Brazilian Journal of Biometrics, **44**, 1-16 (2026)
doi:10.28951/bjb.v44i3.973

BRAZILIAN JOURNAL OF BIOMΣTRICS
ISSN:2764-5290

ARTICLE

Machine learning approaches for modeling reproductive patterns on count data

Makina Venkata Lavanya*[1] and Begari Muniswamy[2]

[1]Assistant Professor, Department of Mathematics and Statistics, GITAM School of Science, GITAM (Deemed to be University), Visakhapatnam, Andhra Pradesh, India.
[2]Professor, Department of Statistics, Andhra University, Visakhapatnam, Andhra Pradesh, India.
*Corresponding author. Email: lmakina@gitam.edu;

(Received: August 26, 2025; Revised: May 20, 2026; Accepted: June 16, 2026; Published: September 22, 2026)
Section editor: Luiz Ricardo Nakamura.

Abstract
Reproductive patterns define the trends and behaviours related to childbirth, family size, and the timing of births within a group. They are affected by several variables, including social, economic, and cultural, along with biological components, providing them a crucial area of study in demography and health research. Analyzing reproductive patterns offers insights into societal health, family planning requirements, and population dynamics. The total number of children ever born (CEB) is a crucial metric in these patterns, reflecting a woman's lifelong fertility. This study analyses the factors affecting fertility among Andhra Pradesh's reproductive-age women, focusing on the socio-demographic impacts on the number of CEBs using information obtained from the fifth round of the National Family Health Survey (NFHS-5). This research applies machine learning techniques to identify key predictors of the number of CEB and compares model performance. Feature selection using the Boruta algorithm identified women's age, fertility preference, wealth index, marital status, type of cooking fuel, religion, caste, and residence as the most influential variables. Associations between predictors and fertility outcomes were further explored using chi-square tests. The Support Vector Machine emerged as a precise prediction model. The findings provide insights for demographic research and support evidence-based reproductive health policy planning.

Keywords: Children Ever Born (CEB); Machine Learning Models; NFHS-5; Predictive Analytics; Reproductive Patterns; Socio-Demographic Determinants.

1. Introduction

Understanding reproductive patterns is essential for formulating population policies and addressing health needs across varied socio-economic settings. Bongaarts (1978) defined the total number of children ever born (CEB) as the average number of children a woman has had; CEB is a key fertility metric reflecting reproductive outcomes shaped by socio-demographic and environmental factors. Sobotka *et al.* (2011) noted that global fertility patterns are changing due to advances in healthcare, economic shifts, and new societal norms. Analyzing indicators such as

CEB is therefore vital to understanding these trends. Such indicators clarify family size preferences and reproductive health, reflect socio-cultural contexts, and inform targeted policy development. Ram (2012) illustrated that India's socio-economic disparities and cultural diversity produce pronounced regional and social differences in fertility patterns.

Previous studies, including one conducted using Poisson regression models in Andhra Pradesh, have highlighted the complex relationship between fertility rates and various socio-demographic factors and analyzed the factors associated with the number of CEB among reproductive-age women. Lavanya *et al.* (2024) explored the determinants of fertility in Andhra Pradesh using the Poisson model, revealing complex relationships between fertility and socio-demographic variables among reproductive-aged women. However, while the Poisson model identifies significant predictors, it may be limited by issues related to data under- or over-dispersion. To address these limitations, Lavanya *et al.* (2024) applied the Hurdle Poisson Regression Model for Reproductive Patterns on Count Data, which addresses the challenges of modeling count data with structural zeros and under-dispersion, both of which are common in CEB data and reveals that factors such as age and fertility preference have a significant influence on CEB, emphasizing the importance of creating an enabling environment for informed reproductive decisions. Muniswamy and Lavanya (2025) extended this analysis by applying Zero-Truncated Poisson (ZTP) and Zero-Truncated Generalized Poisson (ZTGP) models to CEB data, specifically focusing on women who have experienced at least one childbirth. The ZTP model has been suitable for datasets that do not include zero values, enabling the research to focus on the factors influencing childbirth among women who are not childless and identifying religion, wealth, place of residence, and fertility preferences as essential determinants.

Traditional statistical models, such as Poisson and Zero-Truncated Poisson (ZTP) regression, assume predefined functional relationships between independent variables (e.g., women's age, wealth index) and fertility outcomes. These models primarily capture linear associations, which may oversimplify real-world fertility patterns. ML models, however, can identify non-linear relationships and interactions, improving predictive accuracy. Decision Trees & Random Forests, automatically detect breakpoints where relationships change, avoiding assumptions of linearity. Gradient Boosting Models (GBM, XGBoost) can iteratively refine predictions, capturing subtle variations and interactions. Support Vector Machines (SVM) can map the non-linearly separable data into a higher-dimensional space to uncover hidden fertility trends. Uddin *et al.* (2024) demonstrated that the advanced machine learning methodologies for fertility determinants revealed that the Support Vector Machine was Bangladesh's best fertility prediction model. Similarly, Balogun *et al.* (2018) effective machine learning algorithms were used to predict the chance of infertility among Nigerian women. Bitew *et al.* (2020), Mfateneza *et al.* (2022), and Raef and Ferdousi (2019) have done several studies highlighting the uses of machine learning methodologies across several demographic fields of study.

The NFHS as defined by IIPS and ICF (2021), is among India's largest population health surveys, providing detailed data on reproductive health, family planning, nutrition, and socio-economic conditions in both urban and rural contexts. Basumatary (2023) used NFHS data to examine the effects of marital status, economic conditions, age, and cultural norms on fertility patterns. Prior research shows that socio-economic status is associated with lower fertility rates, with urbanization, wealth, and education particularly associated with lower fertility. Moreover, variables such as religion, cooking fuel type, and place of delivery have emerged as significant determinants of fertility trends, with rural areas often registering higher birth rates attributed to socio-cultural and economic influences.

Traditional statistical methods, such as Poisson regression, have been widely used in fertility research. Koshy *et al.* (2022) reviewed AI-based approaches for infertility prediction and supported the use of advanced models in reproductive health. Similarly, Liu *et al.* (2023) developed ML models to predict live births in IVF patients, demonstrating the accuracy of these techniques. Rahman *et al.* (2021) applied ML algorithms to infant mortality data in Bangladesh, demonstrating their utility in fertility and public health research. Machine learning methods are increasingly significant in public health, as they identify and interpret nonlinear relationships and variable interactions.

With a large dataset of over 10,522 observations from NFHS-5, this study focuses on analyzing socio-demographic and health factors that influence fertility among women in Andhra Pradesh. Although the number of Children Ever Born (CEB) is naturally a count variable, this study reformulates the problem as a multiclass classification task. Women are categorized into fertility groups based on the number of children ever born (0, 1, 2, 3, and 4 or more). This formulation enables the application of supervised machine learning classification algorithms and facilitates the evaluation of predictive performance under class-imbalance conditions. Therefore, the objective of this study is to classify fertility levels and identify key socio-demographic determinants associated with different fertility categories.

