Brazilian Journal of Biometrics, **44**, 1–13 (2026)
doi:10.28951/bjb.v44i2.948

BRAZILIAN JOURNAL OF BIOMETRICS
ISSN:2764-5290

ARTICLE

Machine Learning for predicting the number of children among maternal health in Inimutaba – Minas Gerais

Aurélio de Aquino Araújo*[,1] and Ricardo Cerri[2]

[1]Pontificia Universidad Católica de Chile, Facultad de Matemáticas, Santiago, Chile.
[2]Universidade de São Paulo (USP), Instituto de Ciências Matemáticas e de Computação (ICMC), São Carlos, SP, Brazil.
*Corresponding author. Email: aurelio.aquino94@gmail.com

(Received: August 2, 2025; Revised: November 13, 2025; Accepted: December 3, 2025; Published: May 21, 2026)
Section Editor: João Domingos Scalon

Abstract

This study describes the profile of low-income mothers who already had children and received prenatal care in the municipality of Inimutaba, Minas Gerais, between January 2020 and December 2023. Sociodemographic variables were analyzed to identify the factors most strongly associated with the number of children per woman. The sample included 134 women with an average age of 29 years, the majority of whom self-identified as *Brown* (86.56%), living on a minimum wage (94%), and having completed high school (94%).

To investigate the determinants of fertility within this population, we applied an Ordinary Least Squares (OLS) regression and a Classification and Regression Tree (CART) algorithm. Although the in-sample OLS model initially exhibited a high coefficient of determination ($\hat{R}^2 \approx 0.91$), internal validation through 10-fold cross-validation demonstrated substantially lower generalization performance ($\hat{R}^2 \approx 0.37$), indicating overfitting. The most relevant predictors after model refinement were family size, household income, marital status (single), and age.

The CART model showed similar patterns. The fully grown tree overfitted the training data, whereas cost-complexity pruning with an optimal parameter of $\alpha = 0.113$ improved test accuracy to 74%. Despite this improvement, the model exhibited limited discriminative ability (AUC = 0.55), reinforcing the importance of careful interpretation in small epidemiological datasets.

Overall, the findings indicate that fertility patterns among low-income mothers in Inimutaba are closely linked to socioeconomic vulnerability. Machine learning and statistical models, when combined with rigorous internal validation, can help identify high-risk subgroups and support targeted interventions for maternal health.

Keywords: Machine learning; Maternal health; Linear regression; Decision trees; Public health.

Introduction

Historically, and even in current times, the prospect of pregnancy leads women to experience various changes in their daily lives, characterized by metabolic and physical alterations that can significantly affect their living environment (Silva & Souza, 2018; Rodrigues & Nascimento, 2020; Silva & Gomes, 2019). Another concerning factor that can impact maternal health is emotional distress, which in turn affects their financial situation (Oliveira & Lima, 2020). Furthermore, it is essential to note that for a long time, women have been responsible for unpaid domestic tasks. Today, many of them not only maintain this role but also serve as the main providers of the household income, resulting in a double burden in their daily routines (Costa & Almeida, 2019; Almeida & Castro, 2019).

Adolescent pregnancy remains a persistent public health concern in Brazil. Although most births occur among adult women, early pregnancy still affects a substantial number of adolescents, especially those living in socioeconomically vulnerable contexts. According to official data from the Sistema de Informação sobre Nascidos Vivos (SINASC), in 2020 there were 17,526 births among girls aged 10–14 and more than 363,000 births among adolescents aged 15–19 (da Saúde, 2020). These figures highlight a cycle of long-term vulnerability, as early motherhood is associated with higher lifetime fertility, reduced educational attainment, and poorer socioeconomic conditions (Gonçalves & Lima, 2020; Pereira *et al.,* 2021).

In the 2000s, more than 20 million families were headed by low-income women. By the third quarter of 2022, this number had increased to over 37 million, representing a growth of more than 70% over the years (Mainardi & Bidoia, 2022; IBGE, 2022). This trend becomes particularly concerning due to the high number of children, with an average of three children per household headed by such women. This reality contributes significantly to inhumane birth conditions, often linked to lack of maternal education, marital status, and other exploitative aspects of pregnancy, such as teenage pregnancy and unplanned motherhood (Martins & Pereira, 2019; Mainardi & Bidoia, 2022).

The results of this study enable the identification of relevant characteristics that may support efforts to address this issue from a statistical standpoint, focusing on the reality of low-income maternal health among children in the city of Inimutaba, Minas Gerais. The goal is to determine which demographic variables predominate in this context and influence the number of children among these women using a quantitative approach.

In this article, we present statistical and machine learning models applied to a population of low-income maternal health with children in the city of Inimutaba, Minas Gerais. The article is organized into the following sections: Methods, Results, Discussion, Conclusion, and References.

