The Business of Being Formal: Advancing Financial Inclusion for Colombian Microbusinesses

July 31, 2025

MIT: Sabal Ranabhat, Aziz Malouche

UNDP: Santiago Plata Diaz, Diana Gonzalez

Advisor: Yanchong Karen Zheng

Contents

1 Executive Summary

This paper presents a comprehensive analysis of the microbusiness landscape in Colombia, using four indices, Formality, Financial Formality, Digital Readiness, and Business Sophistication, derived from publicly available data in the national household and business surveys (GEIH and EMICRON). Using clustering techniques, Multiple Correspondence Analysis, and Structural Equation Modeling (SEM), we identify the strongest predictors of formality and financial inclusion, and propose actionable, data-driven policy recommendations.

Key findings reveal that Business Sophistication, comprising having a business name, keeping basic accounts, having paid workers, and operating a website, is a powerful predictor of formality. Encouraging microbusinesses to adopt simple practices, such as naming their business and doing basic bookkeeping, offers a low-barrier starting point to formalization. These practices not only drive formalization but are also highly correlated with more productive loan use. Additionally, despite a 94% loan approval rate in 2023, only 18% of microbusinesses apply for loans, primarily due to fear of debt, especially among poorer segments. Expanding access to credit should focus on increasing applications, not approvals. A targeted recommendation is to scale up Colombia's Banca de las Oportunidades, a government program that could be strengthened to deliver tailored financial education and accessible savings products, especially for poor and vulnerable microbusinesses. This also addresses the large gap in formal saving behavior (15% of poor vs. 73% of high-income microbusinesses).

On the digital front, Digital Readiness, a key driver of financial formality, remains limited among microbusinesses. Policies should expand low-cost internet access, subsidize digital devices, and integrate digital literacy training into entrepreneurship programs. A promising model is the Ministry of Education's 2024 initiative to distribute over 57,000 computer terminals to students. A similar program targeting microbusinesses could bridge the digital divide and enhance their ability to access online financial tools and services. Ultimately, this paper argues that gradual, behaviorally informed steps toward formality and financial inclusion, starting with business sophistication and supported by digital readiness, offer an accessible and sustainable pathway to formalization.

To support ongoing analysis and policy design, we also introduce an interactive dashboard that serves as a "living lab" for financial inclusion. Built on a data pipeline integrating GEIH and EMICRON data from 2019 to 2023, the dashboard can be updated annually. It allows users to explore key microbusiness metrics, such as income, expenses, and the four index scores, by region, sector, socioeconomic group, and time. This tool enables policymakers and stakeholders to benchmark progress, identify gaps, and tailor interventions based on real-time, disaggregated data.

2 Introduction

The United Nations Development Program (UNDP), headquartered in New York City, is the United Nations' leading agency for international development. Operating in 177 countries, its core mission is to promote sustainable economic growth and human development. Since 1991, the UNDP has published the Human Development Report, which includes the Human Development Index (HDI)—a composite measure that ranks countries based on achievements in three key areas: longevity and health, education, and standard of living [1].

In Colombia, the UNDP is focusing on the critical role of microbusinesses in driving human development. These enterprises account for 99% of the country's production ecosystem, with most employing fewer than 10 workers. Despite their significance to the economy, approximately 87% of the 5.2 million microbusinesses in Colombia operate informally. This widespread informality limits access to social protections, public support programs, and financial services. Consequently, many microbusiness owners remain impoverished and lack pathways to improve their economic standing. A particularly pressing issue is access to credit. More than 70% of small shop owners are unable to obtain formal credit due to inadequate financial records. Many neighborhood stores—central hubs in local economies—rely instead on informal financial networks. While these networks provide short-term liquidity, they ultimately constrain growth and undermine financial stability. In the absence of formal financial inclusion and institutional support, some microbusiness owners are even driven toward illicit networks to sustain or expand their operations.

Financial inclusion, particularly in middle-income countries, has proven to positively influence the formalization of economic activities. Recognizing this, the UNDP is developing strategies to "score the invisible"—those who lack formal credit histories or traditional credit scores—in order to enable more inclusive financial systems. By enhancing access to credit and support services, the UNDP aims to improve the financial resilience and growth potential of microbusinesses. To this end, the UNDP is working to develop a deeper understanding of the regional dynamics that shape microbusiness success, including demographics, security, and economic conditions. This localized analysis will help identify tailored strategies to promote financial inclusion, support microbusinesses in building operational capacity, and help them integrate into the formal economy and broader financial markets.

To ensure that development efforts reach those most in need, the UNDP places a strong focus on poor and vulnerable populations. These groups—often excluded from formal employment, financial systems, and social protections—face persistent barriers to improving their living conditions. In Colombia, around 55% of microbusiness owners fall within this category. Many operate informally, lack access to credit or public services, and remain economically marginalized despite their contributions to local economies. Addressing the unique challenges faced by these populations is essential to achieving inclusive and equitable development.

One existing public policy intended to support poor and vulnerable populations in Colombia, for instance, is the socioeconomic stratification system. This system categorizes neighborhoods into six strata (from 1 to 6) based not on individual income but on the physical and urban characteristics of their housing. Its original purpose was to enable the government to target utility subsidies more effectively—offering financial support to lower-income households (strata 1, 2, and 3) while charging full rates or surcharges to wealthier households (strata 4, 5, and 6). However, the stratification system has remained largely unchanged since its introduction nearly 40 years ago. Once a home is assigned a stratum, that classification rarely changes, regardless of changes in occupancy or household economic conditions. As a result, misclassifications are common, with some low-income individuals living in higher-stratum neighborhoods and having to pay higher rates, while some high-income individuals live in lower-stratum neighborhoods and receive utility subsidies. In many cities, these divisions have also evolved into physical and symbolic boundaries that reinforce social segregation and limit social and economic mobility [2].

In response to these challenges, the UNDP is pursuing a more dynamic and inclusive approach to identify and support vulnerable populations, especially microbusiness owners. By rethinking outdated classification systems and expanding access to financial tools, the UNDP aims to increase financial inclusion and encourage formalization of microbusinesses. These efforts are central to accelerating human development and reducing inequality across Colombia.

3 Problem Statement

In effort to achieve this broader goal, this project has mainly two objectives:

- The first objective is to understand the structure of the microbusiness landscape in Colombia, identifying key patterns and drivers behind informality and financial exclusion. By analyzing nationally representative household and business surveys, the goal is to find the characteristics of businesses that remain informal or excluded, as well as those that have achieved high levels of formality, financial inclusion, or success.
- The second objective is to develop quantifiable indices to summarize complex business traits and guide evidence-based policymaking. These indices are designed to simplify targeting, track progress over time, and identify where interventions may be most impactful. The final outcome will be concrete, data-driven policy recommendations to promote business formalization and financial inclusion, ultimately enabling more microbusinesses to access growth opportunities and contribute more effectively to Colombia's economy.

4 Data Sources

Dataset	Primary Use	Key Features
GEIH (Gran Encuesta Integrada de Hogares)	Household and labor profiling of microbusiness operators	Demographics, labor activity, household income, urban/rural status, access to technology and social programs
EMICRON (Microbusiness Survey)	Core dataset for analyzing microbusiness characteristics and behavior	Business operations (sector, financing, digital presence), owner motivations, financial behavior, marketing, formal practices
Neighborhood Stores Survey	Micro-level behavioral and operational insights for small urban retailers	Sales patterns, customer behavior, technology use, supply chain, business planning, owner characteristics, location context

Table 1: Summary of Data Sources

We draw from three main data sources to analyze the landscape of microbusinesses in Colombia and the characteristics of their operators. The Large Integrated Household Survey (GEIH) and the Microbusiness survey (EMICRON), maintained by the National Administrative Department of Statistics of Colombia (DANE), offer a macro level understanding of different microbusinesses and their operators. The Neighborhood Stores Survey, assembled as part of a synthetic control study in the past, provides a more comprehensive micro level lens to micro businesses.

4.1 Large Integrated Household Survey (GEIH)

GEIH is a multi-module comprehensive household survey that DANE develops and implements to produce basic statistics related to the demographic, social, and economic situations across Colombia. Originally a result of a structural framework of the Social Survey System over a 10-year period comprising of the core labor and income modules, GEIH has evolved to encompass broader topics and has expanded coverage to include all the departmental capital cities nationwide. For our purpose, GEIH is primarily used to profile the household and labor characteristics of microbusiness operators. Some relevant variables used from this dataset include:

- **Demographics & Geography:** age, gender, education, household structure, region/urbanicity
- **Labor Activity:** income sources, labor activity, hours worked
- **Household Indicators:** total household income, dependency, access to technology, access to community and social programs

4.2 Microbusiness Survey (EMICRON)

EMICRON is a specialized module within the larger GEIH survey that DANE started deploying from second quarter of 2013 that focuses specially on operators of microenterprises to gain a detailed understanding of the operating methods of these businesses, as well as their labor and financial characteristics. As our central dataset for analyzing the dynamics of micro-businesses, EMICRON offers detailed information on:

- **Business Characteristics:** sector, years of operation, business location, has formal name
- **Ownership, Operations and Finances:** gender, motivations, number of workers, sources of funding, use of financial services, use of technology, access to credit, revenues and costs

- **Business Practices:** use of signage, formal registration, digital presence, marketing presence

We integrate this data with GEIH data at the operator level to create a more complete profile of micro-business operators and their operations. These datasets also include an expansion factor, which is a value used to scale each observation in a survey sample to represent the number of people or units it corresponds to in the overall population. Hence, this factor is used to extrapolate any information across Colombia. Note: Both the GEIH and EMICRON dataset that was used for all the analysis does not include the population of the departments of Amazonas, Arauca, Casanare, Guainía, Guaviare, Putumayo, Vaupés and Vichada. Hence, for the purpose of this study, these departments are not included when all the statistics are reported.