2. Data and Methodology

This analysis employs data from the fifth National Family Health Survey (NFHS-5), carried out between 2019 and 2021, focusing on nutrition, health, and population across all states and union territories in India. The NFHS offers essential insights into reproductive and child health, encompassing socio-economic characteristics, fertility rates, early child mortality, family planning, water and sanitation, nutritional status, child immunization, gender-based violence, women's empowerment, and specific non-communicable diseases. The NFHS National Report, co-prepared by the International Institute for Population Sciences (IIPS) in Mumbai and the Statistics Division of the Ministry of Health and Family Welfare (MoHFW), will significantly enhance the country's demographic and health database. Comprehensive documentation of the sample and survey technique associated with the Demography and Health Surveys (DHS) has been released, and all survey data categorized as secondary data is publicly accessible. The NFHS-5 (2019–21) dataset was obtained from the DHS repository and restricted to women aged 15–49 years. Observations containing missing values in the response or predictor variables were excluded, resulting in a final dataset of 10,522 complete records. All categorical predictor variables were encoded using one-hot encoding. The dataset was divided into 80% for training and 20% for testing. To improve predictive performance and reduce overfitting, hyperparameters for each machine learning model were tuned using five-fold cross-validation. Because the response variable (CEB) was imbalanced, the Synthetic Minority Over-sampling Technique was applied to the training dataset.

The main study variable, CEB, measures the total number of children born to a woman at the time of the survey.

Table 1. Independent Variables

Variables	Description
X_1	Place of Residence (Urban, Rural)
X_2	Religion (Hindu, Muslim, Christian)
X_3	Type of Cooking Fuel (Electricity, LPG, Natural gas, Biogas, Kerosene, Coal & lignite, Charcoal, Wood, Straw/Shrubs/Grass, Agricultural crop, Animal dung, Other)
X_4	Wealth Index Combined (Poorest, Poorer, Middle, Richer, Richest)
X_5	Place of Delivery (Home, Public, Private)
X_6	Women Age (15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49)
X_7	Current Marital Status (Single, Married, Widowed, Divorced)
X_8	Fertility Preference (Have another, Undecided, No more, Sterilized, Declared infecund, Never had sex)
X_9	Caste (Schedule Caste, Schedule Tribe, OBC)
X_{10}	Husband Age (18-27, 28-37, 38-47, 48-57, 58 & above)

Table 1 presents the independent variables. The variables associated with CEB, namely women's age, fertility preference, wealth index, marital status, caste, and place of residence, are consistent with earlier demographic studies on Indian fertility, confirming that socio-economic and cultural factors strongly influence reproductive behavior.

This study seeks to investigate risk factors associated with the number of CEB to understand reproductive patterns in count data. The methodology uses several machine learning techniques for classification, including Random Forests (RF), Decision Trees (DT), k-Nearest Neighbours (KNN), Support Vector Machine (SVM), XGBoost, LightGBM, and Neural Networks (NN). An 80% training set and a 20% testing set have been generated from the dataset. The training data set was used to develop machine-learning models, while the test data set was used to evaluate how effectively these approaches performed. The complete dataset was then used to predict using the trained models. The testing and comparison of the models for prediction involved metrics, including the macro- and weighted-average scores from the Confusion Matrix, Accuracy, Recall, Precision, F1 Score, and Area Under the Receiver Operating Characteristics Curve (AUROC).

Hyperparameter tuning for all machine learning models was performed using five-fold cross-validation to improve predictive performance and reduce overfitting. The RF model was implemented using multiple decision trees with bootstrap aggregation, while the DT model utilized recursive partitioning for classification. The KNN model was optimized using the optimal k value selected via cross-validation. The SVM model employed a radial basis function (RBF) kernel to capture non-linear relationships in the data. The XGBoost and LightGBM models were tuned using parameters such as learning rate and tree depth. The NN model included hidden layers with nonlinear activation functions for multiclass classification.

Chi-square tests were conducted to explore the association between each predictor variable and the number of CEB. Furthermore, the Boruta algorithm was used to identify the most important characteristics of CEB. This work used Python for machine learning, and the Boruta package in R was used to select the most informative features.

3. Results and Discussion

A total of 10,522 women aged 15 to 49 years were involved in the study, with the highest frequency of respondents reporting having two children, with 4,627 women in this category, which constitutes a significant portion of the sample. The observed count was 2,435, representing 23.1%

of the sample. A proportion of women (2,435) had no children at the time of the survey, indicating that they had not experienced childbirth by the time of the survey. Fewer women reported having three (1,682) or four children (578), indicating a decreasing trend in the number of children as the count increases, as shown in Table 2. This distribution suggests that while most women fall within the lower to middle range of childbearing, there is a substantial minority with no children, which could be linked to various socio-economic or health factors. Understanding this distribution is crucial for identifying fertility trends, informing reproductive health policies, and addressing the region's socio-demographic factors influencing fertility choices.

Table 2. Frequency Distribution of CEB

CEB	0	1	2	3	4
Frequency	2435	1200	4627	1682	578

Table 3 presents the absolute frequencies of women across each predictor category, along with the corresponding percentage distribution for each CEB level, with the chi-square and p-values, which highlight variations in fertility patterns across socio-demographic groups. Rural women have higher fertility rates than urban women, influenced by traditional pronatalist norms, limited access to family planning services, and the economic benefits of larger families in agricultural settings. Religion also plays a role, with Hindus generally having higher fertility, likely influenced by religious and cultural practices. Women using LPG as a cooking fuel tend to have higher fertility, possibly linked to the economic status and lifestyle associated with this fuel. Wealthier women, especially those in the richest wealth index category, show lower fertility, which aligns with trends of delayed childbearing and fewer children in wealthier groups. Fertility preferences also show marked differences, with women wanting no more children or sterilized women having lower fertility rates. In contrast, women wishing to have another child have higher fertility, highlighting the role of fertility intentions. Women's age and marital status also significantly influence fertility, with younger, married women having more children, while widowed and divorced women tend to have fewer children.

The data indicate that a significant proportion of women residing in rural areas (71%) had more children, particularly in the 3 and 4 categories, compared with their urban counterparts (29%). The statistical analysis reveals a significant difference ($\chi^2 = 72.312$, $p < 0.0001$), suggesting that rural women tend to have more children. Hindu women represent the predominant demographic, with 83.4% exhibiting a higher prevalence of having two children. Muslim and Christian women show a higher tendency to have children in the "4" category compared to Hindus, with a statistically significant difference observed between the religious groups ($\chi^2 = 57.685$, $p < 0.0001$). Women using LPG (82.8%) were more likely to have two children than those using wood (13.5%). Users of wood and charcoal had higher rates of having 4 or more children, indicating significant variation in reproductive patterns based on cooking fuel ($\chi^2 = 88.343$, $p < 0.0001$). Women in the Poorest and Poorer categories were more likely to have four or more children. In contrast, those in the Richer and Richest categories had a lower overall number of children. The observed difference is statistically significant ($\chi^2 = 191.014$, $p < 0.0001$). Women who delivered at home (44.1%) and in public facilities (48.1%) generally had more children, whereas those who delivered in private hospitals (7.8%) had fewer. The distribution demonstrates statistical significance ($p < 0.0001$). Women aged 35-49 were more likely to have three or more children, whereas younger women (15-24) primarily fell into the 0 or 1 child category. The correlation between age and fertility is statistically significant ($\chi^2 = 6091.236$, $p < 0.0001$).