Materials and Methods

The data were obtained from records of maternal health who received prenatal care at the Municipal Health Center in the city of Inimutaba, Minas Gerais, from January 2020 to December 2023, totaling 134 participants. Using the electronic medical records of the maternal health, the data were extracted by professionals from the Neonatal Department, who recorded maternal and neonatal information during prenatal follow-up. The dataset included demographic variables such as age, education level, number of family members, number of children, household income, marital status, and race/skin color.

According to the 2022 census, the city of Inimutaba, located in the state of Minas Gerais, in the Central Mineira Region of the Alto Médio São Francisco — Microrregion of Médio Rio das Velhas, 167 km from the state capital, Belo Horizonte, had a population of 7,371 inhabitants. The population density was 13.99 inhabitants per square kilometer, composed of an urban zone with eight neighborhoods and a rural zone with twelve localities (IPEA, 2019).

Statistical Analysis

The implementation of statistical and data modeling analyses was carried out using Python 3.6 as the programming language. D'Agostino and Pearson's tests (D'Agostino & Pearson, 1973) were applied to assess whether the variables followed a normal distribution. Categorical variables were summarized using absolute and relative frequencies (percentages). The chi-square test was used to assess differences between groups of categorical variables.

Linear regression models were used to evaluate which variables - such as age, education level, number of family members, household income, marital status, race/skin color, height, and weight - had an influence on the number of children among these women. In each step, the least significant variable (with a significance level above 5%) was removed from the model. Three models were constructed, based on different groups of variables characterizing maternal health.

Additionally, a Decision Tree algorithm was applied to compare its predictive performance with that of the best regression model, including pre-pruning strategies. Accuracy and the area under the Receiver Operating Characteristic (ROC) curve were calculated. The ROC curve graphically represents the trade-off between True Positive Rate and False Positive Rate, providing an illustration of the model's performance as the classification threshold varies. Other performance metrics included precision, recall, and F_1-score, which respectively represent the model's sensitivity and harmonic mean of precision and recall (**a29**; **a30**; **a31**; **a32**; Oliveira & Fernandes, 2021).

Model selection followed recommendations from the statistical learning literature, which emphasizes combining linear parametric models and interpretable non-parametric models when analyzing sociodemographic determinants in public health studies. Ordinary Least Squares (OLS) remains the standard baseline for continuous outcomes because of its transparency, ease of interpretation, and suitability for small epidemiological datasets (Harrell Jr, 2015). However, linearity assumptions may not fully capture hierarchical or nonlinear behavioral patterns. For this reason, we also employed a Decision Tree model, which is widely recommended for health services research due to its interpretability, ability to detect interaction structures, and direct usefulness in profiling vulnerable subpopulations (Breiman *et al.,* 1984; Loh, 2011).

To ensure robust generalization and mitigate overfitting — a recurrent concern in biomedical predictive modeling — both approaches were evaluated using 10-fold cross-validation, a validation strategy recognized as the most reliable internal method when sample sizes are limited (Kuhn & Johnson, 2013; Arlot & Celisse, 2010). The use of cross-validation aligns with best practices in epidemiological modeling, as it provides an unbiased estimate of predictive performance while minimizing variance induced by data partitioning, thereby producing more reliable inferences for maternal-health applications.

Assessment of potential overfitting

The initial in-sample performance of the linear regression model suggested the possibility of overfitting, with an R^2 = 0.6765 and RMSE = 0.6064. Although these values indicate a reasonably good fit to the observed data, in-sample metrics are well known to produce optimistic estimates of predictive accuracy, especially in small epidemiological datasets with correlated sociodemographic variables (Harrell Jr, 2015; Babyak, 2004). To obtain a more reliable estimate of generalization capacity, we conducted 10-fold cross-validation following established recommendations for internal validation in biomedical modeling (Arlot & Celisse, 2010; Kuhn & Johnson, 2013).

The cross-validated mean R^2 decreased to 0.3737, with a standard deviation of 0.5383, indicating substantial variability across folds. This instability demonstrates that the model's predictive structure is highly sensitive to data partitioning, a characteristic often observed when sample size is limited and explanatory variables exhibit low variance or multicollinearity (Hastie *et al.,* 2009). The cross-validated RMSE (mean = 0.6399; SD = 0.1979) was also higher than the in-sample RMSE, further supporting the interpretation that the original model overestimated its true predictive performance.

Taken together, these findings confirm that the in-sample R^2 overstated the explanatory power of the model and that cross-validation provides a more conservative and methodologically robust estimate. This evaluation mitigates the influence of overfitting and provides a more accurate understanding of the sociodemographic determinants associated with fertility outcomes in maternal health contexts.

Table 1 summarizes the 10-fold cross-validated metrics for both models, providing a more realistic assessment of generalization performance compared to the in-sample estimates.

