ABSTRACT

This project, entitled **Socioeconomic Determinants of Antenatal Care Utilization in India: An NFHS-Based Analysis**, uses data from the National Family Health Survey (NFHS) to examine how some socioeconomic factors influence the use of antenatal care (ANC) services among Indian women. The study starts with importing, cleaning, and pre-processing a large dataset and incorporating variables for education, household wealth, region, place of residence, religion, caste, marital status, age of mother, and children ever born.

Key variables are consistently renamed from coded NFHS formats (e.g., `v106`, `v190`, `v025`) to descriptive labels (e.g., `education level`, `wealth index`, `residence type`) for clarity. The desired outcome, the number of antenatal check-ups attended (`antenatal_checkups`), is examined against these predictors through both univariate and bivariate analyses using visualization methods including bar plots and distribution charts. The dataset is also weighted with sampling weights to ensure representativeness.

Early results indicate that wealth index and level of education are good predictors for ANC use, with more educated and richer women receiving more antenatal visits. Rural women, in contrast, have decreased utilization compared to urban women, while religious and caste composition also seem to contribute to variations in ANC utilization, highlighting underlying structural and social inequalities.

The study provides useful evidence for public health policymakers seeking to enhance maternal healthcare provision in India. It stresses the need for targeted interventions—particularly for poorer social and economic groups—to facilitate equitable access to antenatal care and enhance maternal and child health outcomes throughout the country.

CONTENT

INTRODUCTION

Antenatal care (ANC) is an integral part of maternal health care and is essential in ensuring healthy outcomes of pregnancy as well as overall health of new-borns and mothers. In spite of tremendous improvement in maternal and child health over the last decades, millions of women in low- and middle-income countries like India continue to be deprived of access to sufficient and timely ANC. The World Health Organization advises at least four antenatal visits during pregnancy to check the health of the mother and foetus, detect complications, and offer vital services like immunization, nutritional counselling, and health education. Access to these services continues to be unevenly distributed, tending to mirror wider socioeconomic disparities.

India, with its large and heterogeneous population, is challenged to provide universal ANC coverage. While institutional delivery and maternal health indicators have made significant strides, there remain disparities along region, wealth, education, and social identity. The National Family Health Survey (NFHS), a large nationally representative dataset, offers rich data on these dimensions, allowing for closer analysis of the determinants of ANC use.

Previous research has consistently highlighted that women's education, household wealth, urban-rural residence, caste, and religion are significant determinants of maternal healthcare access. However, quantifying these associations and understanding how they interact at the population level requires systematic analysis using rigorous statistical tools. In this context, our study aims to identify and quantify the socioeconomic determinants of antenatal care utilization in India using NFHS data.

The data were cleaned and processed to keep important variables that represent maternal demographics, social status, and healthcare behaviour. Education level, wealth index, residence type, region, maternal age, caste, religion, and number of children born were the variables employed to determine their effect on ANC visit frequency. Descriptive and exploratory analysis were performed to reveal patterns in the data, and then statistical modelling to determine the strength and direction of association between socioeconomic determinants and ANC uptake.

By this project, we seek to contribute to the general knowledge of maternal health inequities in India by unearthing major impediments to antenatal care use. The implications of the findings are for health policy, particularly in the planning of targeted interventions for underserved populations. Through synthesis of data-driven analysis with public health needs, this research upholds the continued striving to attain Sustainable Development Goal 3, which focuses on ensuring healthy lives and promoting well-being for all.

Background and Theory

1. Dataset Balance or Imbalance

In classification tasks, including those predicting ANC utilisation, balanced data is key for model performance. A balanced dataset is one in which we have approximately the same number of samples for each class (e.g. women with a sufficient ANC vs those without). On the other hand, a dataset is imbalanced when one class greatly surpasses the other. There could be several methods that the model tends to overfit; for instance, if most women in the dataset were given ANC and only few didn't receive ANC, the model may become biased toward the majority.

- **Why It Matters in This Study:**

Unjust learning: The imbalance of the data set can lead to the down-prediction of the minority class (in the present case, minority class can be women from underprivileged or downtrodden region) of the model.

Public Health Relevance The need to be able to identify those women who will have poor antenatal care is essential for targeted intervention, and failing to identify this group is more than an academic issue.

- **Solution Applied:**

SMOTE as employed to solve the imbalance problem. SMOTE oversamples the minority class examples and aids the model in learning discriminative decision boundaries.

2. Missing Values

The NFHS dataset, like many large-scale surveys, contains missing values due to skipped questions, data entry errors, or nonresponses.

- **Sources of Missingness in NFHS:**

Respondent refusal or ignorance.

Skipped items based on branching logic.

Incomplete household records.

- **Handling Approach:**

Deletion was done to resolve this problem.

- **Imputation Methods used:**

Mean/Median Imputation for continuous variables.

Mode Imputation for categorical variables.

In some cases, model-based imputation was explored for higher accuracy.

3. Outliers

Outliers in features like age, number of children, or household income can distort model training.

- **Outlier Detection:**

Boxplots and Z-score method for normally distributed features.

Interquartile Range (IQR) for skewed variables.

- **Treatment**

Outliers were not universally removed but were capped or transformed where necessary to preserve information while reducing distortion.

4. Correlation Analysis

To examine the linear associations between antenatal care use and important socioeconomic factors, a Pearson correlation analysis was used. The aim was to determine how various factors like education, income, maternal age, and fertility affect the number of antenatal visits.

- **Visualization of Correlation:**

1) Pair plot: Scatterplots for all pairs of features to detect relationships.
2) Heatmap: Displays the correlation matrix with colours indicating the strength and direction of correlations.

5. Feature Scaling

Feature Scaling is very important to have all features contribute equally to the model, and even more for distance-based models (e.g. KNN and SVM).

- **Types of Scaling Techniques:**

- **Absolute Maximum Scaling:**
  Normalizing data to between −1 and 1 by dividing each element by the largest absolute value.

$$X_{\text{scaled}} = \frac{X_i - \max(|X|)}{\max(|X|)}$$

- **Min-Max Scaling:**
  Scales X with a fixed range, often [0, 1].

$$X_{\text{scaled}} = \frac{X_i - X_{\min}}{X_{\max} - X_{\min}}$$

- **Normalization:**
  Apply mean and range based normalization to the data.

$$X_{\text{scaled}} = \frac{X_i - X_{\text{mean}}}{X_{\max} - X_{\min}}$$

- **Standardization:**
  centre's the data set to have a mean value of 0 and standard deviation of 1.

$$X_{\text{scaled}} = \frac{X_i - X_{\text{mean}}}{\sigma}$$

Where σ is the standard deviation.

- **Method Used in Project:**
  In this project, Standardization was applied to scale numeric variables.