Table 3. Descriptive analysis of background predictor variables

Variables	Number of CEB							
	Total	0	1	2	3	4	χ2	p-value
Place of residence								
Urban	3049(29.0)	814(33.4)	371(30.9)	1353(29.2)	367(21.8)	144(24.9)	72.312	<0.0001
Rural	7473(71.0)	1621(66.6)	829(69.1)	3274(70.8)	1315(78.2)	434(75.1)		
Religion								
Hindu	8779(83.4)	1999(82.1)	1026(85.5)	3962(85.6)	1350(80.3)	442(76.5)		
Muslim	807(7.7)	192(7.9)	80(6.7)	313(6.8)	154(9.2)	68(11.8)	57.685	<0.0001
Christian	936(8.9)	244(10.0)	94(7.8)	352(7.6)	178(10.6)	68(11.8)		
Type of cooking fuel								
Electricity	83(0.8)	26(1.1)	7(0.6)	33(0.7)	12(0.7)	5(0.9)		
LPG	8717(82.8)	2025(83.2)	1005(83.8)	3913(84.6)	1341(79.7)	433(74.9)		
Biogas	9(0.1)	2(0.1)	4(0.3)	2(0.0)	1(0.1)	0(0.0)		
Kerosene	18(0.2)	7(0.3)	3(0.3)	3(0.1)	3(0.2)	2(0.3)		
Coal, lignite	19(0.2)	3(0.1)	2(0.2)	11(0.2)	2(0.1)	1(0.2)		
Charcoal	129(1.2)	27(1.1)	14(1.2)	45(1.0)	28(1.7)	15(2.6)	88.343	<0.0001
Wood	1416(13.5)	312(12.8)	156(13.0)	561(12.1)	272(16.2)	115(19.9)		
Straw/Shrubs/Grass	46(0.4)	11(0.5)	2(0.2)	21(0.5)	8(0.5)	4(0.7)		
Agricultural crop	81(0.8)	22(0.9)	7(0.6)	34(0.7)	15(0.9)	3(0.5)		
Animal dung	2(0.0)	0(0.0)	0(0.0)	2(0.0)	0(0.0)	0(0.0)		
Other	2(0.0)	0(0.0)	0(0.0)	2(0.0)	0(0.0)	0(0.0)		
Wealth index combined								
Poorest	427(4.1)	91(3.7)	53(4.4)	150(3.2)	85(5.1)	48(8.3)		
Poorer	2022(19.2)	466(19.1)	217(18.1)	804(17.4)	379(22.5)	156(27.0)		
Middle	3369(32)	736(30.2)	344(28.7)	1500(32.4)	590(35.1)	199(34.4)	191.014	<0.0001
Richer	3136(29.8)	756(31.0)	358(29.8)	1405(30.4)	476(28.3)	141(24.4)		
Richest	1568(14.9)	386(15.9)	228(19.0)	768(16.6)	152(9.0)	34(5.9)		
Place of delivery								
Home	4641(44.1)	1090(44.8)	523(43.6)	2051(44.3)	712(42.3)	265(45.8)		
Public	5064(48.1)	1139(46.8)	589(49.1)	2227(48.1)	838(49.8)	271(46.9)	6.600859	0.5802
Private	817(7.8)	206(8.5)	88(7.3)	349(7.5)	132(7.8)	42(7.3)		
Women age								
15-19	1267(12)	1174(48.2)	70(5.8)	22(0.5)	1(0.1)	0(0.0)		
20-24	1386(13.2)	683(28.0)	275(22.9)	362(7.8)	62(3.7)	4(0.7)		
25-29	1697(16.1)	259(10.6)	262(21.8)	907(19.6)	222(13.2)	47(8.1)		
30-34	1520(14.4)	103(4.2)	172(14.3)	916(19.8)	265(15.8)	64(11.1)	6091.236	<0.0001
35-39	1664(15.8)	81(3.3)	153(12.8)	1031(22.3)	310(18.4)	89(15.4)		
40-44	1382(13.1)	72(3.0)	122(10.2)	717(15.5)	338(20.1)	133(23.0)		
45-49	1606(15.3)	63(2.6)	146(12.2)	672(14.5)	484(28.8)	241(41.7)		
Current marital status								
Single	1792(17)	1750(71.9)	13(1.1)	19(0.4)	6(0.4)	4(0.7)		
Married	7943(75.5)	603(24.8)	1053(87.8)	4262(92.1)	1528(90.8)	497(86.0)	6874.102	<0.0001
Widowed	679(6.5)	42(1.7)	102(8.5)	321(6.9)	141(8.4)	73(12.6)		
Divorced	108(1.0)	40(1.6)	32(2.7)	25(0.5)	7(0.4)	4(0.7)		
Fertility preference								
Have another	1376(13.1)	517(21.2)	602(50.2)	196(4.2)	46(2.7)	15(2.6)		
Undecided	238(2.3)	56(2.3)	55(4.6)	98(2.1)	23(1.4)	6(1.0)		
No more	745(7.1)	94(3.9)	226(18.8)	291(6.3)	94(5.6)	40(6.9)	10161.57	<0.0001
Sterilized	6179(58.7)	33(1.4)	242(20.2)	3930(84.9)	1466(87.2)	508(87.9)		
Declared infecund	298(2.8)	70(2.9)	71(5.9)	102(2.2)	46(2.7)	9(1.6)		
Never had sex	1686(16)	1665(68.4)	4(0.3)	10(0.2)	7(0.4)	0(0.0)		
Caste								
Schedule Caste	3376(32.1)	797(32.7)	391(32.6)	1436(31.0)	535(31.8)	217(37.5)		
Schedule Tribe	1032(9.8)	231(9.5)	135(11.3)	440(9.5)	159(9.5)	67(11.6)	20.32046	0.0092
OBC	6114(58.1)	1407(57.8)	674(56.2)	2751(59.5)	988(58.7)	294(50.9)		
Husband age								
18-27	2155(20.5)	488(20.0)	251(20.9)	966(20.9)	337(20.0)	113(19.6)		
28-37	1977(18.8)	464(19.1)	202(16.8)	875(18.9)	333(19.8)	103(17.8)		
38-47	2188(20.8)	504(20.7)	265(22.1)	954(20.6)	345(20.5)	120(20.8)	9.640516	0.8847
48-57	2189(20.8)	490(20.1)	262(21.8)	959(20.7)	353(21.0)	125(21.6)		
58 & above	2013(19.1)	489(20.1)	220(18.3)	873(18.9)	314(18.7)	117(20.2)		

