R. J. (2003). Exploding productivity growth: Context, causes, and implications. *Brookings Papers on Economic Activity, 2003*(2), 207–298. https://doi.org/10.1353/eca.2004.0006

Gu, J., Yang, B., Brauer, M., & Zhang, K. M. (2021). Enhancing the evaluation and interpretability of data-driven air quality models. *Atmospheric Environment, 246*, Article 118125. https://doi.org/10.1016/j.atmosenv.2020.118125

Harvey, D., Leybourne, S., & Newbold, P. (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting, 13*, 281–291. https://doi.org/10.1016/S0169-2070(96)00719-4

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2, pp. 1–758). New York: springer. https://doi.org/10.1007/978-0-387-84858-7

Herrera, G. P., da Costa, R. B., de Moraes, P. M., Mendes, D. R. F., & Constantino, M. (2017). Smallholder farming in Brazil: An overview for 2014. *African Journal of Agricultural Research, 12*(17), 1424–1429. https://doi.org/10.5897/AJAR2017. 12137

Hidalgo, A., Gabaly, S., Morales-Alonso, G., & Urueña, A. (2020). The digital divide in light of sustainable development: An approach through advanced machine learning techniques. *Technological Forecasting and Social Change, 150*, Article 119754. https://doi.org/10.1016/j.techfore.2019.119754

Hyndman, R. J., & Athanasopoulos, G. (2018). *Forecasting: Principles and practice*. OTexts.

IBGE. (2021). *Panorama [outlook]*. Rio de Janeiro, RJ, Brazil: Instituto Brasileiro de Geografia e Estatística [Brazilian Institute of Geography and Statistics. https://cidades.ibge.gov.br/brasil/panorama.

IMF. (2021). *International monetary fund world economic outlook: Recovery during a pandemic—health concerns, supply disruptions, price pressures*. Washington, DC: October.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (7th ed.). New York: Springer.

Johannessen, J. A., & Olsen, B. (2010). The future of value creation and innovations: Aspects of a theory of value creation and innovation in a global knowledge economy. *International Journal of Information Management, 30*(6), 502–511. https://doi.org/10.1016/j.ijinfomgt.2010.03.007

John-Mathews, J. M. (2022). Some critical and ethical perspectives on the empirical turn of AI interpretability. *Technological Forecasting and Social Change, 174*, Article 121209. https://doi.org/10.1016/j.techfore.2021.121209

Kingma, D. P., & Ba, J. (2014). *Adam: A method for stochastic optimization*. arXiv preprint arXiv:1412.6980.

Lantz, B. (2013). *Machine learning with R Learn how to use R to apply powerful machine learning methods and gain an insight into real-world applications*. Packt Publ.

Li, R., Shinde, A., Liu, A., Glaser, S., Lyou, Y., Yuh, B., … Amini, A. (2020). Machine learning–based interpretation and visualization of nonlinear interactions in prostate cancer survival. *JCO Clinical Cancer Informatics, 4*, 637–646. https://doi.org/10.1200/CCI.20.00002

Llorent-Vaquero, M., Tallon-Rosales, S., & de las Heras Monastero, B. (2020). Use of information and communication technologies (ICTs) in communication and collaboration: A comparative study between university students from Spain and Italy. *Sustainability, 12*(10), 3969. https://doi.org/10.3390/su12103969

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. December. In *Proceedings of the 31st international conference on neural information processing systems* (pp. 4768–4777).

Malaquias, F., Jacobi, L. A. D. S., & Lopes, J. E. F. (2021). *Antecedents and outcomes of ICTs adoption by mompreneurs: Empirical evidence from Brazil*. Information Development. https://doi.org/10.1177/02666669211047925

Matsubayashi, M. O., Oikawa, R. T., Alves, V. L. P., Popova, N. D., & Silva, G. B. F. (2019). *Insights about digital transformation and ICT opportunities for Brazil: Report and recommendations*. London, United Kingdom: Deloitte Touche Tohmatsu Limited.