Model Validation Using 10-fold Cross-Validation

To ensure reproducibility and robust estimation of generalization error, all models were evaluated using 10-fold cross-validation with fixed random seeds. This approach is recommended when working with small epidemiological datasets because it minimizes bias associated with single train–test splits and produces more reliable estimates of predictive performance (Kuhn & Johnson, 2013; Arlot & Celisse, 2010).

Table 1. Cross-validated performance metrics for OLS and Decision Tree models

Model	$\hat{R}^2$ (Mean)	$\hat{R}^2$ (SD)	RMSE (mean $\pm$ SD)
OLS Regression	0.3737	0.5383	0.6399 ± 0.1979
Decision Tree	0.3515	0.3642	0.6721 ± 0.2440

Results

Sample Characteristics

Between January 2020 and December 2023, a total of 134 maternal health who received prenatal care at the Municipal Health Center of Inimutaba, Minas Gerais, were included in this study.

The participants had a mean age of 29.04 years and lived in households with an average of 4 family members. Most women already had more than one child. The sample was predominantly composed of individuals who self-identified as Brown (86.56%), had completed high school (94%), and were single (79%), as detailed in Table 2.

To assess distributional properties of quantitative variables (Age, Number of Family Members, Number of Children, and Income), the Shapiro–Wilk normality test was applied. All variables showed statistically significant results (p-*values* < 0.05), indicating that their distributions deviate from normality. Therefore, subsequent analyses involving these variables rely on statistical techniques robust to non-normality.

Regarding qualitative variables, pairwise associations were evaluated using the chi-square test. Statistically significant associations were identified between Race/Color and Education Level (p-*value* $= 0.0012$) and between Education Level and Marital Status (p-*value* $= 2.357 \times 10^{-6}$), suggesting dependence between these sociodemographic characteristics. Conversely, no significant association was found between Race/Color and Marital Status (p-*value* $= 0.1920$), indicating independence between these variables in the analyzed population.

Ordinary Least Squares Regression Model (OLS)

For the construction of this model, we use multiple linear regression, which predicts the quantitative dependent variable Y based on other predictor variables $(X_1, X_2, \ldots, X_n)$, and can be defined as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + \varepsilon.$$

Table 2. Demographic characteristics of the maternal health population who underwent prenatal care at the Health Center of Inimutaba from January 2020 to December 2023

Variable	Count (n=134)	Percentage (%)
Age		
17 to 19 years	6	4.48
20 to 29 years	75	55.97
30 to 39 years	46	34.33
40+ years	10	7.46
Race/Color		
Brown	116	86.56
White	13	9.70
Asian (Yellow)	4	2.99
Black	1	0.75
Education Level		
High School	126	94.03
Higher Education	8	5.97
Marital Status		
Single	106	79.10
Married	19	14.18
Divorced	9	6.72
Number of Children		
1	63	47.01
2	41	30.60
3	23	17.16
4+	7	5.22
Number of Family Members		
2	3	2.24
3	40	29.85
4	54	40.30
5+	37	27.61
Family Income		
1 Minimum Wage	127	94.78
2 Minimum Wages	7	5.22

In this case, the parameters $\beta_0, \beta_1, \ldots, \beta_n$ are unknown and must be estimated from the data. The fitted model is written as:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_n X_n,$$

where $\hat{Y}$ denotes the predicted value of the response variable.

The estimation procedure relies on the method of least squares, which seeks to find the coefficient vector that minimizes the Residual Sum of Squares (RSS):

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

thereby quantifying the variability of the residuals.

This framework also allows formal testing of whether there is any association between the dependent and independent variables through the following hypothesis system:

$$H_0 : \beta_0 = \beta_1 = \cdots = \beta_n = 0,$$
$$H_1 : \text{at least one } \beta_j \text{ is different from zero.}$$

To obtain an initial, intuitive assessment of model performance and maintain comparability with the Decision Tree analysis, the dataset was split into training and testing sets. The training set was used to fit the regression model, and the testing set was used to obtain an out-of-sample performance estimate based on a single partition. In our study, 70% of the data were used for training and 30% for testing (**a33**; **a34**; **a35**). However, as discussed in the subsection on model validation, the main evaluation of generalization relies on 10-fold cross-validation rather than on this single split.

Note on model validation. To avoid misinterpretation of apparent model performance, all in-sample $\hat{R}^2$ values reported in this section are accompanied by their corresponding 10-fold cross-validated estimates. This ensures that the evaluation of explanatory capacity reflects true generalization rather than overfitting, which is particularly relevant in small epidemiological datasets with correlated sociodemographic predictors.

Tables 3, 4, and 5 summarize the successive OLS specifications considered, showing how the set of predictors was refined by removing non-significant variables and how this refinement affected the in-sample behavior of the models.