Married women exhibited the highest fertility rates, with 75.5% having children, particularly in the two and three-child categories, whereas single women predominantly reported having no children. This factor is significant ($\chi^2 = 6874.102$, $p < 0.0001$). Sterilized women constituted 58.7% of the sample and were predominantly represented in the two- and three-child categories; they also constituted 87.9% of women in the four-child category. The observed difference is highly significant ($\chi^2 = 10161.57$, $p < 0.0001$). Women from Scheduled Castes and Scheduled Tribes exhibited marginally elevated fertility rates in categories 3 and 4, demonstrating a statistically significant difference based on caste ($p < 0.0001$). Women with husbands aged 48-57 had the most children in the fourth category. The correlation between the husband's age and fertility is statistically significant ($p < 0.0001$). This study demonstrates that socio-demographic factors, including residence, religion, fuel type, wealth, marital status, and age, significantly affect reproductive patterns among women.

3.1 Selection of the Best Features

Figure 1 illustrates the Boruta algorithm, confirming that 8 of the predictor variables tested were identified as key predictors of the number of CEB. The most important factors in predicting the number of CEB include fertility preference, current marital status, women's age, wealth index, type of cooking fuel, religion, residence, and caste. Fertility preferences are the strongest determinant of CEB, highlighting family planning intentions. Marital Status displays that married women exhibit higher fertility rates than single or widowed women. Women’s age directly impacts fertility, with declining rates after age 35. Wealth index showed that economic stability correlates with lower CEB due to education and healthcare access. The type of cooking fuel is a proxy for household wealth and influences reproductive behavior. Cultural and religious beliefs shape fertility norms. Rural women tend to have more children than urban women. Caste affects fertility choices through economic and cultural constraints.

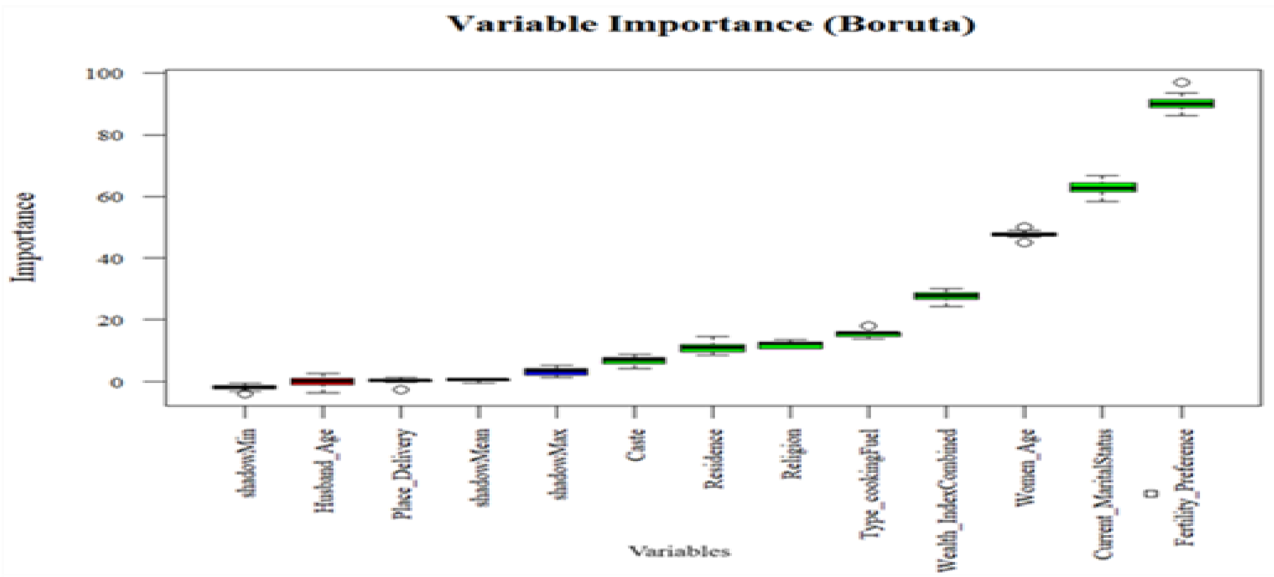

Figure 1. Feature selection using Boruta Algorithm.

Figure 2 illustrates the importance scores of the most influential variables, highlighting their relative impact on the number of CEB. Fertility preference is the most significant factor, with current marital status and women's age serving as the strongest determinants of the number of

CEB, thereby influencing reproductive patterns. Factors such as wealth, cooking fuel, religion, and residence exhibit moderate impacts, whereas caste remains statistically significant despite having the least influence.

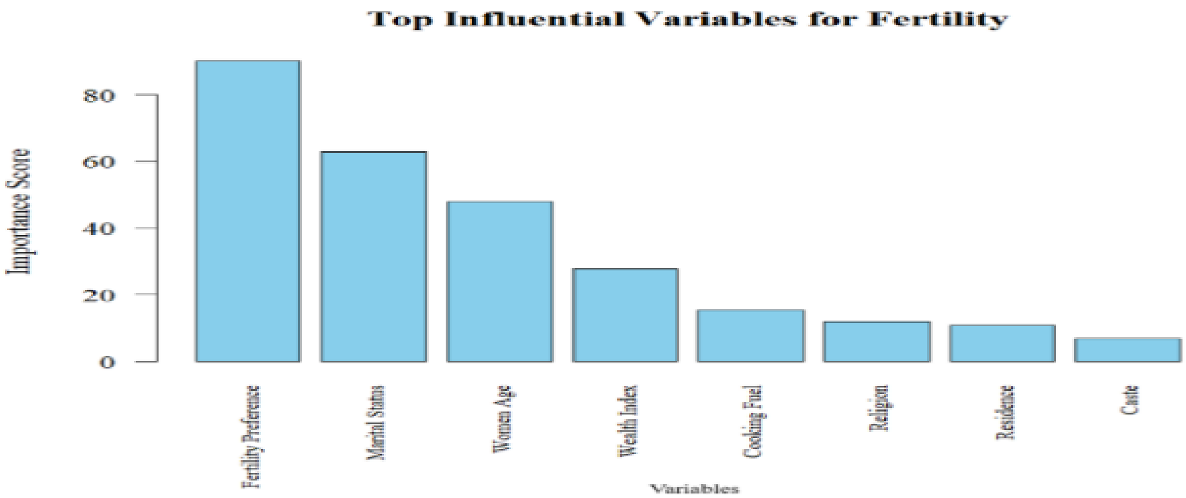

Figure 2. Influential variables for fertility.