4.3 Neighborhood Stores Survey

While GEIH and EMICON covers large part of Colombia, it still provides a more macro-level information without going too detailed on these businesses. Hence, we were provided with the Neighborhood Stores Survey data, which was originally developed for a synthetic control study. This dataset focuses specially on small retailers in urban areas of Colombia and captures the following-like information.

- **Geography:** socio economic stratum of where the business resides, structure, location
- **Owner Demographics:** gender, marital status, relationship to the head of household, education, access to various social benefits, family situation
- **Business Origin and Motivation:** funding source, registration type, business plan
- **Sales and Financials:** monthly sales, average transaction price, number of transactions, sales variability, profit margin, customer retention, use of technology for finance and sales
- **Products and Inventory:** product type, product diversity, use of technology for product,
- **Suppliers:** number of suppliers, supplier delivery frequency, volume discount percentage, credit with the supplier, supply behavior
- **Operation Behavior:** tracking, business outlook, perception, communication behavior

Since this dataset is structurally different from GEIH and EMICRON, we did not aim to directly integrate it with the other two datasets. The goal was to build qualitative and operational profiles of microbusinesses from macro level information and understand micro level information from this dataset as a complementary resource using those profiles.

4.4 Challenges with Data

Despite their richness, working with these datasets presented several challenges. We sourced data directly from the DANE repository across multiple years, which required building a rigorous data cleaning pipeline. The raw files used numeric codes rather than descriptive variable names, and the structure and naming conventions of variables varied significantly across years. To harmonize the data, we manually reviewed DANE's technical documentation to identify, rename, and align variables across files. Merging GEIH and EMICRON required additional cleaning steps to ensure consistency in variable definitions and to correctly apply expansion factors for national-level estimation. To address these issues, we implemented standardized weighting procedures based on DANE's guidelines, applied appropriate filters, and cross-validated our derived statistics against official DANE reports. The resulting data pipeline not only enabled consistent integration of complex survey data, but also offers a reusable framework that can inform future cleaning workflows and motivate better data structuring practices to minimize preprocessing effort.

The Neighborhood Stores Survey data presented a different set of challenges. Since it was constructed as part of a synthetic control analysis, this dataset was highly symmetrical and lacked much variability that real-world data usually presents. This made it difficult for us to extract meaningful trends or behavior.

5 Landscape of Micro-businesses in Colombia

Colombia's microbusiness population is diverse, spanning multiple sectors and social groups. Approximately 64% of microbusiness owners are male, and the majority, about 70%, operate in urban areas, with the remaining 30% in rural settings. These businesses are concentrated in four key sectors, with services being the largest, followed by trade, agriculture, and manufacturing.

In 2019, Colombia had over 6 million microbusinesses, but that number has since declined to 5.2 million in 2023. Of these 5.2 million, 55% are owned by individuals from poor or vulnerable populations, while only 4% come from high-income groups. Despite this, the wealth gap remains stark: high-income microbusinesses earn an average net profit of 6–7 million pesos per year, compared to just 300,000–400,000 pesos for their poor or vulnerable counterparts. This analysis will primarily focus on the economic realities of poor and vulnerable microbusiness owners in 2023.

5.1 Only 18% of Microbusinesses Apply for Loans

- **A small amount of microbusinesses in Colombia apply for a loan annually.**
- Only 15% of poor microbusinesses apply for a loan each year.
- In contrast, 20% of average-income microbusinesses apply for loans annually.
- Loan-applying poor and vulnerable microbusinesses spend and sell 1.8x more per month than those who dont apply.
- Loan applicants had an average income of 1,250,000/month vs. 997,000/month for non-applicants.
- Regional disparities are significant:
  - Antioquia: Lowest loan application rate at just 6.4%.
  - Atlántico: Highest rates, with around 30% of microbusinesses applying for loans.

5.2 43% Avoid Loans Due to Fear of Debt

- **Fear of debt is the top reason microbusinesses don't apply for loans.**
- Among poor and vulnerable populations:
  - 48% cite fear of debt as the main reason for not applying.
  - 17% say they don't meet loan requirements.
- Among high-income populations:
  - Only 29% cite fear of debt as their primary reason.
  - The majority simply report not needing a loan.

5.3 94% Loan Approval Rate for Microbusinesses

- **Most microbusinesses that apply are approved across all income levels.**
- Loan distribution by sector:
  - The trade sector receives 30% of loans, despite making up only 24% of microbusinesses.
  - The services sector receives 40% of loans, though it comprises 44% of microbusinesses.
  - The agriculture and manufacturing sectors follow a similar pattern to the services sector
  - This suggests trade-sector businesses may be more loan-eligible or appealing to lenders.

- Loan approval is largely unrelated to standard formality indicators:
  - 17% of businesses whose loan request are approved are registered, compared to 16% that are not.
  - 21% of businesses whose loan request are approved keep accounting records, the same percentage as those who don't.
  - 31% of businesses whose loan request are approved have a tax ID (RUT), while 35% of approved ones do not.

5.4 Only 19% Report Saving Money

- **Few microbusinesses in Colombia reported saving money in the past year.**
- Savings trends over time:
  - The percentage of businesses saving declined from 2019 to 2021.
  - Since 2021, the number has been steadily rising.
- Savings by income level:
  - Only 16% of poor and vulnerable microbusinesses reported saving.
  - 39% of high-income microbusinesses reported saving.
- Why businesses don't save:
  - 97% of non-saving businesses say it's because they don't have enough money.
- Regional differences in saving behavior:
  - Tolima had the highest saving rates, with 40% of microbusinesses saving.
  - Caqueta had the lowest saving rates, with only 8% reported saving in the past year.

5.5 Only 24% of Savers Save Through Financial Institutions

- **Few microbusinesses will save their money in banks and other financial institutions**
- Most microbusinesses save informally:
  - 67% of those who saved kept their savings at home.
  - Only 28% saved at a financial institution.
- Access to formal savings varies by income:
  - 73% of high-income businesses that save do so at financial institutions.
  - Only 15% of poor and vulnerable businesses that save use financial institutions.

There are two key takeaways from these exploratory findings. First, the main barrier to loans isn't approval—it's motivation. Most microbusinesses that apply are approved, but fear of debt keeps many from applying in the first place. Policy efforts should therefore focus on encouraging applications, not just improving eligibility. Second, when it comes to savings, the issue isn't just capacity—among those who can save, poor and vulnerable businesses are far less likely to use financial institutions. Policies should aim to shift these savings into formal channels, as high-income groups already do.

6 Methodology

To uncover the underlying patterns of financial inclusion and business formalization among Colombian micro-businesses, we structure our analysis into three sequential methodological phases. These phases leverage the principal component framework to combine dimensionality reduction, clustering, latent construct development, and structural modeling to provide a data-driven, yet policy-relevant framework.

- **Phase I: Clustering** — We apply the k-means clustering algorithm to uncover distinct micro-business profiles based on their underlying characteristics. To interpret and contextualize these clusters, we characterize them using external socioeconomic classifications. This approach enables us to identify meaningful typologies and highlight key features that effectively differentiate micro-business segments.
- **Phase II (a): Constructing Indices** — We use Multiple Correspondence Analysis (MCA) to examine the relationships among a set of categorical variables identified through clustering, with the goal of understanding how these variables align with dimensions such as formality, financial behavior, and business practices. By analyzing the spatial distribution of variables in the MCA space, we identify conceptually coherent groupings that serve as candidate indices for constructing latent dimensions.
- **Phase II (b): Understanding Indices through Clustering** - We analyze the coherent groupings identified through MCA and reapply clustering techniques to assess whether these derived features clearly reveal varying levels of maturity within each latent construct.
- **Phase III: Modeling and Index Finalization** — In the final stage, we apply a combination of modeling techniques to validate the relationships between the constructed indices, enabling us to assess how different dimensions of business behavior contribute to financial inclusion and to generate policy-relevant insights.

These phases are described in detail later on.

6.1 Social Classes

To better understand the economic positioning of these micro-business owners in Colombia, we adopt the income-based social classification that is used by UNDP and DANE. This framework categorizes an individual into four social classes based on their monthly household income.

- **Poor**: Households earning less than 396,864 Colombian pesos per month, official below the national poverty line
- **Vulnerable**: Households with monthly income between 396,864 and 704,201 pesos, above the poverty line but economically fragile and at risk of slipping back into poverty
- **Average**: Households with monthly income between 704,201 and less than 3,791,851 pesos, representing stable yet modest living conditions
- **High**: Households earning above 3,791,851 pesos per month, signifying greater resilience to economic fluctuations

These thresholds reflect per-household income levels, and they serve as a practical segmentation tool, allowing us to investigate patterns of formality, financial access, and business characteristics across distinct economic strata.

6.2 Literature Review

We draw on a range of prior studies that provide theoretical grounding and conceptual inspiration for understanding the pathways toward financial inclusion among micro-businesses, particularly in developing contexts like Colombia. These studies not only inform the selection and construction of indices used in our analysis, but also support the theoretical validity of the structural pathways later examined through modeling.