6. Encoding of Categorical Variables

In preparation for machine learning analysis in this research on Socioeconomic Determinants of Antenatal Care Utilization in India, categorical variables were encoded into numerical formats by applying the following two encoding methods:

Label Encoding:

Label encoding was used for ordinal categorical variables whose categories have an ordered or ranking meaning. For example, variables such as level of education (No education, Primary, Secondary, Higher) and wealth index (Poorest to Richest) were labelled with growing integer values (e.g., 0, 1, 2, 3, etc.). This maintains the ordinal order between the categories so that machine learning models can understand the underlying progression or hierarchy in the data. Label encoding is well-supported by models capable of utilizing orderings like such.

One-Hot Encoding

For nominal categorical variables with no inherent order, one-hot encoding was employed. This process generates a unique binary column per category, and a value of 1 implies the existence of that category and a value of 0 for its absence. For instance, the religion variable with categories such as Hindu, Muslim, Christian, and Others was transformed into several binary columns. One category is usually excluded (using drop_first=True) to prevent multicollinearity. One-hot encoding guarantees that models process every category separately, without suggesting any ordinal relationship.

7. Model Used

To check ANC and Independent variables, several machine learning classification algorithms were implemented, each offering unique strengths in handling different types of data structures and complexities. The models used are as follows:

- **Logistic Regression:**

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, logistic regression is a predictive analysis. It is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval, or ratio-level independent variables. Logistic regression uses the logistic function (also called the sigmoid function) to map the output of the linear combination of features to a probability value between 0 and 1.

$$P(y = 1|x) = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + \cdots + b_n x_n)}}$$

Where;

1. P(y=1|x) is the probability of the target variable being 1 given input x.
2. e is the base of the natural logarithm.
3. b0, b1, bn are the coefficients of the logistic regression model.
4. x1, xn are the independent variables (features).

The logistic regression model is trained using optimization techniques such as maximum likelihood estimation (MLE) or gradient descent. The goal is to find the optimal coefficients that maximize the likelihood of the observed data. In binary classification, logistic regression uses a decision boundary to classify the input into one of the two classes based on the predicted probability. By default, the decision boundary is set at 0.5, but it can be adjusted depending on the specific requirements of the problem.

- **SVM:**

SVM is a supervised learning model for classification (and regression) that finds the optimal hyperplane separating classes with the maximum margin.

Key Concepts:

- **Hyperplane**

A decision boundary (e.g., a line in 2D, plane in 3D) that separates classes.

Defined by: $w^T x + b = 0$

where: w = weight vector (normal to the hyperplane), b = bias term.

- **Support Vectors**

The closest data points to the hyperplane (they "support" the margin).

- **Margin**

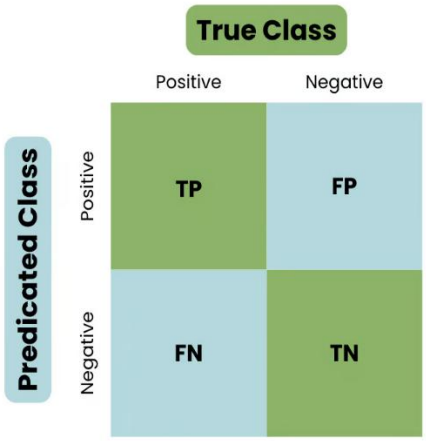
The distance between the hyperplane and the nearest data points of either class.

SVM maximizes this margin for better generalization.

8. Evaluation Metrics

To measure the performance of classification models that forecast antenatal care usage, some measurement criteria were utilized. These metrics give a complete picture of how well the models can differentiate between the positive and negative classes.

1. **Confusion Matrix:** A confusion matrix is a table that illustrates the accuracy of a classification algorithm. It plots actual vs. predicted results:



2. **Accuracy:** Accuracy measures the overall correctness of the model.

$$\textbf{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

3. **Precision:** Precision measures how many of the predicted positive cases were actually positive.

$$\textbf{Precision} = \frac{TP}{TP+FP}$$

4. **Recall (Sensitivity or True Positive Rate):** Recall measures how many of the actual positive cases were correctly identified.

$$\text{Recall} = \frac{TP}{TP+FN}$$

5. **F1 Score:** F1 Score is the harmonic mean of precision and recall, especially useful when classes are imbalanced.

$$\text{F1} = \frac{2\ X\ Recall\ X\ Precision}{Recall + Precision}$$

6. **ROC Curve and AUC (Area Under Curve):** ROC Curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various thresholds.
   - **TPR (Recall) and FPR are defined as:**

$$\text{TRP} = \frac{TP}{TP+FN}$$

$$\text{FPR} = \frac{FP}{FP+TN}$$

   - **AUC (Area Under the ROC Curve)**: Summarizes the model's ability to discriminate between classes.
   - **AUC values range from 0 to 1:**

   AUC = 1: Perfect model
   AUC = 0.5: No discriminative ability (random guessing)
   AUC < 0.5: Worse than random

Data Description

This analysis utilizes data from the National Family Health Survey (NFHS-5) carried out in India between 2019 and 2021. NFHS-5 is a nationally representative cross-sectional survey that offers critical information on population health, nutrition, and demographic indicators. The current analysis centers on a sub-group of 15–49-year-old women who had experienced a live birth in the five years prior to the survey.

Following data cleaning and the exclusion of inconsistent or missing values, the final analysis sample consists of 172,702 women. The study's objective is to investigate the socioeconomic determinants of antenatal care (ANC) use, informed by a set of demographic and household-level variables.

- **Major Variables Included:**
  A. Dependent Variable:

Antenatal Check-ups: This is the measure of the number of ANC visits in the respondent's last pregnancy. According to WHO criteria, ANC use is generally ranked as "adequate" (≥4 visits) and "inadequate" (<4 visits).

B. Independent Variables:

Education Level: Categorical variable representing the highest education level the woman has achieved (e.g., no education, primary, secondary, higher).

Years of Education: Continuous variable indicating the number of years the woman has spent in formal schooling.

Wealth Index: Asset-based composite index (poorest to richest), a measure of household economic status.

Region: Geographical grouping of India's states into larger regions (e.g., North, South, East, West, Northeast, Central).

Type of Residence: Urban or rural residence classification of the woman's dwelling.

Mother's Age: Age of the woman at the time of the survey, regarded either as a continuous or as a categorized variable (e.g., 15–24, 25–34, 35–49).

Religion: Categorical variable referring to the religious affiliation of the respondent (e.g., Hindu, Muslim, Christian, Other).

Cast: Social stratification variable (e.g., Scheduled Caste, Scheduled Tribe, Other Backward Class, General).

Total Children Born: Number of children the woman has given birth to.

Marital Status: Reveals whether or not the woman is married, previously married, or never married.