Table 4 displays the Root Mean Square Error (RMSE) values obtained from the predicted class labels and the observed fertility categories. Although the study is formulated as a multiclass classification problem, RMSE is reported only as a supplementary numerical error indicator to show the average deviation between predicted and observed class values. However, since the number of Children Ever Born (CEB) was reformulated as a multiclass classification problem, the machine learning models were trained to predict fertility categories rather than directly minimize numerical prediction error. Therefore, classification-based evaluation metrics such as accuracy, precision, recall, balanced accuracy, and macro-averaged F1-score are considered the primary criteria for model comparison.

Table 4. Root Mean Squared Error (RMSE) values of machine learning models

Model	RF	DT	KNN	SVM	XGBoost	LightGBM	NN
RMSE	0.764	0.777	0.818	0.726	1.285	1.282	1.283

3.2 Predicting the Number of CEB Using Machine Learning Approaches

The following machine learning techniques were used to construct a predictive model for the number of CEB: Random Forests (RF), Decision Trees (DT), K-Nearest Neighbours (KNN), Support Vector Machine (SVM), XGBoost, LightGBM, and Neural Network (NN) methods. Each predictive model was trained on 80% of the dataset and tested on the remaining 20%. The present research describes several predictive models' performance metrics, including accuracy, recall, precision, and F1 score. These metrics are evaluated using two averaging methods: weighted average and macro average.

The RF model demonstrated moderate classification performance in predicting fertility categories. Although the model achieved an overall accuracy of 60%, greater emphasis was placed on balanced accuracy (43%) and macro-averaged F1-score (43%), as these metrics provide a more reliable evaluation under class imbalance conditions. The weighted average recall, precision, and F1-score were 60%, 56%, and 58%, respectively, indicating satisfactory performance for the majority of fertility classes. However, the comparatively lower macro-average measures suggest reduced predictive performance for minority fertility categories, particularly higher-order births. The RF model correctly classified 408 women with no children, 90 with one child, 722 with two children, 47 with three children, and 6 with four or more children. Misclassification across minority classes reflects the demographic heterogeneity and lower frequency of higher fertility outcomes observed in the dataset (Table 5).

The DT model predicted reproductive patterns with an accuracy of 58%, which is relatively low. The DT model demonstrated limited effectiveness in predicting fertility, achieving a weighted average recall of 58%, a precision of 54%, and an F1 score of 56%. The macro average recall, precision, and F1 scores were 42%. The DT model provided precise predictions regarding the number of CEB: 413 with no children, 74 with 1 child, 661 with 2 children, 58 with 3 children, and 7 with 4 or more children. The prediction inaccurately identified 892 children from various classes or groups (Table 5).

The KNN model predicted reproductive patterns with an accuracy of 62%. The KNN model exhibited a balanced performance in predicting fertility, achieving a weighted average recall of 62%, a precision of 56%, and an F1 score of 58%. Furthermore, it shown a macro average recall of 44%, a precision of 45%, and an F1 score of 43%. The KNN model provided precise predictions regarding the number of CEB: 407 individuals without children, 91 with one child, 747 with two children, 47 with three children, and 4 with four or more children. The estimation of 809 children born to various classes or groups has been provided inaccurately (Table 5).

The SVM model demonstrated strong classification performance in predicting fertility categories. Although the model achieved an overall accuracy of 65%, balanced accuracy (44%) and macro F1-score (41%) were considered more appropriate measures due to class imbalance. The weighted average recall, precision, and F1-score were 65%, 53%, and 57%, respectively. The model effectively classified dominant fertility categories but showed lower performance for minority fertility groups, particularly higher-order births, reflecting an imbalance in fertility outcomes. The SVM model accurately predicted the number of CEB: 392 as having no children, 109 as having one child, and 870 as having two children. It has inaccurately predicted 734 children born to various classes or groups (Table 5).

The XGBoost model showed balanced classification performance across fertility categories based on balanced accuracy and macro-averaged evaluation measures. The XGBoost model demonstrated balanced performance in predicting fertility, achieving a weighted average recall of 64%, a precision of 59%, and F1 score of 59%. Additionally, the macro average recall was 45%, precision was 49%, and F1 score was 44%. The XGBoost model accurately predicted the number of CEB: 402 individuals with no children, 97 with one child, 795 with two children, 39 with three children, and 5 with four or more children. It has been inaccurately estimated that 767 children were born across various classes or groups (Table 5).

The LightGBM model predicted reproductive patterns with an accuracy of 64%. The LightGBM model showed a balanced performance in predicting fertility, achieving a weighted average recall of 64%, precision of 59%, and F1 score of 59%. Additionally, the macro average recall was 45%, precision was 49%, and F1 score was 44%. The LightGBM model accurately predicted the number of CEB: 399 individuals with no children, 104 with one child, 817 with two children, 30 with three children, and 3 with four or more children. It has inaccurately predicted 752 children born to various classes or groups (Table 5).

The NN model demonstrated a predictive accuracy of 65% regarding reproductive patterns. The NN model demonstrated general effectiveness in predicting fertility, achieving a weighted average recall of 65%, precision of 57%, and F1 score of 57%. Furthermore, it revealed a macro average recall of 44%, a precision of 48%, and an F1 score of 41%. The NN model provided precise predictions for the number of CEB: 397 individuals without children, 104 with a single child, 857 with two children, and 2 each with three and four or more children. It has made an inaccurate prediction regarding the birth of 743 children across various categories or groups (Table 5).

Table 5 presents the performance comparison of machine learning models using classification-based evaluation metrics, including accuracy, balanced accuracy, weighted average precision, recall, and F1-score, as well as macro-averaged precision, recall, and F1-score. Balanced accuracy is included to provide a more reliable evaluation in the presence of class imbalance across fertility categories, as it computes the average recall over all classes.

Table 5. Predictive models of the performance of a number of CEB

Predictive Models Performances			Weighted Average			Macro Average		
	Accuracy	Balanced Accuracy	Recall	Precision	F1 score	Recall	Precision	F1 score
RF	60%	43%	60%	56%	58%	43%	44%	43%
DT	58%	42%	58%	54%	56%	42%	42%	42%
KNN	62%	44%	62%	56%	58%	44%	45%	43%
SVM	65%	44%	65%	53%	57%	44%	39%	41%
XGBoost	64%	45%	64%	59%	59%	45%	49%	44%
LightGBM	64%	45%	64%	59%	59%	45%	49%	44%
NN	65%	44%	65%	57%	57%	44%	48%	41%

Because the response variable (CEB) has multiple classes (0, 1, 2, 3, and 4) and is highly imbalanced, overall accuracy alone is not an appropriate measure of predictive performance. Therefore, balanced accuracy and macro-averaged F1-score were used for model comparison, as these metrics evaluate the performance of each fertility category equally and provide a more reliable assessment. The results indicate that although several models achieve similar accuracy levels, balanced accuracy provides a more robust evaluation by accounting for class-wise prediction ability. The relatively low F1-scores suggest that minority fertility classes are not being predicted accurately, indicating the need for improved preprocessing and class balancing. The models showed broadly similar classification performance, although differences were observed across balanced accuracy and macro-averaged F1-score.