- **Formality:** We adopt the definition of formality from the UNDP policy note [3], which includes four key dimensions: Registration, Social Benefits, Tax Compliance, and Accounting Systems. These dimensions form the basis of our latent construct of formality and are confirmed to exhibit coherent structure during the exploratory phase via MCA.
- **Formality and Financial Inclusion:** According to Bruhn and McKenzie (2014), formalization increases a firm's likelihood of accessing formal financial services, as registered businesses are more visible to institutions and more likely to meet eligibility criteria [4]. While we include this as one marker, we do not treat formalization as the sole mechanism for financial inclusion. Instead, we aim to explore alternative pathways—such as digital adoption and operational maturity—that may offer comparable or complementary signals of trustworthiness and creditworthiness.
- **Digital Readiness and Financial Access:** The World Bank (2022) and GSMA Mobile Money reports emphasize that digital readiness—such as internet access, smartphone ownership, and the ability to accept digital payments—plays a pivotal role in expanding financial service uptake in emerging economies [5, 6]. Businesses with digital infrastructure are better positioned to engage with banks, fintechs, and e-government platforms.
- **Business Practice and Financial Inclusion:** Hoinaru et al. (2022) highlight that foundational business practices—including bookkeeping, forward planning, and external communication—are significantly associated with higher levels of financial inclusion, especially when formal registration is absent or partial [7]. These practices signal professionalism and operational readiness, thereby increasing the likelihood of credit access and financial trust. This inspires our treatment of business practices as an independent pathway toward financial inclusion, distinct from formal registration.
- **Social Class and Opportunity Access:** Literature on stratified access to financial tools, such as Giné and Townsend (2004), shows that household income and wealth influence how microbusinesses can benefit from digital and financial innovations [8]. We incorporate social class as a moderating factor in our structural modeling framework to account for baseline disparities in opportunity and resilience.

Building on these insights, we view microbusiness development and financial inclusion through the following conceptual lenses:

- *Business Best Practices and Personal Initiatives*
- *Digital Readiness*
- *Formality*
- *Financial Inclusion*

6.3 Phase I: Clustering

We begin our analysis by clustering micro-businesses across the various conceptual dimensions outlined in the literature review. To do this, we aggregate and pre-process relevant features from the GEIH and EMICRON datasets, encompassing both categorical and continuous variables.

For the clustering algorithm, we employ **k-means clustering**, a widely used unsupervised learning technique that partitions observations into k groups by minimizing the within-cluster sum of squares (WSS). The goal is to group businesses with similar characteristics together, thereby uncovering meaningful behavioral profiles across the micro-business landscape. Prior to clustering, all continuous variables—such as average business income, monthly sales, expenses, owner's age, number of employees, and months worked—are standardized to ensure equal contribution to the distance calculations.

Choosing the number of clusters, k, is a key step in k-means. We use the **elbow method**, which plots the within-cluster sum of squares (WSS) against different k values to find the point where improvements level off—the "elbow." This balances model simplicity with clustering quality. Based on this method, a k between 4 and 6 provides a good trade-off between cohesion and separation.

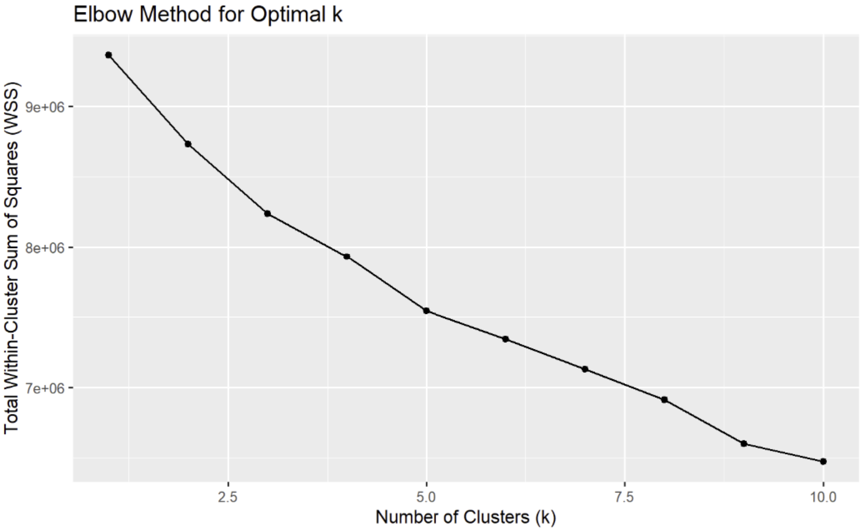


Figure 1: Elbow Method

After clustering, we analyze the resulting groups by examining how key categorical variables—such as income-based social class and geographic region—differ across clusters. Notably, one cluster accounted for only 2% of the microbusiness population but exhibited substantially higher income and profit levels than the others. This group closely aligned with the previously defined "high-class" segment, offering external validation of the clustering output, as shown below.

Variable	**High Cluster**	**Other Clusters**
monthly_expenses_GASTOS_MES	3,502,841	227,545
prev_month_costs_COSTOS_MES_ANTERIOR	14,378,168	729,961
prev_year_costs_COSTOS_ANIO_ANTERIOR	140,339,952	7,898,507
prev_year_sales_VENTAS_ANIO_ANTERIOR	247,457,642	19,195,306
business_net_profit_past_year_P550	28,838,104	5,064,326
avg_business_income_per_month_P3072	4,479,459	692,542

The following variables were selected as initial candidates for use in the indices due to the clear separability observed between the 'high cluster' and the other clusters.

Candidates for Formality Index Variables

Variable	High Cluster	Other Clusters
business_is_registered_with_chamberce_of_commerce_P1055 (yes)	70.1%	10.8%
business_registered_as_P1056 (legal entity)	21%	4.3%
business_has_single_tax_registry_RUT_P1633 (yes)	86.8%	23.6%
filed_income_tax_return_past_year_P2991 (yes)	64.5%	13.8%
filed_VAT_past_year_P2992 (yes)	27.6%	2.6%
filed_ICA_past_year_P2993 (yes)	49.3%	16.8%
source_of_social_security_P6110 (pays full fee)	72.3%	40.3%
business_does_accounting_or_daily_bookkeeping_P6775 (yes)	54.1%	6%
record_type_used_to_keep_accounts_P640 (Balance sheet or P/L)	29.4%	1.2%
contract_type_verbal_or_written_P6450 (written)	78%	39.1%
paid_ARL_P3090 (yes)	37%	4.4%
paid_compensation_fund_or_SENA_ICBF_P2989 (yes)	14.4%	1%

Candidates for Financial Formality Index Variables

Variable	High Cluster	Other Clusters
source_of_founding_money_P3052 (bank loans)	25.3%	7.7%
have_applied_credit_or_loan_last_year_P1765 (yes)	35.8%	18%
what_did_you_use_the_loan_for_P1570 (invest in business)	82.2 %	59.1%
where_did_you_save_P1771 (financial inst.)	67.5%	22.6%
paid_health_or_pension_last_month_P3088 (yes)	52.3%	6.8%
accepts credit card (yes)	37.0%	4.4%
accepts digital payment(yes)	52.1%	12.4%

Candidates for Digital Readiness Index Variables

Variable	High Cluster	Other Clusters
has_email_P3000 (yes)	44.7%	12.2%
business_has_presence_on_social_network_P1559 (yes)	33.8%	10%
business_has_internet_service_P2524 (yes)	78.9%	35.6%
business_has_website_P2532 (yes)	14.1%	1.5%
use_laptop_or_tablet_for_business_P4001 (yes)	57.6%	10.2%
use_cellphone_for_business_P976 (yes)	92.6%	64.8%

Candidates for Business Sophistication Index Variables

Variable	High Cluster	Other Clusters
business_is_visible_to_public_P469 (yes)	84.2%	65.3%
has_business_name_P3035 (yes)	67.4%	15.4%
has_exclusive_space_in_house_for_this_work_P3095 (yes)	72.7%	39.3%
highest_level_of_education (university)	25.6%	13%
place_mainly_worked_in_P6880 (fixed office)	65.6%	17.3%
role_in_work_P6430 (employer)	61.6%	7.6%
have_people_who_help_you_P1800 (yes)	70.5%	16.2%
workers_who_receive_payments_P3032_1 (yes)	86.9%	46%

In this way, the clustering phase not only uncovers underlying patterns in the data but also serves as a foundation for the subsequent dimensionality reduction and latent construct development.

6.4 Phase II: Exploration and Filtration through Multiple Correspondence Analysis

Building on the observed features that proved useful in clustering businesses across different behavioral dimensions in Phase I, we next explore how well these variables align with interpretable latent patterns. Given that the GEIH and EMICRON surveys offer rich, predominantly categorical data—from household demographics to detailed business practices—we apply Multiple Correspondence Analysis (MCA), a dimensionality reduction technique specifically designed for categorical variables.

MCA extends the principles of Correspondence Analysis by mapping multiple categorical variables into a lower-dimensional Euclidean space. In this geometric framework, each level of a categorical variable is represented as a point, and proximity between points reflects the frequency with which they co-occur across observations. The first few dimensions capture the majority of variation in response patterns, providing interpretable axes for latent structure.

We evaluate the MCA output by examining both the proportion of variance explained by each dimension and the contribution of individual variables to the inertia (variance) of each dimension. This helps determine which features meaningfully differentiate micro-businesses across underlying behavioral constructs.