Methodology

- **Data Understanding:**
  The data employed in this research was based on NFHS-5 (National Family Health Survey - 5) and was centred on Indian women's antenatal care (ANC) use. The data consisted of several socioeconomic and demographic attributes like education level, years of schooling, wealth index, zone, residence type, mother's age, religion, caste, number of children born, and marital status. The target attribute was binary that signifies whether or not the woman received proper ANC (1/0). 2.
- **Missing Value Imputation:**
  There were missing values in a couple of variables in the dataset. These were addressed by either: Deleting rows with too much missingness, or Imputing values by the mode for categorical and median for numerical variables, taking care that imputation did not skew the dataset distribution.
- **Feature Selection:**
  Variables were chosen based on domain expertise and appropriateness to ANC utilization. Characteristics such as `education level`, `wealth index`, `residence type`, and `mother age` were kept since they had been shown to impact maternal health outcomes. Very correlated or redundant variables were excluded to avoid multicollinearity.
- **Encoding of Categorical Variables:**
  Label Encoding was applied for ordinal variables such as education level and wealth index.
  One-Hot Encoding was used for nominal variables like region, religion, caste, and residence type, to transform them into a model-input-friendly format.
- **Data Splitting:**
  The data was divided into training and test sets with an 80:20 ratio to make the models generalizable. The training data was utilized for building the models, and the test data for evaluation.
- **Class Imbalance Handling:**
  As the target variable was slightly imbalanced (there were more women who received poor ANC), class imbalance was handled with SMOTE (Synthetic Minority Oversampling Technique) on the training dataset to balance the class distribution to enhance model performance on the minority class.
- **Model Building:**
  Two classification models were employed:
  Logistic Regression: A basic statistical model for binary classification problems.
  Support Vector Machine (SVM): A more sophisticated model that is able to deal with non-linear decision boundaries via kernel tricks (e.g., RBF kernel).
- **Model Evaluation:** Models were compared using the following metrics:
  Accuracy: Correctness overall.
  Precision and Recall: To measure performance on the positive class.
  F1-Score: Harmonic mean of precision and recall.

ROC-AUC: For assessing model ability to discriminate between thresholds.

- **Feature Importance Analysis:**
  For Logistic Regression, the coefficients were interpreted to identify direction and magnitude of association between independent variables and ANC utilization. For SVM, feature importance was derived using model-agnostic methods like permutation importance.
- **Model Comparison and Selection:**
  Between the two models, Logistic Regression did a tad bit better in the aspects of interpretability and similar performance measures. SVM performed more strongly on precision and F1-score after SMOTE balancing, however. Final model selection involved making a compromise between interpretability and predictive ability, and SVM was used instead if accuracy and recall were what mattered most.

Data Pre-processing

Data pre-processing is a core activity in the data science workflow that prepares the dataset to be clean, consistent, and ready for machine learning model building. In this project, various methods were used to pre-process raw survey data to a form that was appropriate for predictive modelling.

1. Missing Data Handling:

There were missing values in some of the variables in the dataset. Upon close examination, it was discovered that these were mostly in non-critical areas or due to logically missing data (e.g., first births with no previous child-related data). Rather than deleting a significant amount of the data, a considerate imputation plan was used. For instance, missing values in columns such as `age_at_death` and `age_at_death_months` were replaced with zeros to indicate the non-applicability of such attributes in firstborn children. Some of the columns, such as `preceding_birth_interval`, were removed completely since they were not considered important to the task of predicting ANC (antenatal care) usage.

2. Encoding Categorical Variables:

Machine learning models need numerical input; hence, categorical variables were accordingly transformed:

Label Encoding was used on ordinal variables like `education level` and `wealth index`, where the levels have a natural rank or ordering. This retained the ordinal order between the values, enabling algorithms such as Logistic Regression to estimate the gradient of effect between the levels.

One-Hot Encoding was applied to nominal variables like `region`, `residence type`, `religion`, `caste`, and `marital status`. These do not have intrinsic order, so binary columns were established for each category in order to provide a good, non-hierarchical representation for the models.

3.Target Variable Construction:

The target of the project was to predict whether a woman had sufficient antenatal care. A binary target variable was defined on this basis, with:

1 indicating sufficient ANC use, and **0** indicating insufficient ANC use.

This was done to pose the problem as a binary classification problem, appropriate for use with algorithms like Logistic Regression and Support Vector Machines (SVM).

4. Feature Scaling:

Even though Logistic Regression is quite insensitive to the scaling of features, Support Vector Machine (SVM) needs features to be scaled similarly because it's based on distance-based computation. Hence, numerical features like `mother age` and `years_of_education` were scaled using Z-score normalization (mean = 0, standard deviation = 1). This was done to make all the numerical attributes equally influential while training the model.

5. Data Splitting:

In order to train and test the model, the data was split into training (80%) and testing (20%) sets through a stratified split in order to maintain the class distribution of the target variable. This ensures that the model can be tested on unseen data and thereby its generalizability.

6. Handling Class Imbalance:

The target variable's class distribution was slightly unbalanced, with fewer occurrences of proper use of ANC. In order to reduce the bias caused due to this imbalance, SMOTE (Synthetic Minority Over-sampling Technique) was used on the training set. SMOTE creates new minority class examples by synthesizing them, thus allowing the models to learn balanced decision boundaries and enhancing their performance on the underrepresented class.

EXPLANATORY DATA ANALYSIS

EDA was conducted to understand the distribution and relationships between key variables affecting antenatal care (ANC) utilization. This involved the use of bar plots, count plots, and box plots to analyse how various socioeconomic and demographic variables influence ANC check-up behaviour.

1. Summary Statistics:

Descriptive Statistics for Numerical Variables

	years_of_education	mother_age	total_children_born
count	11888	14811	14811
mean	4.225942	27.17683	2.22463
std	1.676221	5.090391	1.375528
min	0	15	0
25%	3	24	1
50%	4	26	2
75%	5	30	3
max	8	49	16

- ### years_of_education:

  Count: There are 11,888 valid responses for years of education.
  Mean: The mean number of years of education is about 4.23 years.
  Standard Deviation: The standard deviation of 1.68 years reflects a moderate dispersion around the mean.
  Minimum: The minimum number of years of education is 0.
  25th Percentile: 25% of the people have 3 or less years of education.
  50th Percentile (Median): 50% of the people have 4 or less years of education. This is a bit less than the mean, indicating a slight skew towards lower levels of education.
  75th Percentile: 75% of the people have 5 or less years of education.
  Maximum: The highest number of years of education reported is 8.