Figure 3 displays the confusion matrices for several models: RF, DT, KNN, SVM, XGBoost, LightGBM, and NN. Each model exhibits differing performance, with RF and DT achieving robust true positive rates across classes; however, DT experiences marginally higher misclassification rates. KNN exhibits competitive accuracy but tends to misclassify certain instances, particularly those in lower-frequency classes. SVM shows proficiency in classifying a single class while exhibiting difficulties with the other fertility categories. Although XGBoost and LightGBM perform strongly in dominant classes, they exhibit certain misclassifications. Neural networks demonstrate significant predictive capability overall, although their accuracy may be diminished in specific cases.

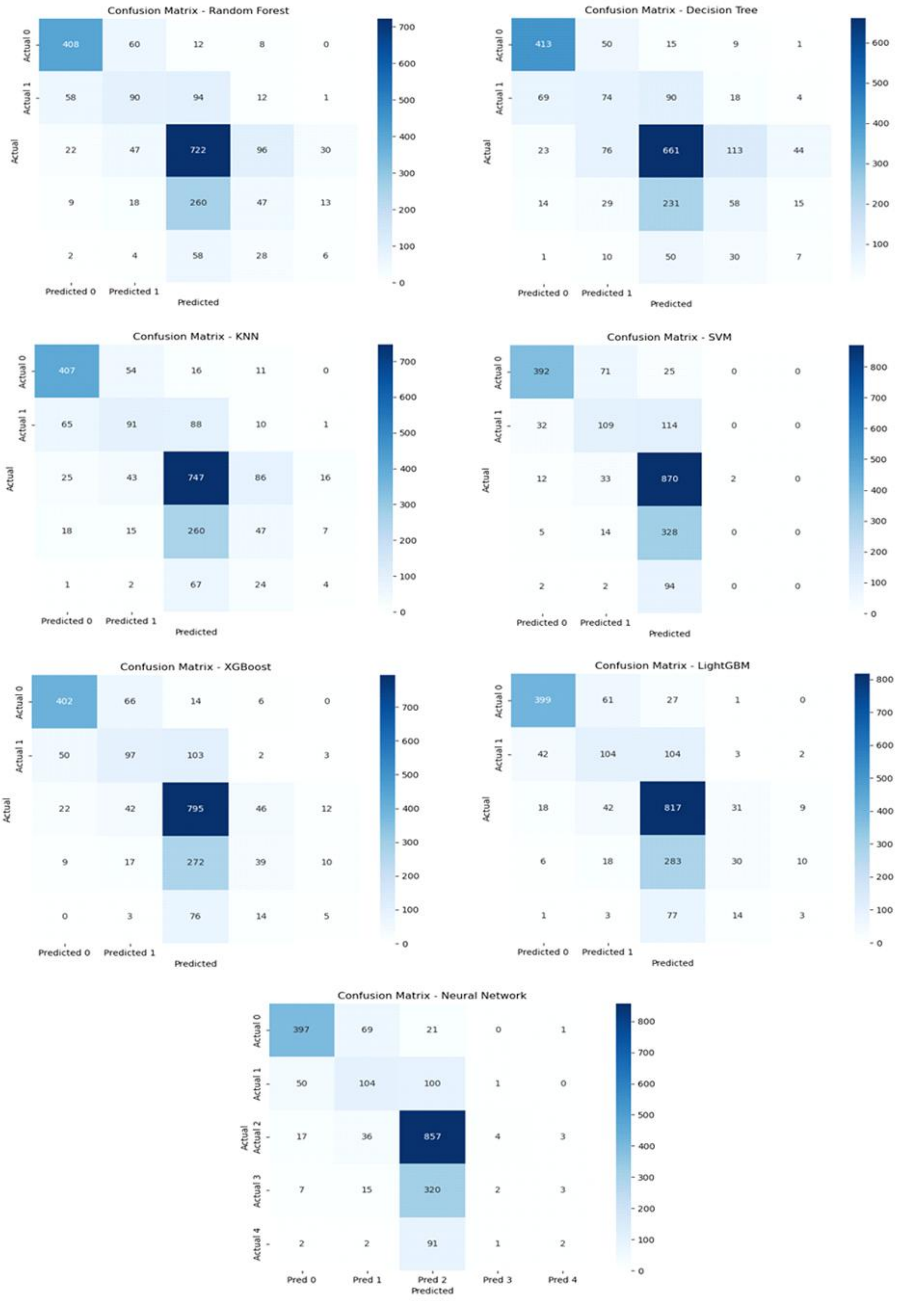

Figure 3. Confusion matrix of classification results.

Figure 4 presents the ROC curves demonstrating the performance of several machine learning models, including RF, DT, KNN, SVM, XGBoost, LightGBM, and NN, in predicting the number of CEB for fertility outcomes. The ROC curve illustrates the relationship between the true positive rate (sensitivity) and the false positive rate (1-specificity) across different threshold values. The Random Forest (RF) model exhibited moderate predictive accuracy, displaying strengths in specific scenarios while demonstrating variability in others. The Decision Tree (DT) model demonstrated effective classification capabilities but exhibited constraints in sustaining consistent overall accuracy. The K-Nearest Neighbours (KNN) model demonstrated satisfactory accuracy, especially in cases where data patterns were more discernible. The Support Vector Machine (SVM) model demonstrated significant predictive capability, exhibiting high accuracy across various cases, indicating its reliability for this analysis. XGBoost and LightGBM yielded varied outcomes; XGBoost showed strong performance in specific contexts, whereas LightGBM exhibited less reliability, suggesting difficulties in managing difficult classification challenges. The Neural Network (NN) model emerged as a leading performer, effectively determining complex patterns in the data and demonstrating high predictive accuracy, positioning it as a viable option for complex fertility prediction tasks.