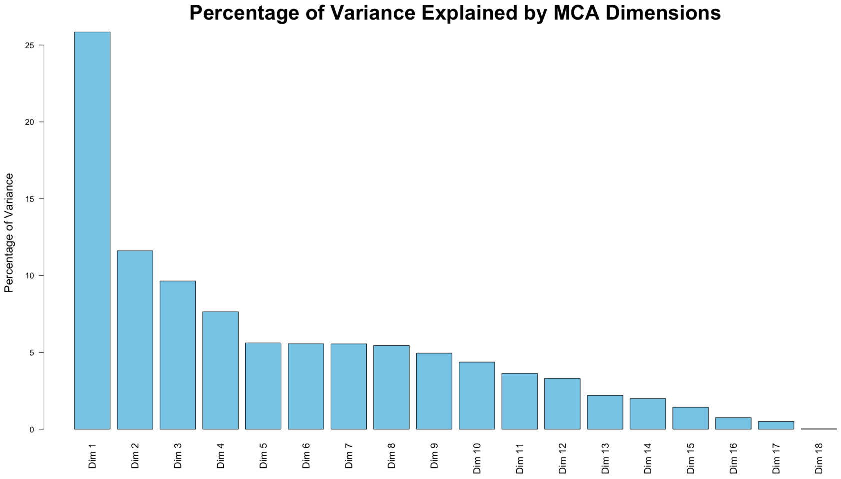


Figure 2: Percentage of Variance Explained by MCA Principal Components

To interpret the latent dimensions visually, we plot the MCA coordinates of key variables identified in Phase I. As shown in the chart below, Dimension 1 appears to represent a gradient of business formality. Variables such as `filed_income_tax`, `has_tax_registry`, and `is_registered` show clear spatial separation between "No" (near the origin) and "Yes" (further along the positive axis), suggesting that Dimension 1 captures variation along the formal–informal spectrum. Conversely, variables like `has_social_security` cluster near the center for both "Yes" and "No" responses, indicating weak discriminatory power and minimal contribution to latent differentiation. These are excluded from subsequent index construction. Similarly, while variables such as `filed_ICA` and `filed_VAT` shift toward the formal end of the dimension, their lack of internal contrast across categories suggests they add limited explanatory value and are also excluded.

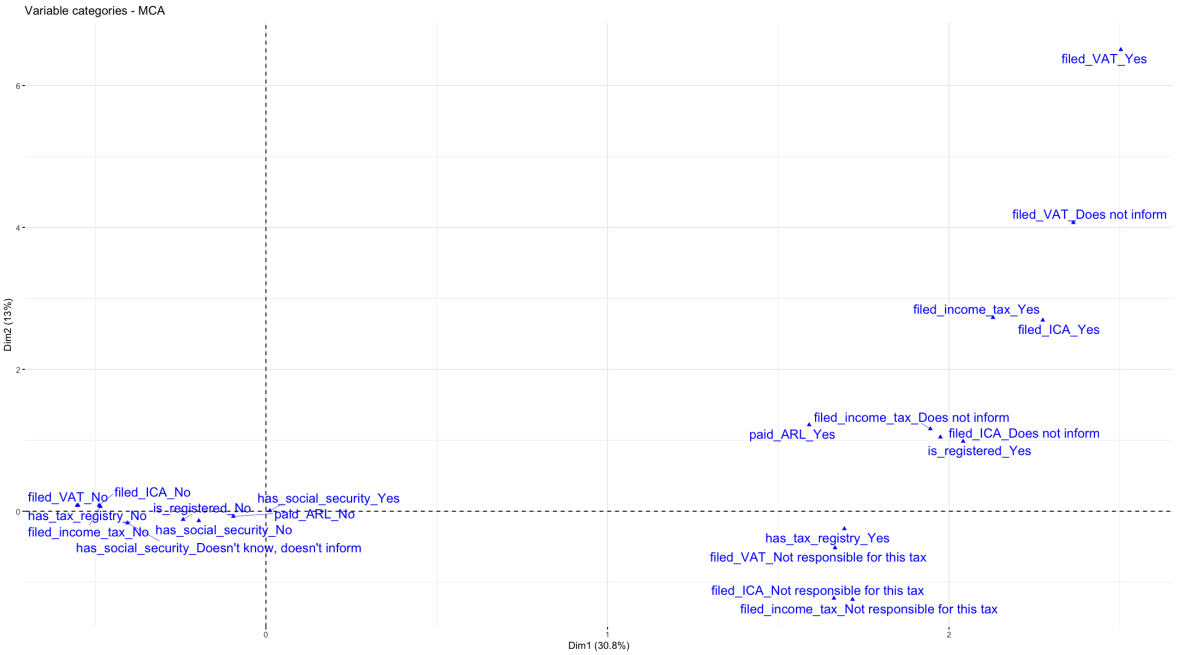


Figure 3: Projection of Active Variable Categories in MCA Space

To further interpret the latent dimensions uncovered by MCA, we incorporate a range of demographic and behavioral characteristics as *supplementary variables*. While the *active variables*—as shown in the previous plot—contribute directly to the construction of the MCA dimensions, supplementary variables are projected onto the same geometric space to help interpret those dimensions in terms of additional features not used in their creation. We include several continuous and ordinal variables—such as income, age, sales, and expenses—as supplementary variables to assess whether these features show meaningful alignment with the latent dimensions. This aids in determining whether such variables might be valuable additions to composite indices or can serve as contextual descriptors.

For instance, in the plots below, we examine the projection of ”Average Monthly Business Income” onto the MCA space. The plot on the left displays the direction and strength as a supplementary quantitative variable, visualized as an arrow. The direction of the arrow—pointing toward the right—suggests that higher income is modestly associated with businesses exhibiting greater formality. However, the short length of the arrow indicates that this correlation is relatively weak. The plot on the right reinforces this interpretation by displaying individual businesses in MCA space, with color indicating income level. While we observe that some of the higher-income observations appear more often on the right side of the space, the majority of businesses have very low income, and there is no strong or continuous income gradient across the dimension.

We employ such tools to study relationship and exclude variables from our index construction when it does not contribute meaningfully to the latent dimensions we aim to capture and could introduce noise or bias in composite scoring.

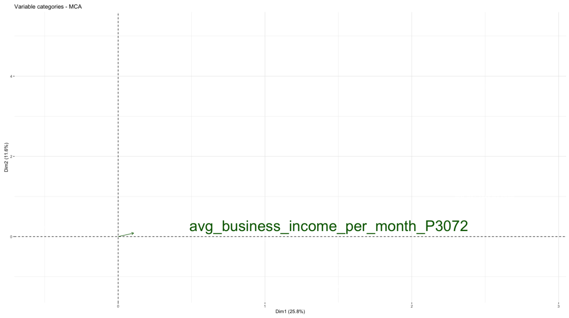


Figure 4: Direction and Strength of Average Monthly Income in MCA Space

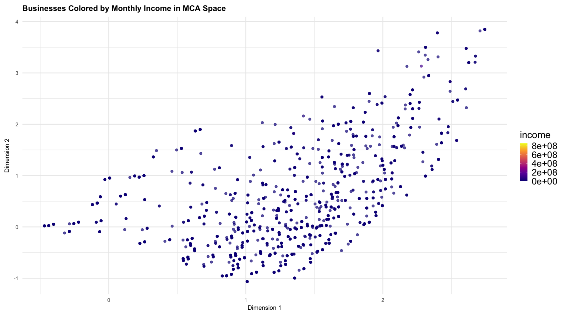


Figure 5: Distribution of Individual Businesses Colored by Average Monthly Income

Similarly, the plot below shows "Highest Education Level" of the business owner, treated as an ordinal supplementary variable. A clear gradient emerges along Dimension 1: businesses with more highly educated owners tend to align more closely with the formal end of the axis, providing further evidence that this dimension meaningfully captures variation in business formality.

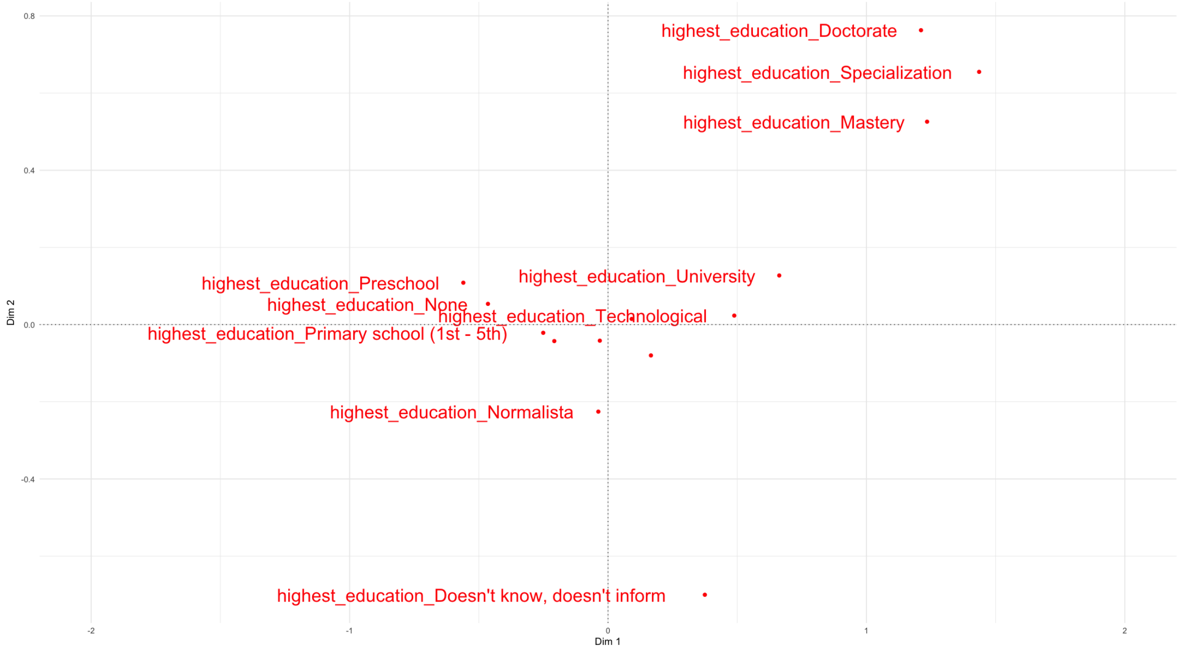


Figure 6: Projection of Supplementary Variable Categories in MCA Space

We repeat this process across multiple thematic domains—such as financial access, digital readiness, and operational maturity—to systematically identify which features contribute substantively to the latent constructs of interest. Through this process of visual interpretation and variable filtration, we construct a set of conceptually grounded and empirically distinct dimensions. These dimensions serve as the foundation for downstream clustering, profiling, and modeling in the next phase.