- ### mother_age:

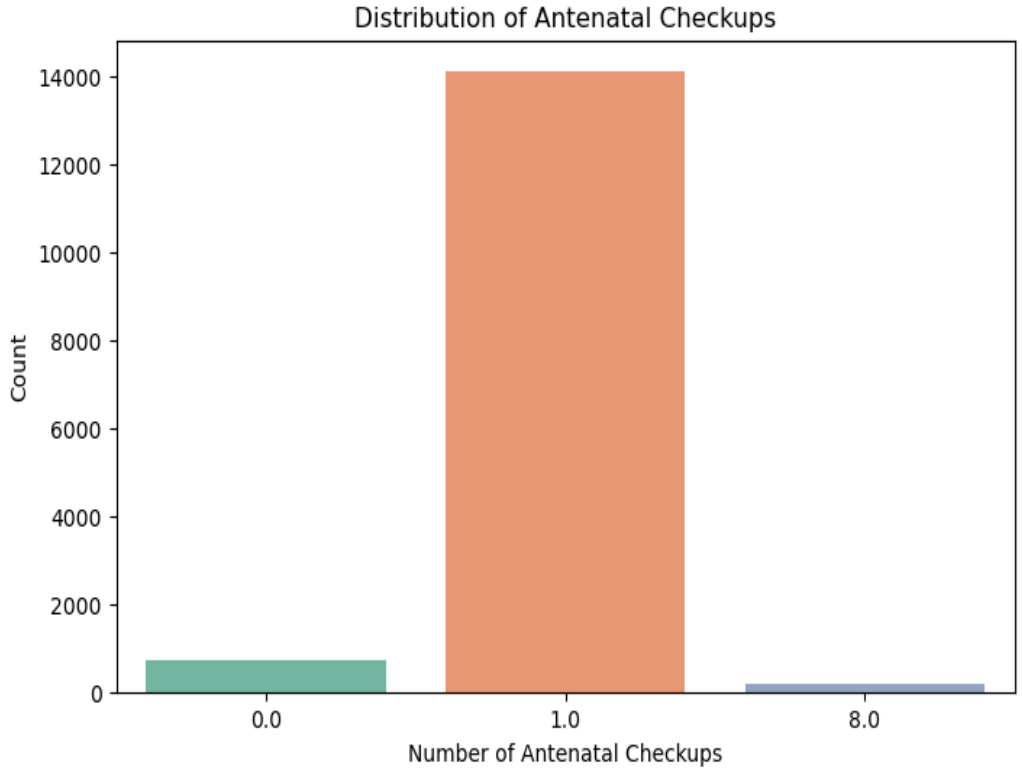
  Count: There are 14,811 valid records for the age of the mothers.
  Mean: The mean age of the mothers in the data set is around 27.18 years.
  Standard Deviation: A moderate range in the age of the mothers is reflected in the standard deviation of 5.09 years.
  Minimum: The minimum age of the mother in the data set is 15 years.
  25th Percentile: 25% of the mothers are 24 years or less.
  50th Percentile (Median): Fifty percent of the mothers are 26 years and under. The median is somewhat lower than the mean, which indicates a minor positive skew (more young mothers).
  75th Percentile: Seventy-five percent of the mothers are 30 years and under.
  Maximum: The maximum age of a mother in the data is 49 years.

- ### total_children_born:

  Count: There are 14,811 valid observations for the total number of children born.
  Mean: On average, these mothers have given birth to about 2.22 children.

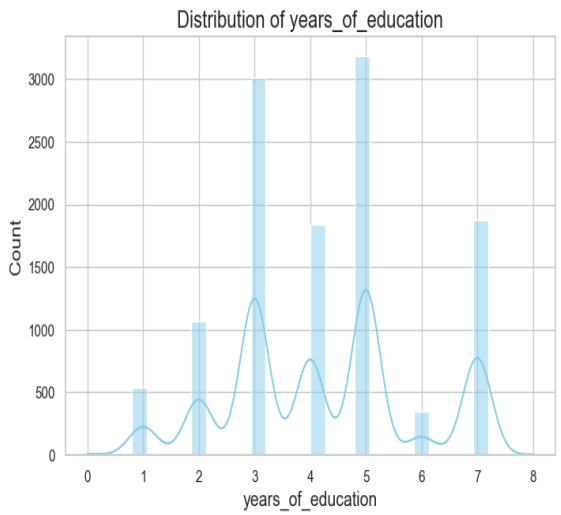
Standard Deviation: A standard deviation of 1.38 children implies a moderate range in the number of children given birth to.
Minimum: There are mothers who have given birth to no children (0).
25th Percentile: 25% of the mothers have had 1 or fewer children.
50th Percentile (Median): 50% of the mothers have given birth to 2 or less children. The median is practically equal to the mean.
75th Percentile: 75% of the mothers have given birth to 3 or less children.
Maximum: The highest number of children born to any mother in this data is 16, which is an outlier.

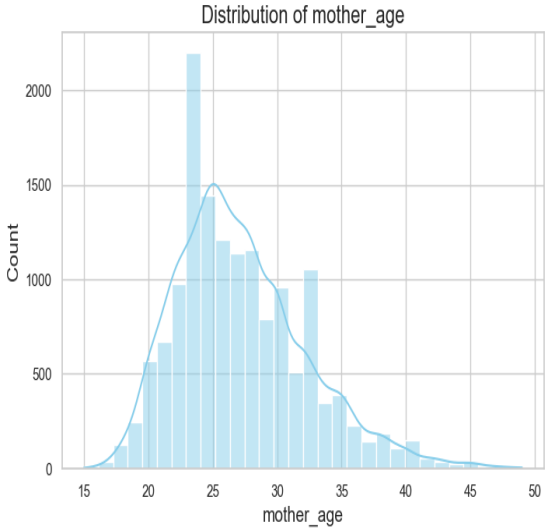
2. Distribution of Antenatal Check-ups before data processing:

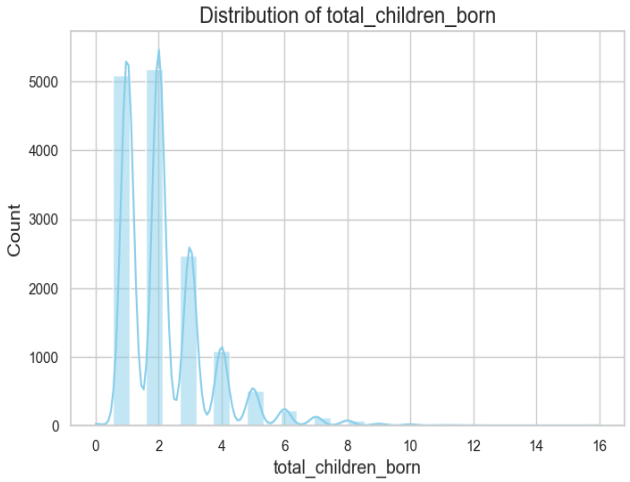


- **Observations:**
  I. Label 1.0 is the most frequent in the distribution, comprising around 14,000 women.
  II. Few women are under 0.0, meaning no antenatal check-ups — it is around 1,000 respondents.
  III. Even fewer fall under 8.0, which could mean good or optimal number of ANC check-ups (the norm is usually WHO recommends 8 or more visits).
- **Interpretation:**
  I. The majority of women had at least one antenatal check-up (designated as 1.0), indicating minimal awareness and availability of maternal health care.
  II. Yet, the fact that there is a non-negligible proportion with 0 check-up's indicates that some section of the population is totally outside the radius of maternal care, and this is problematic.
  III. The extremely low number at 8 check-up's indicates that effective ANC coverage is unusual, and this implies problems of continuity and completeness of maternal care.