First OLS Model

Table 3. Results of the initial linear regression model for the population of maternal health from January 2020 to December 2023

Variable	Coefficient	*p*-value	Confidence Interval
Intercept	-0.5255	0.054	[–1.060; 0.009]
Age	0.0232	0.045	[0.001; 0.046]
Family Members	0.7600	< 0.001	[0.614; 0.906]
Household Income	-0.5869	0.024	[–1.093; –0.081]
Race [Asian]	-0.2386	0.514	[–0.962; 0.485]
Race [White]	-0.1868	0.433	[–0.659; 0.285]
Race [Brown]	-0.1001	0.573	[–0.452; 0.252]
Race [Black]	7.5×10^{-17}	0.420	$[-1.1 \times 10^{-16}; 2.6 \times 10^{-16}]$
Education [High School]	-0.2154	0.186	[–0.537; 0.106]
Education [Higher Education]	-0.3101	0.234	[–0.825; 0.205]
Marital Status [Single]	-0.3144	0.024	[–0.586; –0.043]
Marital Status [Married]	-0.1117	0.562	[–0.493; 0.270]
Marital Status [Divorced]	-0.0995	0.691	[–0.596; 0.397]

In this initial specification, the in-sample value of $R^2 \approx 0.66$ suggested a moderate association between the set of predictors and the number of children. However, Table 3 shows that several covariates (particularly race and education categories, as well as some marital status levels) were not statistically significant, with *p*-values well above the 5% level. This combination of a moderate in-sample fit and multiple non-significant predictors suggested that the model was overparameterized for the available sample size, indicating that a simpler specification would be more appropriate.

From a modeling perspective, this first OLS model was therefore used as an exploratory step to screen variables rather than as a definitive explanatory model. Guided by the significance patterns in

Table 3, we proceeded to remove non-contributory predictors and to construct a more parsimonious specification.

Second OLS Model

Table 4. Results of the first adjusted linear regression model for the population of maternal health from January 2020 to December 2023

Variable	Coefficient	*p*-value	Confidence Interval
Age	0.0137	0.167	[–0.006; 0.033]
Family Members	0.7244	< 0.001	[0.595; 0.854]
Household Income	-0.8825	< 0.001	[–1.223; –0.542]
Marital Status [Single]	-0.3533	0.021	[–0.653; –0.054]

After removing the non-significant race, education, and marital status categories, the second model retained age, family members, household income, and marital status (single). In this reduced specification, the in-sample coefficient of determination was $\hat{R}^2 \approx 0.91$, which at first glance suggests a strong explanatory capacity.

However, this apparent performance must be interpreted with caution. As detailed in the cross-validation subsection, when the model was evaluated using 10-fold cross-validation, the mean cross-validated $\hat{R}^2$ dropped to approximately 0.37 (see Table 1). This discrepancy between in-sample and cross-validated performance is characteristic of overfitting, especially in small datasets with correlated sociodemographic variables. Thus, while Table 4 identifies family size, income, and single marital status as relevant predictors, the in-sample $\hat{R}^2$ clearly overestimates the true predictive power of the model.

Given that age was no longer statistically significant in this adjusted specification (p-value = 0.167), an additional simplification step was performed by removing this variable, producing a third, more parsimonious model.

Third OLS Model

Table 5. Results of the second adjusted linear regression model for the population of maternal health from January 2020 to December 2023

Variable	Coefficient	*p*-value	Confidence Interval
Family Members	0.7733	< 0.001	[0.664; 0.883]
Household Income	-0.7457	< 0.001	[–1.026; –0.465]
Marital Status [Single]	-0.3426	0.026	[–0.643; –0.042]

The third model retains only the variables that remained consistently significant: number of family members, household income, and marital status (single). In-sample, this specification yielded $\hat{R}^2 \approx 0.90$ and a high F-statistic, which might suggest a very good fit if considered in isolation. Nonetheless, as with the second model, the 10-fold cross-validated $\hat{R}^2$ remained close to 0.37 (Table 1), indicating that the apparent 90% explanation is largely an artifact of overfitting.

Therefore, Table 5 should be interpreted primarily in terms of the direction and statistical significance of the coefficients rather than as evidence that the model explains 90% of the variability in the number of children. The cross-validated results show that, once internal validation is taken into account, the model attains only moderate explanatory power. Even so, the signs and magnitudes

of the coefficients are consistent with the social vulnerability literature: a larger number of family members and lower income are associated with higher parity, while being single is negatively associated with the number of children compared to the reference marital category.

In summary, the sequence of OLS models demonstrates that: (i) a naive interpretation of in-sample $\hat{R}^2$ would lead to an inflated perception of explanatory power; (ii) internal validation via 10-fold cross-validation reveals a substantially lower, more realistic predictive performance; and (iii) family size, income, and marital status emerge as the most robust sociodemographic determinants of the number of children among the maternal health in this study.