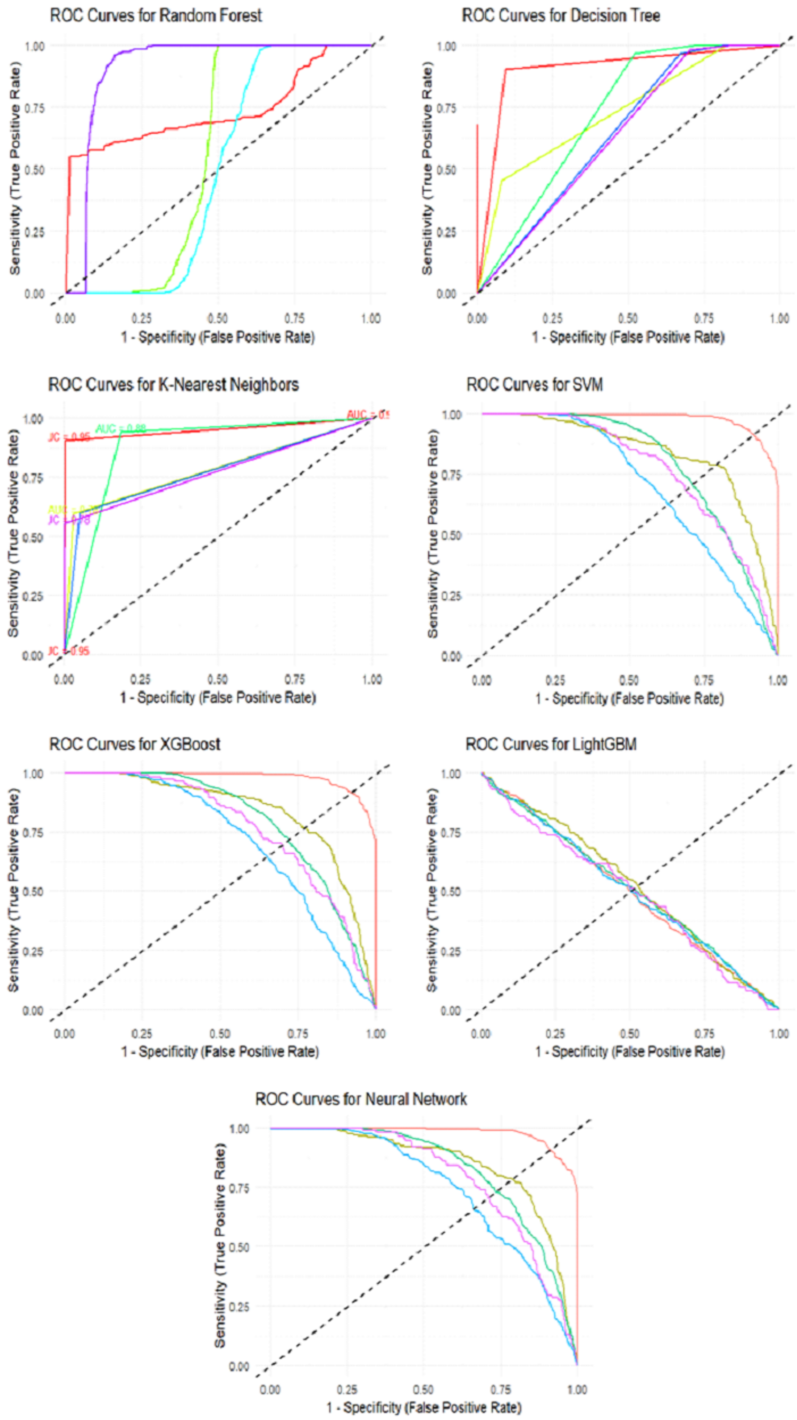

Figure 4. ROC to predict fertility using RF, DT, KNN, SVM, XGBoost, LightGBM, and NN.

Figure 5 presents a graphical representation of characteristics found by an SVM classifier, facilitating the distinction and emphasis of predictor factors that influence the number of CEB and, hence, informing reproductive patterns in AP. The SVM method in machine learning identifies the features influencing CEB, including fertility preference, marital status, women's age, wealth index, caste, and residence.

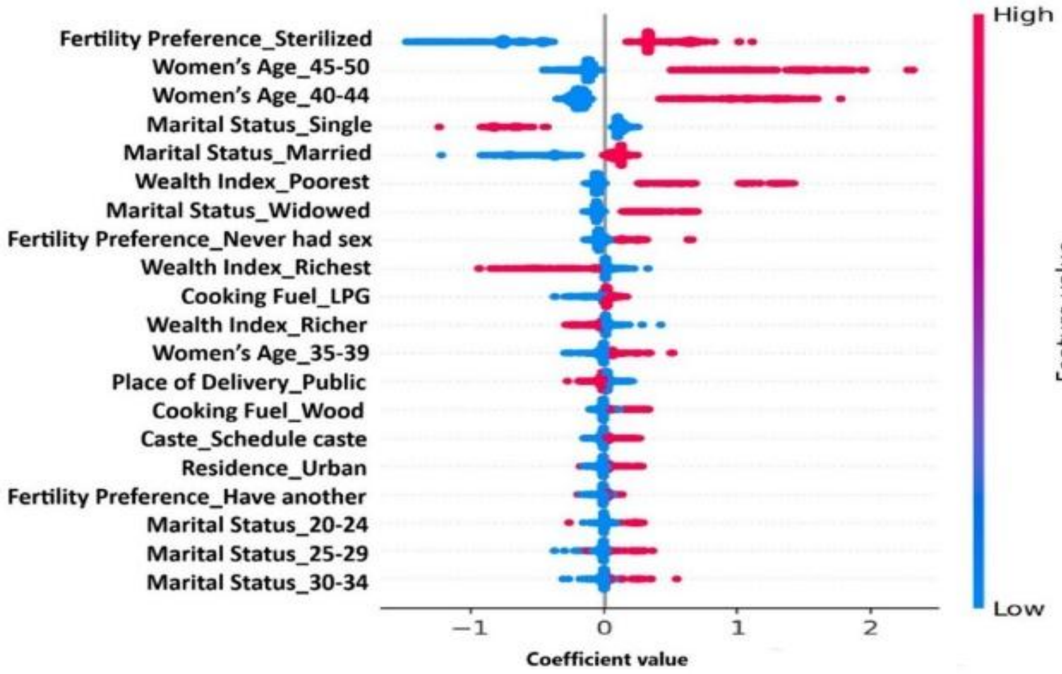

Figure 5. Important features visualization using ML SVM model.

This research examines reproductive patterns among women aged 15 to 49 in Andhra Pradesh, analyzing data from 10,522 participants in the fifth round of the National Family Health Survey (NFHS-5) regarding the number of CEB. The findings indicate a notable discrepancy between the observed and expected counts of women with children, indicating that a proportion of women had not experienced childbirth by the time of the survey and a societal trend towards smaller family sizes, specifically a preference for two children. The model showed better prediction for the more frequent fertility categories but had limited predictive performance for the less frequent higher-order fertility categories. The observed variation in predictive performance across fertility categories may be attributed to underlying demographic and socio-economic heterogeneity, as reported in previous studies on fertility behavior in developing regions. In particular, lower predictive accuracy for higher-order births reflects the relatively smaller sample size and greater variability in these categories. These findings are consistent with existing literature, which highlights the challenges of modeling rare fertility outcomes and emphasizes the importance of socio-economic and cultural determinants in shaping reproductive behavior. Statistical analyses reveal significant associations between CEB-related outcomes and various socio-demographic factors. Rural women tend to have larger families, and religious differences affect fertility preferences. Economic factors, including the cooking fuel used, influence family size, with women using LPG typically having smaller families. Furthermore, age and marital status are significant factors, as older and married women exhibit higher fertility rates. Sterilization status illustrates the relationship between fertility intentions and family planning practices.