3. Distribution of Years of education, Mother age, Total Children ever born:

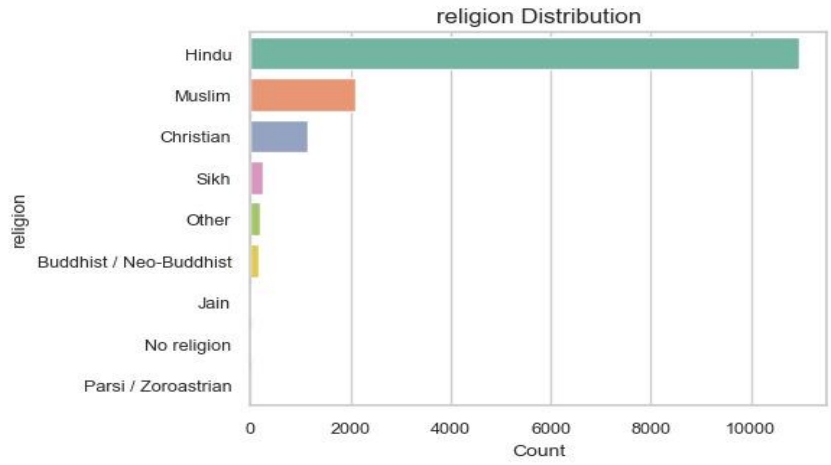




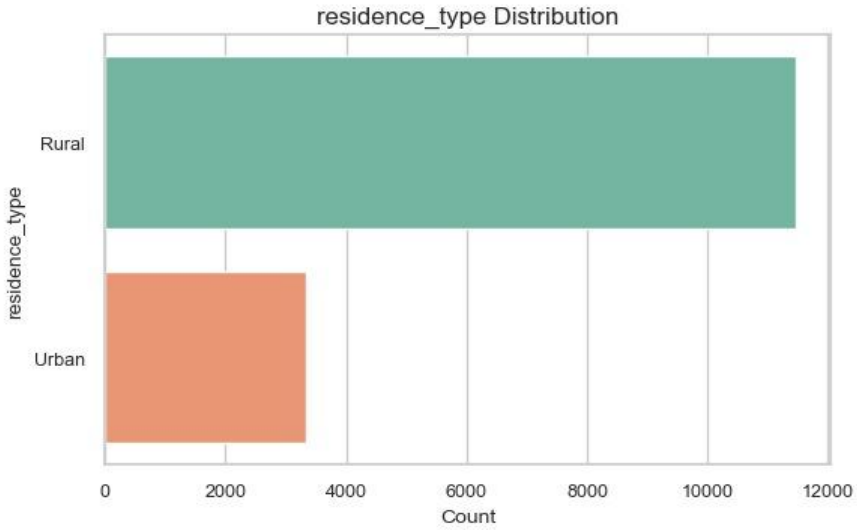


- **Interpretation:**
  I. **Years of Schooling:** Experienced several peaks, reflecting groups of people with the same years of schooling, with significant concentrations in 3, 5, and 7 years.
  II. **Mother's Age:** Shown to have a right-skewed distribution, with the highest frequency in the mid-twenties and a steady decline in frequency towards higher ages, with a small peak in the early thirties.
  III. **Total Children Born:** Displayed a number of clear peaks at lower values (1, 2, and 3), with the distribution falling off sharply as the number of children grew, showing an inclination towards having fewer children.

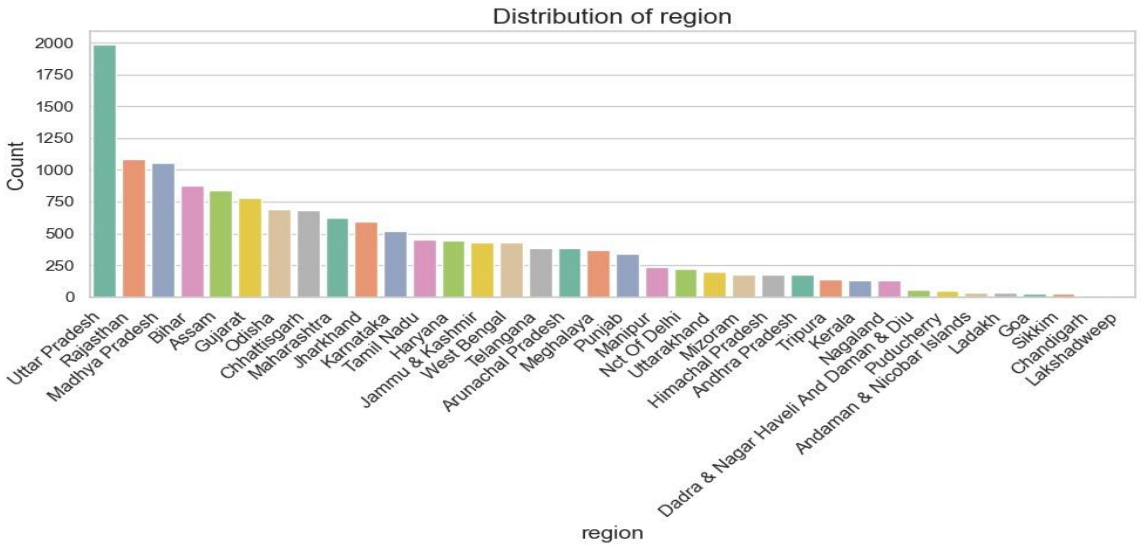
4. Distribution of other Independent variables before data processing:



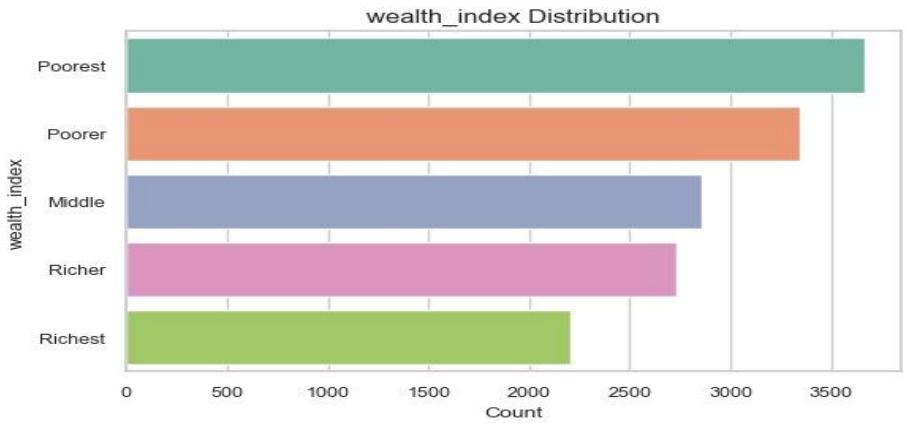
Interpretation: The religious distribution in this data set is strongly skewed toward Hinduism, which is the most represented religion by a wide margin. Islam is the second largest group, although significantly smaller. Christianity and Sikhism have fewer followers, while "Other," Buddhist/Neo-Buddhist, Jain, no religion, and Parsi/Zoroastrian communities make up very small portions of the data set



Interpretation: The bar chart graphically shows there is a disparity in residence types distribution. The much longer bar for "Rural" reflects that most people in this set live in rural areas. By contrast, the shorter bar for "Urban" shows that there is a far smaller percentage of the people in this set that live in urban areas. The visual disparity in the bar lengths indicates a significant imbalance, rural residency being much more prevalent than urban residency in the population addressed by this data.

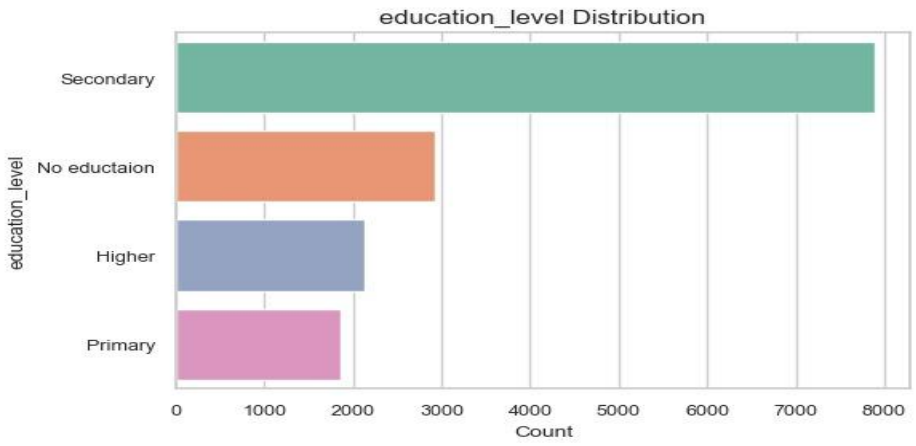


Interpretation: The bar chart illustrates the breakdown of people across regions. Uttar Pradesh has the highest number, reflecting the largest number of representatives from this area. Then come Rajasthan and Madhya Pradesh, with relatively large numbers, albeit fewer than Uttar Pradesh. There is a considerable dip to the next group of regions, including Bihar, Assam, Gujarat, Odisha, and Chhattisgarh, which all have similarly moderately high figures. Moving forward, Maharashtra, Jharkhand, Karnataka, Tamil Nadu, and Haryana depict progressively diminishing, though substantial, representations. All the rest of the regions, beginning from Jammu & Kashmir and going all the way up to Lakshadweep, reflect a progressive and then sudden diminution in the number of people, with the last few regions having extremely limited representation in the dataset. This distribution shows a significant disparity in the populations that come from various regions, with a strong predominance of the more populous northern and central states.



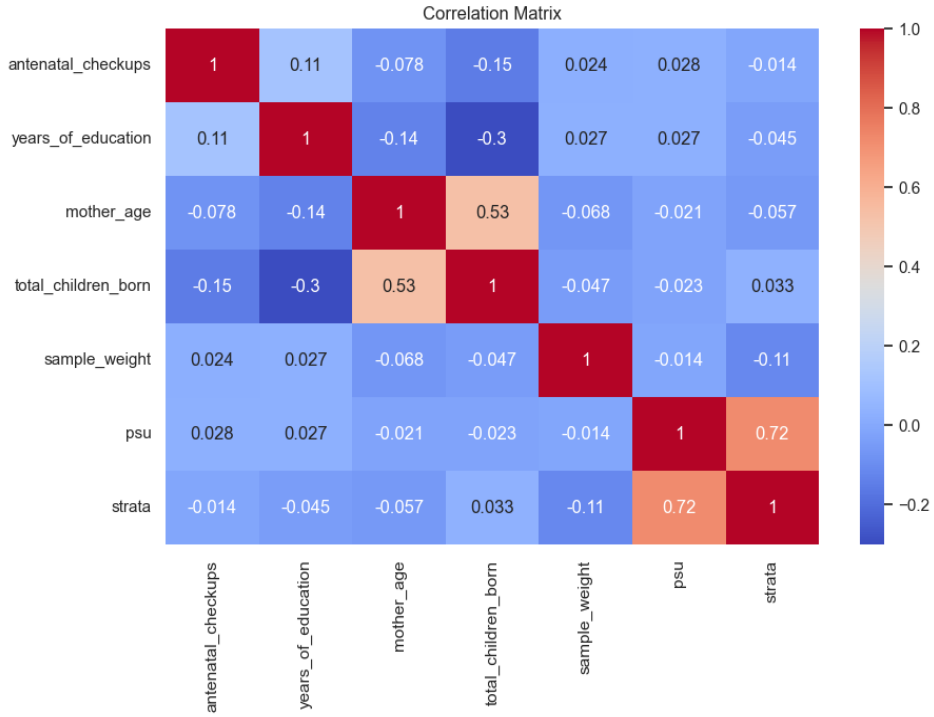
Interpretation: The horizontal bar chart displays the breakdown of the wealth index by category. The category "Poorest" has the longest bar, reflecting the largest number of people belong to this wealth category. The "Poorer" category also has a large number, although slightly lower than the "Poorest" category. The "Middle" wealth index has a relatively small number in comparison to the two poorer categories. Both the "Richer" and "Richest" groups have the shortest bars, indicating that these wealth groups have the smallest fractions of the population

in this data. Generally, the distribution indicates a greater number of people in the lower wealth index groups, declining by wealth index.



Interpretation: The horizontal bar chart displays the breakdown of education levels across the dataset. The "Secondary" level has the longest bar, meaning it is the most represented group. Next to "Secondary," "No education" is the most common category. "Higher" education is smaller than "No education," and "Primary" education has the shortest bar, meaning the least representation among the four categories. In short, from the data, most people possess a secondary level of education, followed by a considerable number with no education, followed by a lesser number with higher education, and the least number with primary education only.

5. Correlation matrix:



Interpretation: This is a correlation matrix graphically representing pairwise linear correlations among various variables. The numerical value and intensity of colour provide evidence of correlation direction and strength, respectively.