6.4.1 Further Clustering to Understand these Indices

To explore potential sub-segments within the latent domains constructed through MCA, we reapplied clustering techniques to assess the diversity of business behaviors within each domain. The goal was to uncover varying levels of maturity or engagement within each thematic construct by identifying distinct, domain-specific business profiles.

Since few latent domains included a mix of categorical and continuous variables, we employed **Factor Analysis of Mixed Data (FAMD)**—a dimensionality reduction method that combines the principles of PCA and MCA to handle mixed-type datasets. The quality of FAMD results was evaluated based on the cumulative variance explained by the first few dimensions and the interpretability of how input features aligned with these axes. After filtering for dimensions that collectively explained at least 70% of total variance, we applied the elbow method to determine the optimal number of clusters.

This approach offered a useful lens into the heterogeneity of micro-businesses within specific behavioral or financial domains. For example, when clustering based on household usage of financial products, we observed a discernible trend: some businesses clustered around credit card use without access to home or vehicle loans, while others showed no engagement with financial products at all.

However, as illustrated in figure below, the separation between clusters was not always distinct. The limited granularity of features within certain domains led to overlapping data points, reducing the clarity of the segmentation. As a result, these domain-specific clusters served primarily as exploratory tools for profiling, rather than definitive typologies.

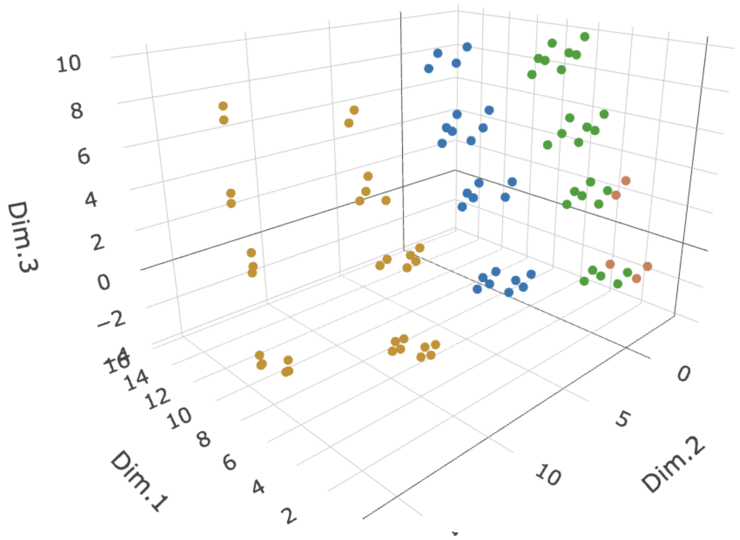


Figure 7: Clustering of Microbusinesses Based on Credit Utilization Characteristics

6.5 Phase III: Modeling and Index Finalization

While the latent domains and their associated profiles serve as powerful descriptive tools for policy development, we extend our analysis to examine the interrelationships among these constructs and their influence on financial inclusion. This phase seeks to formalize the conceptual connections between domains and quantify how they interact to shape financial inclusion outcomes.

6.5.1 Structural Equation Modeling (SEM)

Structural Equation Modeling (SEM) is a statistical framework that integrates factor analysis and path analysis to estimate both measurement properties and directional relationships between latent constructs. SEM comprises two components: the *measurement model* and the *structural model*.

Measurement Model (Confirmatory Factor Analysis) In the measurement model, we use Confirmatory Factor Analysis (CFA) to validate the structure of the latent constructs identified in earlier phases. Conceptually, this can be expressed as:

$$\mathbf{x} = \Lambda \boldsymbol{\eta} + \boldsymbol{\epsilon}$$

Where: - $\mathbf{x}$ is a vector of observed variables (e.g., `is_registered`, `paid_ARL`) - $\boldsymbol{\eta}$ represents latent variables (e.g., *Formality*) - Λ is the matrix of factor loadings, indicating how strongly each observed variable is associated with its corresponding latent variable - $\boldsymbol{\epsilon}$ captures measurement error

We use the dimensions constructed through clustering and MCA to define these latent variables and validate their coherence through CFA, confirming that the selected indicators consistently represent the intended underlying constructs.

Structural Model (Path Analysis) In the structural model, we model the directional relationships between latent constructs. This component can be represented as:

$$\boldsymbol{\eta} = \mathbf{B}\boldsymbol{\eta} + \boldsymbol{\Gamma}\boldsymbol{\xi} + \boldsymbol{\zeta}$$

Where: - $\boldsymbol{\eta}$ are the endogenous latent variables (dependent constructs) - $\boldsymbol{\xi}$ are the exogenous latent variables (independent constructs) - $\mathbf{B}$ is the matrix of regression coefficients among endogenous constructs - $\boldsymbol{\Gamma}$ represents regression weights from exogenous to endogenous constructs - $\boldsymbol{\zeta}$ denotes structural error terms

The objective is to estimate parameters such that the model-implied covariance matrix closely approximates the observed sample covariance matrix. This enables us to test theoretically informed relationships between domains—such as *Formality*, *Digital Readiness*, and *Business Sophistication*—and quantify how these dimensions contribute to financial inclusion. By estimating the strength and direction of these pathways, SEM allows us to construct composite indices that reflect the empirical contribution of each domain, grounded in both data and theory. These indices, in turn, can support targeted policy design and measurement of progress in financial inclusion initiatives.

Drawing on insights from both prior literature and our exploratory analyses, we construct a structural model to test specific directional pathways between latent constructs. Our hypothesized model includes paths from *Business Sophistication* to *Formality*, *Digital Readiness*, and *Financial Formality*, as well as a direct path from *Digital Readiness* to *Financial Formality.* We also assess potential indirect effects, such as the mediated pathway:

$$\text{Business Sophistication} \rightarrow \text{Digital Readiness} \rightarrow \text{Financial Formality}$$

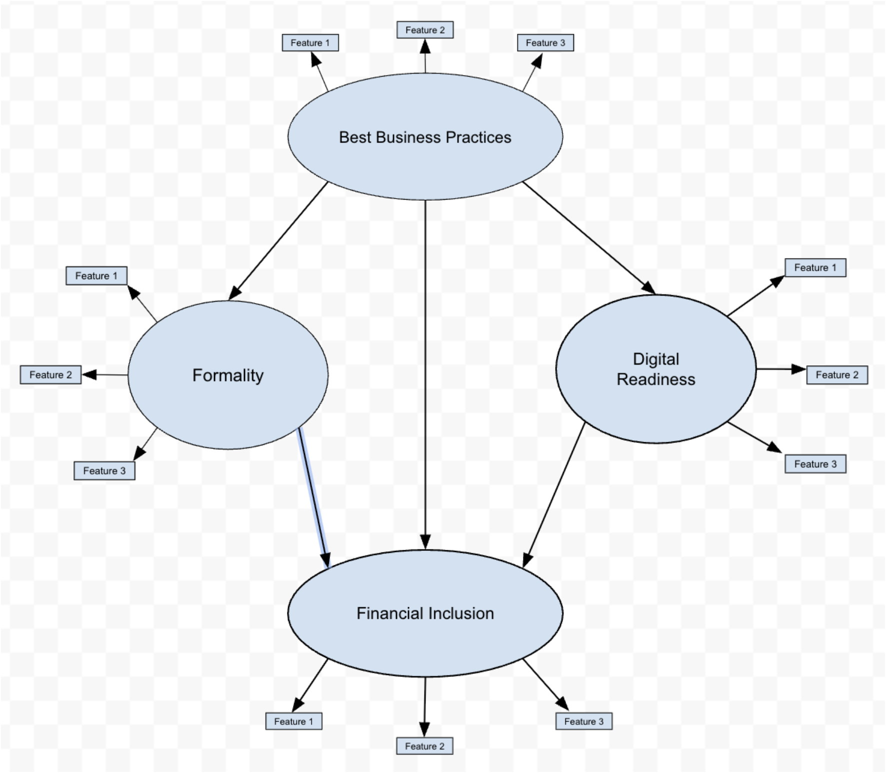


Figure 8: Path Diagram of Structural Equation Model Linking Formality, Digital Readiness, Best Business Practices, and Financial Inclusion

To estimate this model, we minimize a discrepancy function between the sample covariance matrix ($\mathbf{S}$) and the model-implied covariance matrix ($\Sigma(\theta)$), where θ represents the set of model parameters. Formally, the model is fit by solving:

$$F(\theta) = \text{discrepancy}(\mathbf{S}, \Sigma(\theta)) \to \min$$

Given the presence of categorical and ordinal variables, we use the **Weighted Least Squares Mean and Variance adjusted (WLSMV)** estimator, which is appropriate for modeling non-continuous data. WLSMV internally relies on polychoric correlations to estimate covariances between ordinal variables.