Njoh, A. J. (2018). The relationship between modern information and communications technologies (ICTs) and development in Africa. *Utilities Policy, 50*, 83–90. https://doi.org/10.1016/j.jup.2017.10.005

Novak, J., Purta, M., Marciniak, T., Ignatowicz, K., Rozenbaum, K., & Yearwood, K. (2018). *The rise of Digital Challengers: How digitization can become the next growth engine for Central and Eastern Europe*. McKinsey Global Institute.

OECD. (2017). *OECD digital economy outlook 2017*. Paris: OECD Publishing. https://doi.org/10.1787/9789264276284-en

OECD. (2020a). *Going digital in Brazil, OECD reviews of digital transformation*. Paris: OECD Publishing. https://doi.org/10.1787/e9bf7f8a-en

OECD. (2020b). *OECD economic surveys: Brazil 2020*. Paris: OECD Publishing. https://doi.org/10.1787/250240ad-en

Parsa, A. B., Movahedi, A., Taghipour, H., Derrible, S., & Mohammadian, A. K. (2020). Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. *Accident Analysis & Prevention, 136*, Article 105405. https://doi.org/10.1016/j.aap.2019.105405

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research, 12*, 2825–2830.

Richmond, K., & Triplett, R. E. (2018). ICT and income inequality: A cross-national perspective. *International Review of Applied Economics, 32*(2), 195–214. https://doi.org/10.1080/02692171.2017.1338677

Rico-Juan, J. R., & de La Paz, P. T. (2021). Machine learning with explainability or spatial hedonics tools? An analysis of the asking prices in the housing market in alicante, Spain. *Expert Systems with Applications, 171*, Article 114590. https://doi.org/10.1016/j.eswa.2021.114590

Rogers, Everett M. (1962). *Diffusion of innovations* (1st ed.). New York: Free Press of Glencoe. OCLC 254636.

Saidi, K., & Mongi, C. (2018). The effect of education, R&D and ICT on economic growth in high income countries. *Economics Bulletin, 38*(2), 810–825.

Samarasinghe, S. (2016). *Neural networks for applied sciences and engineering: From fundamentals to complex pattern recognition*. Crc Press.

Shapley, L. S. (1953). A value for n-person games. In *Contributions to the theory of games* (pp. 307–317).

Signor, D., Kim, J., & Tebaldi, E. (2019). Persistence and determinants of income inequality: The Brazilian case. *Review of Development Economics, 23*(4), 1748–1767. https://doi.org/10.1111/rode.12598

Silva, T. C., Coelho, F. C., Ehrl, P., & Tabak, B. M. (2020). Internet access in recessionary periods: The case of Brazil. *Physica A: Statistical Mechanics and Its Applications, 537*, Article 122777. https://doi.org/10.1016/j.physa.2019.122777

Thongpeth, W., Lim, A., Wongpairin, A., Thongpeth, T., & Chaimontree, S. (2021). Comparison of linear, penalized linear and machine learning models predicting hospital visit costs from chronic disease in Thailand. *Informatics in Medicine Unlocked, 26*, Article 100769. https://doi.org/10.1016/j.imu.2021.100769

Van Rossum, G., & Drake, F. L. (2009). *Python 3 reference manual*. CA: Scotts Valley.

Venturini, F. (2022). Intelligent technologies and productivity spillovers: Evidence from the fourth industrial revolution. *Journal of Economic Behavior & Organization, 194*, 220–243. https://doi.org/10.1016/j.jebo.2021.12.018

Vu, K. M. (2011). ICT as a source of economic growth in the information age: Empirical evidence from the 1996–2005 period. *Telecommunications Policy, 35*(4), 357–372. https://doi.org/10.1016/j.telpol.2011.02.008

World Bank. (2022). *Global economic prospects, january 2022*. Washington, DC: World Bank. https://doi.org/10.1596/978-1-4648-1758-8

WTO. (2021). *World trade organization - world trade statistical review 2021* (Switzerland).

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B, 67*(2), 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x