Decision Tree Induction Algorithm

To complement the linear regression analysis and to incorporate a non-parametric, highly interpretable predictive model, we fitted a Classification and Regression Tree (CART). Decision trees partition the feature space recursively using splits that maximize impurity reduction at each node, enabling the identification of nonlinear and hierarchical relationships often present in sociodemographic and maternal health data (Breiman *et al.,* 1984; Loh, 2011; Hastie *et al.,* 2009).

All sociodemographic variables were included as predictors. The dataset was divided into training (70%) and testing (30%) subsets. As expected for small samples, the fully grown tree overfitted the training data, achieving high in-sample performance but reduced generalization. Thus, pruning was applied to obtain a more stable and interpretable structure.

Cost-Complexity Pruning

To control overfitting, we applied cost-complexity pruning (Breiman *et al.,* 1984). This approach considers a sequence of subtrees obtained by progressively removing splits that contribute marginally to predictive performance. For a given subtree T, the penalized cost is:

$$R_\alpha(T) = R(T) + \alpha |f(T)|,$$

where $R(T)$ is the misclassification error and $|f(T)|$ is the number of terminal nodes. Larger values of α favor simpler, more generalizable trees.

Figure 1 shows the relationship between α and predictive accuracy. As α increases from zero, the model transitions from an overfitted structure with high variance to a more stable configuration. The highest testing accuracy was observed at $\alpha = 0.113$.

Pruned Decision Tree

Using the optimal value $\alpha = 0.113$, we refitted the model to obtain the final pruned tree shown in Figure 2. This simplified structure removes noisy, high-variance splits while preserving the most informative branching patterns related to family composition and maternal characteristics.

The pruned model achieved improved generalization compared with the unpruned tree, reducing overfitting and offering a more interpretable pattern of fertility determinants. The first splitting variable remained *family members*, consistent with the regression results, reinforcing its central role in the demographic structure of low-income households. This interpretability is particularly relevant for maternal health decision-making, as the model highlights specific sociodemographic profiles that may be prioritized in community-level interventions.

Discussion

Pregnancy is a multidimensional event that deeply affects women's social trajectories, shaping educational, economic, and relational opportunities throughout their lives (Freitas & Santos, 2019). In Brazil, these effects are intensified by structural socioeconomic inequalities that disproportionately

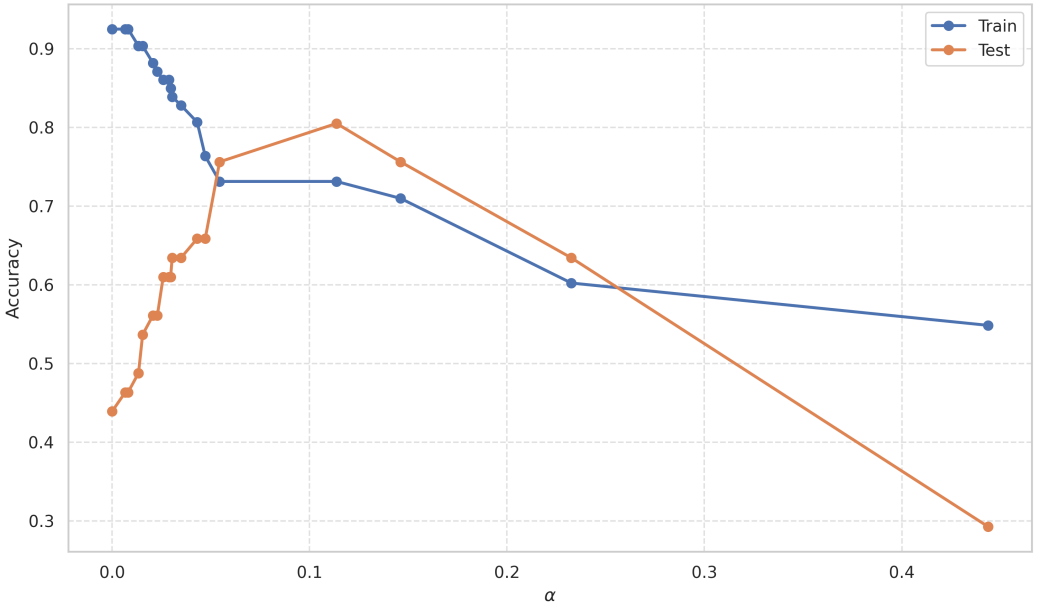

Figure 1. Accuracy of the Decision Tree model across different values of the cost-complexity parameter α for the training and testing datasets.