Model Evaluation We assess model fit using two standard indices: the **Comparative Fit Index (CFI)** and the **Tucker–Lewis Index (TLI)**. CFI compares the specified model to a baseline (null) model assuming no relationships between variables, while TLI penalizes overly complex models. A threshold of 0.95 or higher for both indices is typically considered evidence of good fit. We also validate that latent constructs are well-defined by ensuring that factor loadings are sufficiently high, generally using a threshold of 0.7 or above on the fully standardized scale (`std.all`).

Handling Unbalanced Binary Variables One challenge in our model arises from highly unbalanced binary variables, many of which are dominated by ”No” responses (e.g., only a small share of businesses accept credit cards or have filed VAT). Because WLSMV uses polychoric correlation, which assumes an underlying bivariate normal distribution, it can produce misleadingly high correlation estimates under extreme imbalance.

To address this, we compute **Jaccard similarity** between binary variables to assess whether high correlations are driven by true co-occurrence rather than shared sparsity. Jaccard similarity is defined as:

$$\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

Where A and B are sets of observations where the variable is "Yes." If we observe that two variables have low Jaccard similarity but high polychoric correlation, we consider that correlation unreliable. In such cases, we avoid including both variables simultaneously in the SEM model. Instead, we estimate multiple models with different combinations of such features to assess their individual contribution. Variables are only retained together in the same dimension if both demonstrate meaningful explanatory power.

While alternative approaches such as *Bayesian SEM* or *Conditional SEM* may offer better control over this issue by accounting for prior information or underlying structure, we consider these out of scope for the current study but valuable directions for future research.

Limitations of SEM in Institutional Context A broader limitation of SEM is that it estimates relationships based on existing institutional patterns, which may themselves reflect biased gatekeeping mechanisms. For example, formal registration and tax compliance may emerge as strong predictors of financial inclusion simply because they are used by financial institutions as eligibility filters. In contrast, dimensions like digital engagement or entrepreneurial initiative—though potentially strong alternative indicators—may appear weaker simply because they are not yet recognized or rewarded by formal systems.

This institutional bias can cause SEM to reinforce existing norms rather than identify genuinely inclusive alternatives. Despite this, we believe SEM remains a valuable tool for understanding current pathways to financial inclusion. To further validate that the alternative constructs we model (e.g., digital readiness, business sophistication) can also serve as reliable predictors, we complement this analysis with targeted machine learning models in the next section.

6.5.2 Targeted Modeling

As discussed earlier, a core challenge in financial inclusion in Colombia is not that most businesses are denied credit, but that many never apply for it at all. This suggests that the most impactful interventions should focus on expanding participation—encouraging business owners to engage with the financial system—rather than solely improving approval rates. While the indices developed in this study do not fully resolve this issue, we investigate whether they offer alternative, behaviorally grounded signals of trustworthiness. If effective, such signals could help financial institutions identify and engage currently excluded micro-businesses. This, in turn, opens the possibility for a co-evolving ecosystem in which businesses gain access to credit, and institutions acquire more inclusive and reliable tools for assessing financial risk.

To evaluate whether the indices and features we developed are predictive of financially meaningful outcomes, we complement our SEM analysis with a targeted machine learning approach. Specifically, we train a **random forest classifier** on a subset of businesses that have applied for credit, aiming to distinguish those who used credit *productively*—i.e., for business investment—from those who did not.

We assess model performance using the *out-of-bag* (OOB) error rate and class-specific confusion matrices. The resulting OOB error rate is 23.73%, indicating reasonably good predictive performance for this type of behavioral data. For feature importance, we rely on the **Mean Decrease in Gini** metric, which evaluates how much each variable contributes to improving node purity across decision trees. Variables with higher scores are more useful in splitting the data into homogeneous outcome groups, and are thus considered more informative.

7 Results

This analysis of different features helped us filter down to the following variables in each index:

7.1 Final 4 Indices

Formality Index

- Is the business registered with the chamber of commerce?
- Does the business have a single tax registry (RUT)?
- Has the business paid ARL (occupational risk insurance)?
- Has the business paid compensation fund or SENA/ICBF (social and labor welfare)?
- Has the business paid for health or pension in the last month?

Financial Formality Index

- Does the business save money in a financial institution?
- Does the business accept online payments?
- Does the business accept credit cards as payment?
- Has the business applied for a credit or loan in the last year?

Digital Readiness Index

- Does the business have an email?
- Does the business use a cellphone?
- Does the business use a laptop or tablet?
- Does the business have internet service?

Business Sophistication Index

- Does the business have a name?
- Does the business have workers who receive payments?
- Does the business have a website?
- Does the business do accounting or daily bookkeeping?

Each business receives 1 point for every ”yes” response. The final index score is calculated as the total points divided by the number of variables in the index. When analyzing the relationships between these indices using Structural Equation Modeling (SEM), we observe a Comparative Fit Index (CFI) of 0.972 and a Tucker–Lewis Index (TLI) of 0.968. These values indicate a strong model fit to the specified structure. Based on this validated model, we report the following relationships among the latent constructs:

Latent Variable (lhs)		**Observed Variable (rhs)**	**Estimate**	**Standardized**
Formality	=	is_registered	1.000	0.983
Formality	=	has_tax_registry	0.943	0.926
Formality	=	paid_ARL	0.995	0.978
Formality	=	paid_SENA_ICBF	0.877	0.862
Formality	=	paid_health_and_pension	0.941	0.925
DigitalReadiness	=	has_email	1.000	0.634
DigitalReadiness	=	use_cellphone	1.454	0.921
DigitalReadiness	=	use_laptop_or_tablet	1.466	0.929
DigitalReadiness	=	uses_internet	1.503	0.952
BusinessSophistication	=	has_name	1.000	0.747
BusinessSophistication	=	workers_receiving_payments_P3032_1	0.755	0.564
BusinessSophistication	=	has_website	0.923	0.689
BusinessSophistication	=	does_bookkeeping	0.234	0.175
HouseholdFinancial Engagement	=	household_uses_credit_card_P5222S7	1.000	0.823
HouseholdFinancial Engagement	=	household_uses_free_investment_loan_P5222S6	0.521	0.429
HouseholdFinancial Engagement	=	household_uses_home_purchase_loan_P5222S4	0.679	0.559
HouseholdFinancial Engagement	=	household_uses_vehicle_purchase_loan_P5222S5	0.823	0.678
HouseholdFinancial Engagement	=	household_uses_checking_account_P5222S1	0.552	0.455
HouseholdFinancial Engagement	=	household_uses_savings_account_P5222S2	0.984	0.811
HouseholdFinancial Engagement	=	household_uses_CDT_P5222S3	0.738	0.608
FinancialFormality	=	saved_money_in_financial_institution	1.000	0.437
FinancialFormality	=	accepts_online_payment_P1764_3	1.905	0.832
FinancialFormality	=	accepts_credit_card_P1764_6	1.952	0.853
FinancialFormality	=	applied_credit	0.376	0.164

Table 2: SEM Estimates: Factor Loadings

As shown in the table above, each latent construct is reasonably well-defined, with most factor loadings (standardized, std.all) exceeding the commonly accepted threshold of 0.7. One exception is `has_email`, which loads at 0.634 on the *Digital Readiness* construct. Although this value is still moderate, it suggests that email usage may not be as central to digital readiness among micro-businesses as other indicators such as internet or device use. Similarly, `does_bookkeeping` exhibits a notably weak loading of 0.175 on the *Business Sophistication* construct. This likely reflects poor co-occurrence with other defining indicators of business sophistication, such as having a business name or website. This misalignment suggests that the bookkeeping variable may either be measuring a separate behavior or requires more targeted contextualization within this construct.

Examining the structural pathways between constructs below, we observe a strong direct effect from *Business Sophistication* to *Formality* (standardized estimate = 0.906), affirming that operational maturity contributes meaningfully to formalization. We also see a substantial link from *Business Sophistication* to *Digital Readiness* (0.721), as well as a moderate direct effect from *Digital Readiness* to *Financial Formality* (0.646), even after controlling for formality. This suggests that policies promoting digital infrastructure and digital capabilities may play a critical role in enabling financial inclusion. On the other hand, the direct effect from *Business Sophistication* to *Financial Formality* is weaker (0.379), which may be partly explained by indirect effects via digital readiness or by the current composition of that index. Enriching this construct with additional behavioral indicators could improve its explanatory power.

Outcome (lhs)		Predictor (rhs)	Estimate	Standardized (std.all)
Formality	∼	BusinessSophistication	1.192	0.906
DigitalReadiness	∼	BusinessSophistication	0.612	0.721
FinancialFormality	∼	DigitalReadiness	0.446	0.646
FinancialFormality	∼	BusinessSophistication	0.221	0.379
FinancialFormality	∼	HouseholdFinancialEngagement	0.038	0.071

Table 3: SEM Estimates: Structural Relationships Between Latent Variables

One index tested but ultimately excluded from the final SEM model was *Household Financial Engagement*, which attempted to capture how business owners' engagement with financial products in their household impact their business financial engagement. While conceptually interesting, this construct yielded a low standardized path coefficient (0.071) in relation to financial formality. This weak relationship suggests that, in its current form, household financial behavior does not serve as a reliable proxy for business-level financial engagement. However, this does not necessarily imply that household financial behavior is unimportant—rather, richer and more targeted indicators would be needed to evaluate its relevance.