This research utilizes Machine Learning methodologies, indicating that variables including residence, religion, cooking fuel, wealth index, women's age, marital status, fertility preference, and caste as significant predictors of fertility, as identified through the Boruta algorithm's feature selection. However, applying the conventional chi-square test revealed that all eight variables were significant.

The results assessed the performance of various machine learning models, demonstrating that these methods can effectively predict fertility categories related to the number of CEB. The models showed broadly similar classification performance, although differences were observed across balanced accuracy and macro-averaged F1-score. However, predictive performance varied across fertility categories and models. The SVM model identified six critical factors, namely fertility preference, marital status, women's age, wealth index, caste, and residence, as important factors associated with reproductive patterns.

Although the number of CEB is naturally a count variable, the present study reformulated the outcome into fertility categories to facilitate multiclass classification and address class imbalance issues. This approach improves interpretability and enables the application of advanced machine learning classification algorithms for identifying fertility groups. However, categorization may also result in some loss of numerical information compared with traditional count-data regression models such as Poisson or Zero-Truncated Poisson regression. While classical count models remain valuable for parameter interpretation and direct estimation of fertility counts, the classification framework adopted in this study primarily focuses on predictive discrimination across fertility categories. Therefore, both approaches may be considered complementary in fertility research.

4. Conclusion

This study applied several machine learning techniques to predict the number of CEB among women in Andhra Pradesh. Among the evaluated models, the Support Vector Machine (SVM) demonstrated good predictive performance. The results indicate that women's age, caste, wealth index, fertility preference, and place of residence are important determinants of reproductive patterns. While machine learning models provide useful predictive insights, classical statistical models for count data remain valuable for interpretation and policy relevance. Future research may compare machine learning and traditional count-data regression models to further assess their relative strengths.

Acknowledgments

The authors appreciate the DHS Program for granting access to the data analyzed in the study.

Ethics Approval and Consent to Participate

This study does not require ethical approval, as it involves secondary analysis of NFHS data from 2019 to 2021. The data were anonymous; no personal data was on the DHS website.

Conflicts of Interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization: MUNISWAMY, B.; LAVANYA, M.V. **Data curation:** LAVANYA, M.V. **Formal analysis:** LAVANYA, M.V. **Funding acquisition:** LAVANYA, M.V. **Investigation:** LAVANYA, M.V. **Methodology:** LAVANYA, M.V. **Project administration:** MUNISWAMY, B.; LAVANYA, M.V. **Software:** LAVANYA, M.V. **Resources:** LAVANYA, M.V. **Supervision:** MUNISWAMY, B. **Validation:** MUNISWAMY, B.; LAVANYA, M.V. **Visualization:** MUNISWAMY, B.; LAVANYA, M.V. **Writing - original draft:** LAVANYA, M.V. **Writing - review and editing:** MUNISWAMY, B.; LAVANYA, M.V.

References

1. Balogun, J. A., Egejuru, N. C., & Idowu, P. A. Comparative analysis of predictive models for the likelihood of infertility in women using supervised machine learning techniques. *Computer Reviews Journal*, **2**(1), 313–330 (2018).
2. Basumatary, K. Fertility pattern and differential in India. *ARPHA Preprints*, **4**, e101097 (2023). https://doi.org/10.3897/arphapreprints.e101097
3. Bitew, H. F., Woldeyohannes, S. M., & Yeshambel, T. W. Machine learning approach for predicting under-five mortality determinants in Ethiopia: Evidence from the 2016 Ethiopian Demographic and Health Survey. *Genus,* **76**, 1–16 (2020). https://doi.org/10.1186/s41118-020-00093-0
4. Bongaarts, J. A framework for analyzing the proximate determinants of fertility. Population and Development Review, **4**(1), 105–132 (1978).
5. International Institute for Population Sciences (IIPS) & ICF. National Family Health Survey (NFHS-5), 2019–21: India. Mumbai: IIPS (2021).
6. Koshy, S., & Anuradha, K. A review on AI methods for the prediction of infertility in women. *International Journal of Engineering Technology and Management Sciences*, **6**, 287–296 (2022).
7. Lavanya, M. V., & Muniswamy, B. Exploring reproductive patterns: A Poisson regression study in Andhra Pradesh. *African Journal of Biomedical Research*, **27**(1S), 1181–1190 (2024a). https://doi.org/10.53555/AJBR.v27i1S.1436
8. Lavanya, M. V., & Muniswamy, B. A study on the Hurdle Poisson regression model for reproductive patterns on count data. *African Journal of Biological Sciences*, **6**(4), 1309–1322 (2024b). https://doi.org/10.33472/AFJBS.6.4.2024.1309-1322
9. Liu, X., Chen, Z., & Ji, Y. Construction of the machine learning-based live birth prediction models for the first in vitro fertilization pregnant women. *BMC Pregnancy and Childbirth*, **23**, 476 (2023). https://doi.org/10.1186/s12884-023-05791-7
10. Mfateneza, E., Rutayisire, P. C., Biracyaza, E., Musafiri, S., & Mpabuka, W. G. Application of machine learning methods for predicting infant mortality in Rwanda: Analysis of Rwanda Demographic Health Survey 2014–15 dataset. *BMC Pregnancy and Childbirth*, **22**, 388 (2022). https://doi.org/10.1186/s12884-022-04759-0
11. Muniswamy, B., & Lavanya, M. V. Zero-truncated Poisson regression model for reproductive patterns on count data. *Reliability: Theory & Applications*, **1**(82), 1070–1088 (2025). https://doi.org/10.24412/1932-2321-2025-182-1070-1088
12. Raef, B., & Ferdousi, R. A review of machine learning approaches in assisted reproductive technologies. *Acta Informatica Medica*, **27**(3), 205–210 (2019). https://doi.org/10.5455/aim.2019.27.205-210

13. Rahman, A., Hossain, Z., Kabir, E., & Rois, R. Machine learning algorithm for analyzing infant mortality in Bangladesh. *In International Conference on Health Information Science*, 205–219 (2021). Springer. https://doi.org/10.1007/978-3-030-72751-2_17
14. Ram, B. Fertility decline and family change in India: A demographic perspective. *Journal of Comparative Family Studies*, **43**(1), 11–40 (2012).
15. Ranjini, K., Suruliandi, A., & Raja, S. P. Machine learning techniques for assisted reproductive technology: *A review. Journal of Circuits, Systems and Computers*, **29**(11), 2030010 (2020). https://doi.org/10.1142/S0218126620300106
16. Sobotka, T., Skirbekk, V., & Philipov, D. Economic recession and fertility in the developed world. *Population and Development Review*, **37**(2), 267–306 (2011). https://doi.org/10.1111/j.1728-4457.2011.00411.x
17. Uddin, M. J., Kabir, A., & Naznin, S. Machine learning approaches for prediction of fertility determinants in Bangladesh: Evidence from the *BDHS* 2017–18 data. *Research Square* (2024). https://doi.org/10.21203/rs.3.rs-3934391/v1