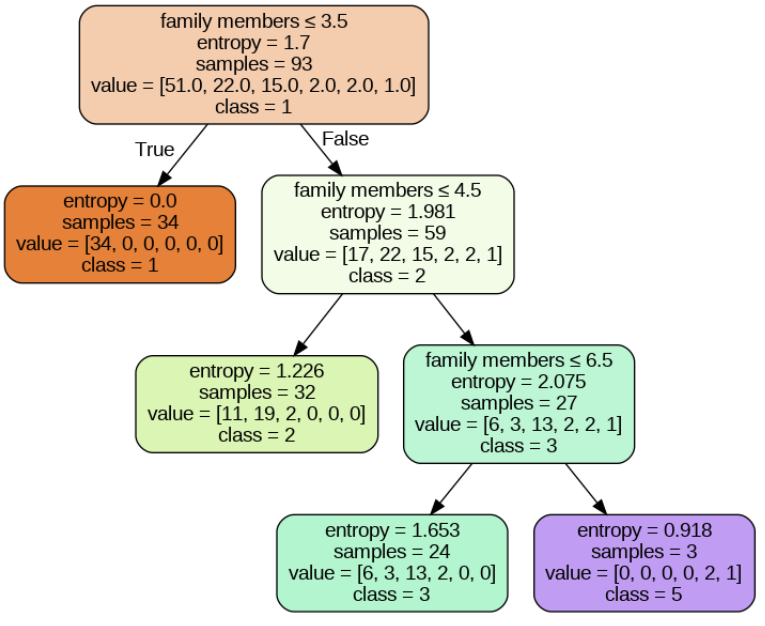

Figure 2. Decision Tree after cost-complexity pruning using $\alpha = 0.113$.

impact women in situations of vulnerability, particularly those with low schooling, unstable marital bonds, and limited access to formal employment (Martins & Pereira, 2019; Souza & Matos, 2020; Barros & Rocha, 2020). Understanding how these determinants interact is crucial for maternal

health planning, particularly in small municipalities where social vulnerability is prevalent.

The demographic profile of maternal health in Inimutaba between 2020 and 2023 reflects patterns commonly described in the national literature. Nearly half (47%) already had at least one child, 79% were single, and 94% lived with a family income equivalent to one minimum wage. These women reached an average per capita income of approximately R$434.00, situating them among the over 60 million Brazilians living under significant material deprivation (IBGE, 2022; da Cidadania, 2021). The social context of low income and precarious living conditions is known to influence reproductive behavior, maternal autonomy, adherence to prenatal care, and long-term child development (Nascimento & Moreira, 2020; Campos & Silva, 2020; Vieira & Lopes, 2019; Silva & Andrade, 2021; Cunha & Oliveira, 2021).

Family size emerged as another central characteristic of vulnerability in this study. Approximately 40% of maternal health lived with four people, and 28% with five or more. Large households are associated with economic stress, limited access to resources, and greater dependency ratios, all of which heighten maternal and child vulnerability (Nascimento & Moreira, 2020; Cunha & Oliveira, 2021). While many families develop adaptive strategies that enable social advancement despite adversity, evidence consistently shows that household overcrowding intensifies exposure to health, nutritional, and educational risks (Campos & Silva, 2020; Silva & Andrade, 2021). Thus, reproductive planning in these contexts is deeply influenced by living conditions, highlighting the need for targeted public policies.

Racial inequalities also permeate these findings. More than a quarter of Brazilian women are Black or Brown, groups that historically present lower educational attainment and are disproportionately impacted by social exclusion (IBGE, 2021). In this study, 86.56% of maternal health identified as Brown and 94% had low schooling. Prior research shows clear gradients: as educational attainment decreases, the average number of children increases, and opportunities for socioeconomic mobility diminish (Cunha & Oliveira, 2021; Santos & Lima, 2017). These structural inequalities are not isolated determinants but components of intergenerational cycles that influence reproductive behavior, parenting challenges, and exposure to gendered social vulnerabilities.

From an analytical standpoint, the statistical modeling revealed that although the initial OLS model appeared to explain a large proportion of variance, this impression was misleading. After removing non-significant predictors and reassessing the model under more rigorous cross-validation, the generalization capacity reduced substantially. The strongest predictors of the number of children were age, family members, family income, and marital status (single), aligning with national-level evidence on social vulnerability among young and low-income mothers (Martins & Pereira, 2019; Mainardi & Bidoia, 2022; Cunha & Oliveira, 2021). The updated cross-validated coefficient of determination was approximately $\hat{R}^2 = 0.37$, indicating moderate explanatory power after controlling for overfitting. This correction was essential to prevent erroneous interpretations of the model's performance, especially given the small sample size.

Teenage pregnancy, a recurring element in vulnerable settings, reinforces long-term disadvantages: early motherhood increases the likelihood of having multiple children across the life course, intensifies financial hardship, and complicates the maintenance of stable relationships, often leading to union dissolution (Cunha & Oliveira, 2021; IBGE, 2021; Moura & Barbosa, 2018). These dynamics, strongly reflected in the significant predictors identified by the OLS model, highlight how reproductive outcomes are embedded within broader psychosocial and economic structures.