Overall, the structural model provides empirical support for a multi-dimensional understanding of financial inclusion, emphasizing that formality, digital readiness, and operational practices each play a role. While some latent constructs may be further strengthened with more granular data, the framework we present is flexible and extensible. As richer datasets become available, the methodology can be adapted to include additional dimensions, making it a practical tool for both monitoring inclusion and guiding policy design.

The targeted modeling we did with random forest highlights several features—such as bookkeeping, use of digital tools, and business visibility—that consistently emerge as strong predictors of productive credit use. Importantly, many of these features align with those identified in the SEM framework, even for businesses that lack formal registration. This convergence reinforces the idea that alternative behavioral indicators can serve as reliable proxies for financial trust and capability, and may form the basis for more inclusive credit scoring approaches in the future.

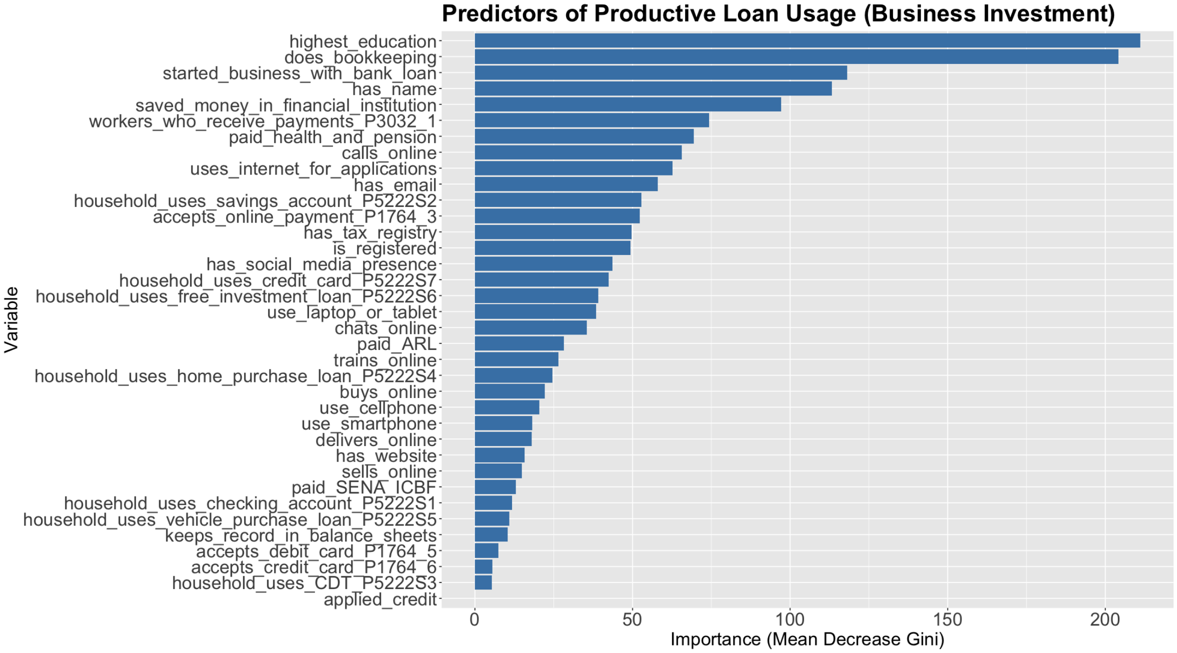


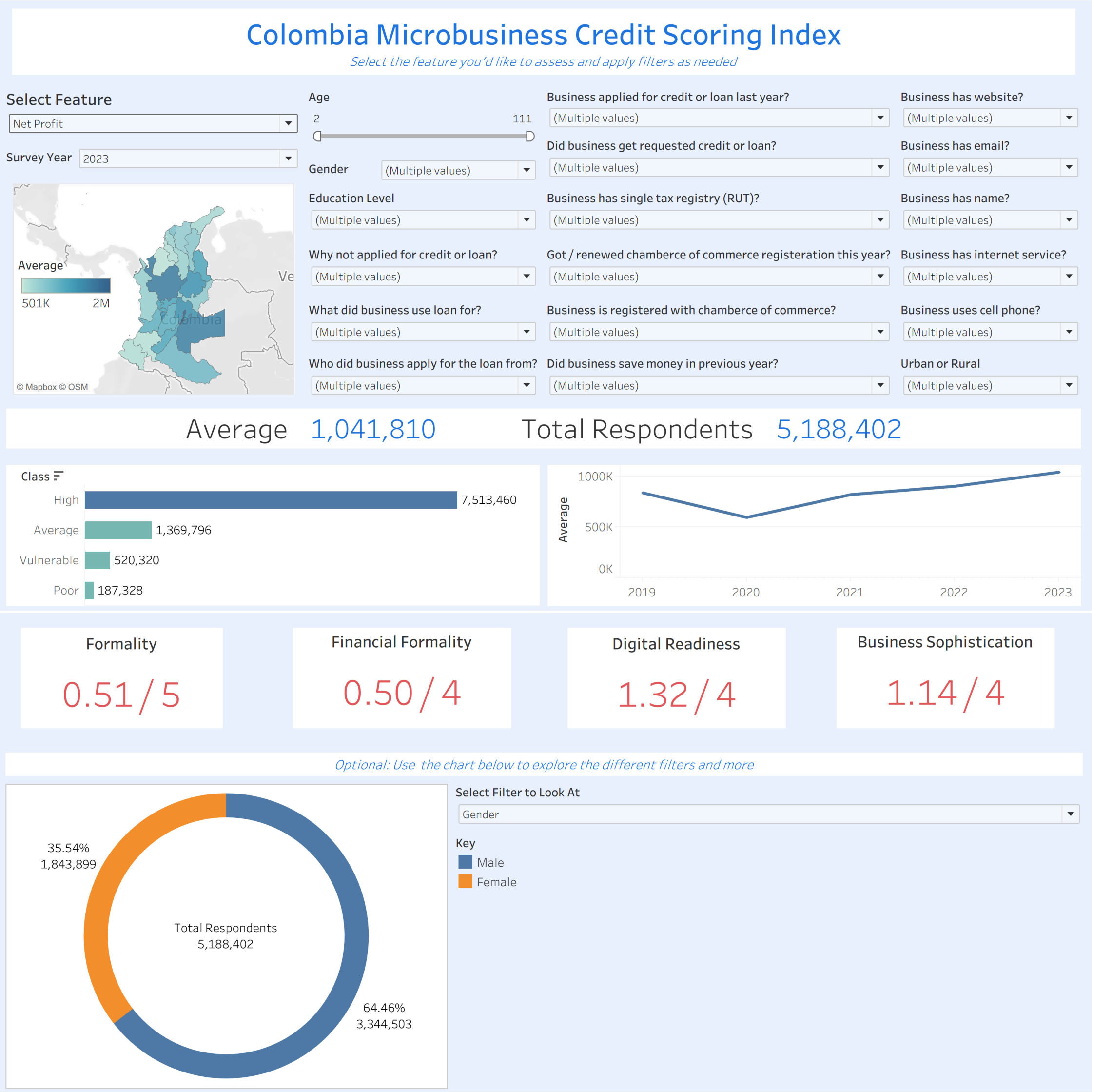
Figure 9: Top Predictors of Loan Usage for Business Development Based on Mean Decrease in Gini Impurity

7.2 Interactive Dashboard with Agentic Support

This interactive dashboard acts as a "living lab" for financial inclusion policy. It integrates EMICRON and GEIH survey data from 2019 to 2023 and is designed to be updated annually. A separate data pipeline can automatically processes raw survey data from DANE into a clean, standardized format compatible with Tableau. This structure ensures consistent variables across years, enabling reliable year-over-year comparisons and longitudinal analysis of microbusiness dynamics.

On the top left of the dashboard, users can explore features like monthly expenses, annual revenue, or net profit. Users can then apply a wide range of filters, such as sector, education level, or registration status, or click directly on the map, income class chart, or timeline to view results by region, socioeconomic class, or year. The dashboard then displays average values, total respondents, and average index scores (formality, financial formality, digital readiness, and business sophistication) for the selected profile, along with a donut chart summarizing key categorical traits like business sector and funding sources. This allows users, whether policymakers or business owners, to identify common characteristics, benchmark performance, and tailor interventions based on the dashboard summaries.

A built-in chatbot powered by LLaMA 3 and RAG also offers support by explaining the indices, suggesting policies, and simulating outcomes—drawing from this report and wider Colombian microbusiness research.