The decision tree analysis corroborated this interpretation. The full unpruned tree showed clear signs of overfitting - a common behavior of CART models in small datasets - where the model achieved nearly perfect training accuracy but poor generalization. After applying cost-complexity pruning, the optimal complexity parameter was $\alpha = 0.1137$, resulting in an improved test accuracy of 74%. The pruned tree retained only the most informative splits, primarily related to family size and age, reinforcing their central role in reproductive patterns. However, the ROC curve

indicated limited discriminative ability (AUC = 0.55), suggesting that although the pruned model is more stable, classification performance remains moderate. This is expected in small epidemiological samples with limited variability in the variables.

Overall, the convergence between the OLS and pruned decision-tree models indicates that the demographic determinants of the number of children among these maternal health outcomes are fundamentally shaped by social vulnerability. These findings have direct implications for maternal health policy. Interventions such as expanded access to prenatal education, reproductive-planning programs, community outreach by Family Health Strategy teams, and targeted support for young and single mothers could mitigate adverse outcomes. Strengthening policies that integrate income support, social protection, and educational opportunities for women may further reduce the long-term vulnerability associated with repeated pregnancies under precarious conditions.

Finally, the contrast between the high in-sample $\hat{R}^2$ values and the substantially lower cross-validated performance underscores the importance of cautious interpretation of statistical models, particularly in small population-based studies. For maternal-health applications, employing internal validation methods and avoiding reliance on apparent explanatory power is essential to ensure reliable, policy-relevant insights.

Conclusion

This study employed linear regression and a CART decision tree algorithm to examine the sociodemographic factors influencing the number of children among low-income mothers in Inimutaba, Minas Gerais. By examining demographic variables such as age, education level, family size, household income, marital status, and race/skin color, the modeling approach provided an overview of the structural conditions influencing fertility patterns in this population.

The results showed that family size, household income, and marital status (single) were consistently associated with the number of children across models. Although age also appeared as a relevant predictor in the initial OLS model, its statistical significance diminished after model refinement. Importantly, the high in-sample performance of the regression models did not translate into strong generalization capacity: internal validation using 10-fold cross-validation reduced the coefficient of determination to approximately $\hat{R}^2 = 0.37$, indicating substantial overfitting. This reinforces the need for cautious interpretation when working with small epidemiological datasets.

Similarly, the decision tree model demonstrated typical overfitting behavior when fully expanded. Cost-complexity pruning with an optimal parameter of $\alpha = 0.113$ improved test accuracy to 74%, but the modest AUC (0.55) indicates limited discriminative ability. Together, these findings suggest that while interpretable machine learning methods can support exploratory analyses of maternal-health contexts, their predictive performance remains constrained by sample size and variable distribution.

Overall, the study highlights the strong influence of socioeconomic vulnerability on reproductive patterns among the women evaluated. Factors such as low income, large household size, and unstable marital relationships reflect broader structural inequalities that shape maternal and child health in small municipalities. Future research should incorporate larger samples and additional psychosocial variables to improve model robustness. Expanding analyses to include methods such as random forests, artificial neural networks, or hierarchical Bayesian models may further clarify the complexity of fertility determinants in vulnerable populations.

These findings underscore the importance of strengthening social protection policies, reproductive-planning initiatives, and community-based health interventions targeting low-income women. Rigorous statistical modeling, combined with public health strategies, can support the better allocation of resources and promote equity in maternal and child health outcomes.