Colombia Microbusiness Credit Scoring Index

Select the feature you'd like to assess and apply filters as needed

Select Feature: Net Profit

Survey Year: 2023

Age: 2 – 111

Gender: (Multiple values)

Education Level: (Multiple values)

Why not applied for credit or loan?: (Multiple values)

What did business use loan for?: (Multiple values)

Who did business apply for the loan from?: (Multiple values)

Business applied for credit or loan last year?: (Multiple values)

Did business get requested credit or loan?: (Multiple values)

Business has single tax registry (RUT)?: (Multiple values)

Got / renewed chamberce of commerce registeration this year?: (Multiple values)

Business is registered with chamberce of commerce?: (Multiple values)

Did business save money in previous year?: (Multiple values)

Business has website?: (Multiple values)

Business has email?: (Multiple values)

Business has name?: (Multiple values)

Business has internet service?: (Multiple values)

Business uses cell phone?: (Multiple values)

Urban or Rural: (Multiple values)

Average 1,041,810 Total Respondents 5,188,402

Class	
High	7,513,460
Average	1,369,796
Vulnerable	520,320
Poor	187,328

Formality	Financial Formality	Digital Readiness	Business Sophistication
0.51 / 5	0.50 / 4	1.32 / 4	1.14 / 4

Optional: Use the chart below to explore the different filters and more

Select Filter to Look At: Gender

Key
- 🟦 Male
- 🟧 Female

Total Respondents 5,188,402

35.54% 1,843,899

64.46% 3,344,503

Figure 10: Interactive Dashboard

8 Policy Recommendations

8.1 Promote Formality via Business Sophistication

While government-backed incentives for formality, such as preferential loan terms and temporary tax exemptions, have proven effective in countries like Indonesia [10], their implementation can complicated. Bureaucratic hurdles, limited outreach, and administrative delays may reduce their accessibility and impact for microbusinesses. Additionally, navigating government programs can be challenging for small business owners, which can hinder timely adoption and limit the overall effectiveness of these incentives.

Business Sophistication, which includes having a business name, doing bookkeeping, paying workers, and having a website, is a strong predictor of formality (standardized loading 0.906). Encouraging businesses to take these manageable steps can be an easier and more intuitive path to formalization than pushing for full registration immediately. Notably, having a business name and keeping basic accounts are both easier steps than paying workers or maintaining a website, and they also stand out as two of the strongest predictors of productive loan usage, meaning loans are more likely invested back into the business rather than spent on personal expenses. This makes them practical, impactful starting points for encouraging formalization.

Starting with these simple actions builds good habits that support growth and loan effectiveness. From there, businesses can gradually progress toward full formality, including registration, tax compliance, and social security payments. This stepwise approach lowers barriers and helps businesses formalize more naturally over time.

8.2 Encourage Digital Transformation

8.2.1 Online Payments

As accepting online and credit card payments accounts for half of the Financial Formality Index, integrating digital payment options is critical for improving financial formality. One promising pathway is through platforms like Bold, a fast-growing Colombian fintech company that helps small businesses accept electronic payments through low-cost, easy-to-use devices. Founded in 2019, Bold aims to promote financial inclusion by making digital payments accessible to small and medium-sized enterprises. Their quick registration process allows merchants to begin accepting payments within minutes [9]. Despite these tools being available, many microbusinesses still don't use digital payment systems. Encouraging the adoption of platforms like Bold—or similar alternatives—can help microbusinesses expand their customer base, build transaction histories, and gain access to formal financial services.

8.2.2 Digital Readiness

A second pillar of digital transformation should focus on improving digital readiness, as measured by the Digital Readiness Index—a strong predictor of financial formality. This includes having access to an email account, cellphone, laptop or tablet, and internet service. To expand access, Colombia can build on existing efforts like the Computadores para Educar (CPE) program, which in 2024 aims to deliver over 57,000 computer terminals to students [12]. A parallel initiative for microbusinesses could provide subsidized or refurbished devices bundled with basic business tools and digital training. Additionally, expanding low-cost internet connectivity programs, offering tax deductions for digital equipment, and integrating digital literacy into current entrepreneurship and financial inclusion programs would help microbusinesses better participate in the digital economy—laying the foundation for greater financial formality and long-term growth.

8.3 Boost Loan Uptake and Savings Accounts

To increase loan applications among microbusinesses, policy should focus on addressing the widespread fear of debt—reported by 43% of microbusinesses, and even higher (48%) among poor and vulnerable groups. Despite a high loan approval rate of 94% in 2023, only 18% of microbusinesses currently apply for credit. Additionally, saving in formal financial institutions is low, with only 15% of poor and vulnerable businesses using these services compared to 73% of high-income businesses. Increasing formal savings is crucial for building creditworthiness and financial resilience. Both credit application and saving at a financial institution are key components of the Financial Formality Index. Improving uptake in these areas would directly strengthen businesses' financial formality and long-term growth potential.

Expanding Colombia's Banca de las Oportunidades offers a strategic policy pathway. This government initiative already promotes financial inclusion through education, credit facilitation, and financial product access. By scaling its reach, especially focusing on microbusinesses in vulnerable communities, Banca de las Oportunidades can provide tailored financial literacy programs to reduce fear of debt, offer hands-on loan application support, and encourage savings through accessible, low-cost products. Such targeted expansion would help close the gap in credit uptake and formal saving, driving broader financial formality and business growth [11].

8.4 Survey Microbusiness Progress, Not Only Status

To better understand the causal drivers of financial inclusion and formality among microbusinesses, it is essential to incorporate a time component into the GEIH and EMICRON survey data. Currently, the data provides only cross-sectional snapshots, capturing businesses at a single point in time, so we can observe correlations (e.g., between digital readiness and financial inclusion) but not causal pathways. Most microbusinesses in Colombia remain financially excluded, and without knowing when or how a business became included, we cannot determine whether certain features led to inclusion or were adopted afterward.

That said, some variables in the current data already hint at a temporal dimension. For instance, 25% of high-income microbusiness owners used bank loans as their startup capital, compared to just 8% in lower-income clusters, suggesting that early access to formal credit could play a foundational role in long-term inclusion. Likewise, higher education is consistently associated with higher business income. Meanwhile, other characteristics, like having a business name, may not be directly causal but are simple to adopt and strongly correlated with higher income: 67% of high-income businesses have a name, versus only 15% in lower-income clusters. Even if causality is uncertain, such low-cost, high-return practices may still be worth promoting.

To move from correlation to causation, future data collection efforts should focus on longitudinal tracking, observing the same businesses over time, to identify which changes precede inclusion and success.

- Track the timing of key milestones (e.g., when a business registers formally) to see if they come before or after increases in income or inclusion.
- Capture the frequency and timing of financial behaviors (e.g., monthly spending habits) to better understand capacity and planning.
- Record when businesses apply for loans, and whether inclusion/formality happened before or after.
- Expand GEIH and EMICRON into surveys that follow the same businesses annually.

8.5 Adoption To Other Countries

This model was possible to develop in Colombia thanks to the country's comprehensive survey data on microbusinesses, coming from the GEIH and EMICRON surveys, which capture detailed information across regions, sectors, and income levels. This dataset allowed for the identification of key features that differentiate high-income microbusinesses from those that are poor or vulnerable, enabling targeted and evidence-based policy development.

Thanks to this strong foundation, similar models can be adapted and scaled to other countries with comparable data collection efforts. Even in places where exact datasets are not available, the methodology—leveraging available business characteristics to identify the most predictive factors of financial inclusion and formality—can guide the creation of customized indices. These indices can then be integrated into credit evaluation frameworks or tested with partner institutions to improve financial access and support for microbusinesses.

Ultimately, this approach offers a path to move beyond generic policies toward precise, data-driven strategies that foster financial inclusion, formalization, and sustainable growth in the microbusiness sector worldwide.

9 References

[1] United Nations. (2022). *Data center: Human Development Reports*. Retrieved from https://hdr.undp.org/data-center/human-development-index#/indicies/HDI

[2] Freixes, J. (2024, December 4). Social Stratification in Colombia: A laudable but flawed system. *Colombia One*. Retrieved from https://colombiaone.com/2024/12/04/colombia-stratification/

[3] UNDP. (n.d.). *Policies and Procedures Portal*. United Nations Development Programme. Retrieved from https://popp.undp.org/

[4] Bruhn, M. (2014). Entry regulation and the formalization of microenterprises in developing countries. *The World Bank Research Observer*. Retrieved from https://academic.oup.com/wbro/article-abstract/29/2/186/1631296

[5] World Bank. (n.d.). *Financial Inclusion Overview*. Retrieved from https://www.worldbank.org/en/topic/financialinclusion/overview

[6] Raithatha, R., & Storchi, G. (2024). *The State of the Industry Report on Mobile Money 2025*. GSMA. Retrieved from https://www.gsma.com/sotir/

[7] Woodruff, C., & McKenzie, D. (2015). Business practices in small firms in developing countries. *World Bank Open Knowledge Repository*. Retrieved from https://openknowledge.worldbank.org/entities/publication/89b6a68a-666f-518e-9e54-a385c9b7b664

[8] Giné, X., & Townsend, R. (2004). Evaluation of financial liberalization: A general equilibrium model with constrained occupation choice. *Journal of Development Economics*. Retrieved from https://www.sciencedirect.com/science/article/abs/pii/S0304387804000161

[9] Lara, M. G. (2024). Bold raises $50M in Series C funding round led by General Atlantic and IFC. *IFC*. Retrieved from https://www.ifc.org/en/pressroom/2024/bold-raises-50m-in-series-c-funding-round-led-by-general-atlantic-and-ifc

[10] Asian Development Bank. (2022). *Modernizing Local Government Taxation in Indonesia*. Retrieved from https://www.adb.org/sites/default/files/publication/791356/modernizing-local-government-taxation-indonesia.pdf

[11] Bank of Opportunities. (2017). *Banca de las Oportunidades*. Retrieved from https://www.bancadelasoportunidades.gov.co/

[12] Government of Colombia. (2024, November 26). *¿Qué es Computadores para Educar?*. Retrieved from https://www.computadoresparaeducar.gov.co/publicaciones/1/que-es-computadores-para-educar/