Acknowledgments

We would like to thank reviewers and editors for their comments.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Almeida, B. & Castro, M. Gravidez e juventude: desafios contemporâneos na saúde pública. *Revista Interface: Comunicação, Saúde, Educação* **23** (2019).
2. Arlot, S. & Celisse, A. A survey of cross-validation procedures for model selection. *Statistics Surveys* **4,** 40–79 (2010).
3. Babyak, M. A. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic Medicine* **66,** 411–421 (2004).
4. Barros, D. & Rocha, V. Educação, gênero e reprodução social em adolescentes grávidas. *Revista Educação & Sociedade* **41,** 1–18 (2020).
5. Breiman, L., Friedman, J., Olshen, R. & Stone, C. *Classification and Regression Trees* (Chapman and Hall/CRC, New York, 1984).
6. Campos, L. & Silva, R. Impactos da pobreza na primeira infância: uma revisão sistemática. *Revista Psicologia em Pesquisa* **14,** 35–43 (2020).
7. Costa, F. & Almeida, J. Desigualdade social e saúde materna: uma análise regional. *Revista Brasileira de Epidemiologia* **22,** e190046. doi:10.1590/1980-549720190046 (2019).
8. Cunha, R. & Oliveira, P. Gênero, trabalho e cuidado: desafios para a equidade nas famílias brasileiras. *Revista de Estudos Feministas* **29,** 1–14 (2021).
9. Da Cidadania, M. Indicadores de pobreza e vulnerabilidade social no Brasil. *Governo Federal* (2021).
10. Da Saúde, M. *Sistema de Informação sobre Nascidos Vivos (SINASC): Estatísticas de Nascimentos por Idade da Mã* https://datasus.saude.gov.br/nascidos-vivos-sinasc/. Acessado em 27 de novembro de 2025. 2020.
11. D'Agostino, R. B. & Pearson, E. S. Tests for normal distribution. *Biometrika* **60,** 613–622. doi:10.1093/biomet/60.3.613 (1973).
12. Freitas, J. & Santos, R. Vivências da gravidez e construção de identidade social. *Revista Psicologia: Teoria e Prática* **21,** 56–68 (2019).
13. Gonçalves, R. & Lima, T. Famílias chefiadas por mulheres e os desafios da desigualdade de renda. *Revista Brasileira de Estudos de População* **37,** e0120 (2020).
14. Harrell Jr, F. E. *Regression Modeling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis* 2nd ed. (Springer, 2015).
15. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* 2nd ed. (Springer, 2009).
16. IBGE. Estatísticas de Gênero: Indicadores sociais das mulheres no Brasil. *Instituto Brasileiro de Geografia e Estatística.* Disponível em: https://www.ibge.gov.br/estatisticas/sociais/populacao/20189-estatisticas-de-genero.html (2021).
17. IBGE. Síntese de Indicadores Sociais: uma análise das condições de vida da população brasileira. *Instituto Brasileiro de Geografia e Estatística.* Disponível em: https://biblioteca.ibge.gov.br/visualizacao/livros/liv101760.pdf (2022).

18. IPEA. Retrato das desigualdades de gênero e raça. *Instituto de Pesquisa Econômica Aplicada.* Disponível em: https://www.ipea.gov.br/retratos/ (2019).

19. Kuhn, M. & Johnson, K. *Applied Predictive Modeling* (Springer, 2013).

20. Loh, W.-Y. Classification and regression trees. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* **1,** 14–23 (2011).

21. Mainardi, P. H. & Bidoia, E. D. Fundamental concepts and recent applications of factorial statistical designs. *Brazilian Journal of Biometrics* (2022).

22. Martins, E. & Pereira, C. Gravidez e exclusão social: uma análise das mulheres em situação de vulnerabilidade. *Revista Katálysis* **22,** 204–212 (2019).

23. Moura, L. & Barbosa, A. C. Consequências sociais da gravidez na adolescência: desafios e perspectivas. *Revista Saúde e Sociedade* **27,** 755–766. doi:10.1590/S0104-12902018170301 (2018).

24. Nascimento, P. & Moreira, C. Infância, desigualdade social e políticas públicas no Brasil. *Revista Brasileira de Educação* **25** (2020).

25. Oliveira, C. S. & Fernandes, J. L. A importância do pré-natal para a saúde materna e neonatal: revisão integrativa. *Revista Brasileira de Enfermagem* **74,** e20200783. doi:10.1590/0034-7167-2020-0783 (2021).

26. Oliveira, C. & Lima, J. A trajetória das mães solo e o impacto na estrutura familiar. *Cadernos de Pesquisa* **50,** 80–100 (2020).

27. Pereira, C. A. B., Nakamura, L. R. & Rodrigues, P. C. Naive statistical analyses for COVID-19: application to data from Brazil and Italy. *Brazilian Journal of Biometrics* **39,** 158–176 (2021).

28. Rodrigues, A. R. & Nascimento, T. Saúde da mulher e maternidade no Brasil contemporâneo. *Revista Brasileira de Saúde Materno Infantil* **20,** 789–800. doi:10.1590/1806-93042020000300005 (2020).

29. Santos, F. & Lima, A. P. Gravidez precoce e vulnerabilidade social: reflexos na vida das adolescentes. *Revista da Escola de Enfermagem da USP* **51,** 1–8. doi:10.1590/S1980-220X2016041903252 (2017).

30. Silva, J. & Andrade, P. A pobreza como violação de direitos sociais. *Revista Direitos Fundamentais* **14,** 45–59 (2021).

31. Silva, M. C. & Gomes, E. Políticas públicas e saúde da mulher: uma análise crítica. *Revista Ciência & Saúde Coletiva* **24,** 1125–1133 (2019).

32. Silva, M. A. & Souza, L. B. Desigualdade de gênero e maternidade no Brasil. *Revista Estudos Feministas* **26,** e48412. doi:10.1590/1806-9584-2018v26n148412 (2018).

33. Souza, A. C. & Matos, F. Maternidade e relações sociais em contextos de pobreza. *Revista Ciência & Saúde Coletiva* **25,** 1927–1934 (2020).

34. Vieira, M. & Lopes, F. Ações de combate à pobreza no Brasil contemporâneo. *Revista de Políticas Públicas* **23,** 115–127 (2019).