- **Strong Positive Correlation:**

- psu and strata (0.72): There exists a high positive correlation between primary sampling unit (psu) and the variable used for stratification (strata). This is not surprising because strata tend to be employed in defining primary sampling units within survey design.
- mother_age and total_children_born (0.53): A moderate positive correlation between the age of the mother and the number of children born. This indicates that, on average, older mothers have more children.

- **Weak Positive Correlation:**

- antenatal_checkups and years_of_education (0.11): There is a very weak positive correlation, indicating a very slight tendency for mothers with more years of education to have more antenatal check-ups.
- years_of_education and sample_weight (0.027), psu (0.027): These two correlations are nearly as close to zero as is possible, which indicates virtually no linear relationship.
- antenatal_checkups and sample_weight (0.024), psu (0.028): These are very weak positive correlations.
- total_children_born and strata (0.033): A very weak positive correlation.

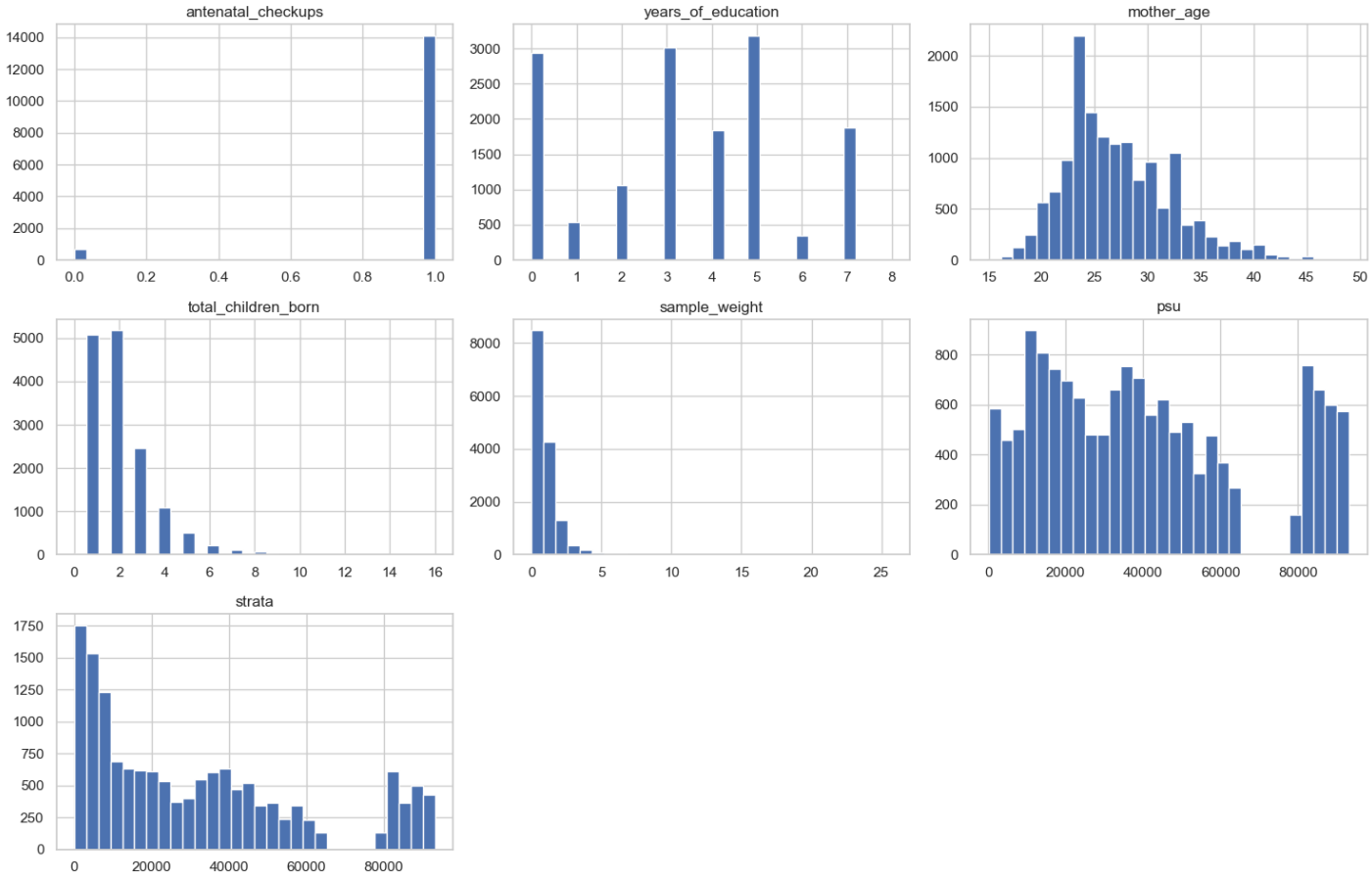
- **Weak Negative Correlation:**

- antenatal_checkups and mother_age (-0.078), total_children_born (-0.15), strata (-0.014): These indicate a very weak tendency for mothers with more children and older mothers to have a bit fewer antenatal check-up's, and a negative correlation with the strata variable that is negligible.
- years_of_education and mother_age (-0.14), total_children_born (-0.30), strata (-0.045): There is a weak negative correlation between years of education and mother's age and a somewhat stronger weak negative correlation with the number of children born. This could indicate that younger mothers and those with fewer children have slightly more years of education in this data set. The correlation with strata is insignificant.
- mother_age and sample_weight (-0.068): A very weak negative correlation.
- total_children_born and sample_weight (-0.047): A very weak negative correlation.
- sample_weight and psu (-0.014), strata (-0.11): Very weak negative correlations.

- **Near Zero Correlation:**

The majority of other pairs of variables have correlations close to zero, meaning a very weak or non-existent linear relationship between them.

6. Histograms:



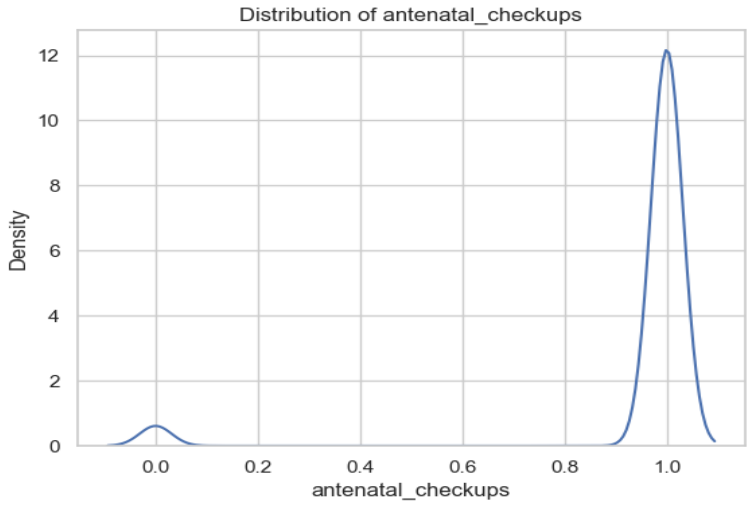
Interpretation:

- antenatal_check-up's: This binary histogram shows the overwhelming majority of observations to be at 1 (probably denoting "yes" or having received antenatal checkups), and there is an infinitesimally small number of observations at 0 (probably denoting "no" or not receiving antenatal check-up's). This indicates that antenatal checkups are extremely prevalent in this data set.
- years_of_education: This histogram also indicates a multimodal distribution, as in the previous density plot. There are clear peaks at some years of education, especially at 0, 3, 5, and 7 years. This suggests that these are the most common levels of education in the sample.
- mother_age: This distribution is quite right-skewed, with the mode being near the mid-twenties. Frequency diminishes as mother's age increases with a tail out towards older ages. This goes along with the earlier observation.
- total_children_born: This histogram is strongly right-skewed with the majority of the data clumped at the lower numbers of children (1, 2, and 3). The frequency decreases sharply as the number of children goes up, which also agrees with our previous interpretation.
- sample_weight: This right-skewed histogram has most of the sample weights bunched at lower values with a long tail spreading to higher weights. This implies that some

observations in the data have weights much higher than others, which is typical in survey data to handle varying probabilities of selection.

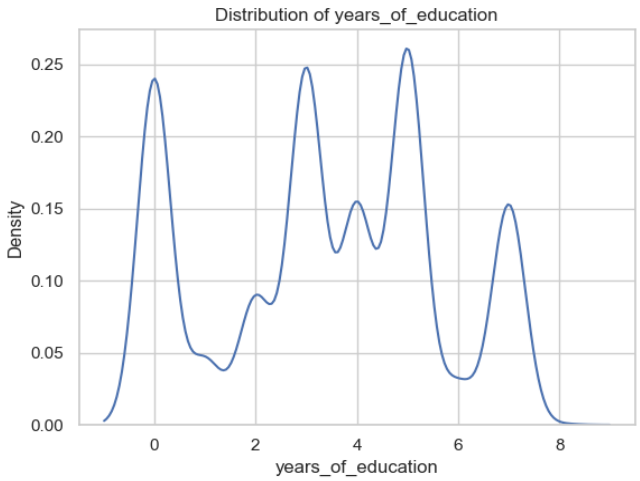
- psu (Primary Sampling Unit): This histogram indicates a generally irregular distribution across various PSU values. There doesn't seem to be a leading PSU; rather, the frequencies fluctuate over a set of PSU identifiers. This is expected since PSUs are different geographical or administrative units sampled during the first stage of a survey.
- strata: This histogram also indicates a diversified distribution across various strata levels. There are several peaks and troughs, which means that the number of observations in each stratum is not the same. Strata are employed to split the population into homogeneous groups prior to sampling, and the distribution here is indicative of the number of sampled units in each specified stratum.

7. Distribution of Antenatal Check-ups After data processing:

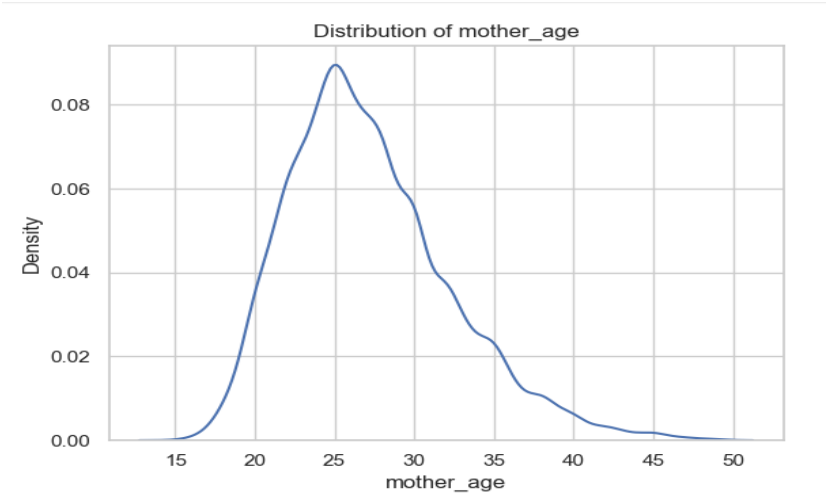


- **Bimodal Distribution:** The plot shows two distinct peaks, indicating a bimodal distribution.
- **Major Peak at 1:** There is a very sharp and tall peak at the value of 1. This suggests a very high concentration of observations at this value. In the context of antenatal checkups, this likely represents individuals who had the maximum number of recorded checkups or simply indicates the presence of checkups (coded as 1).
- **Minor Peak Near 0:** There is a much smaller and broader peak close to the value of 0. This indicates a small proportion of observations at or near zero, likely representing individuals who had very few or no antenatal checkups.
- **Gap in Between:** There is a clear gap or very low density of observations between the two peaks, suggesting that values between approximately 0.1 and 0.8 for "antenatal_checkups" are very rare in this dataset.

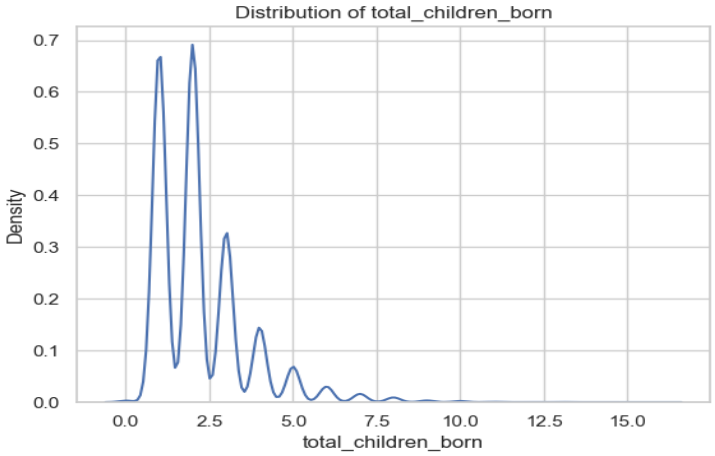
8. Distribution of other Independent variables after data processing:



- **Multimodal Distribution:** The plot clearly shows multiple peaks, indicating a multimodal distribution. This suggests that there are several common levels of educational attainment within the dataset.
- **Peaks at Integer Values:** The most prominent peaks appear to be centred around integer values, specifically 0, 3, 5, and 7 years of education. These peaks suggest a higher density of individuals with these specific durations of schooling.
- **Smaller Peaks/Shoulders:** There are also smaller peaks or "shoulders" around 1, 2, 4, and 6 years of education, indicating some concentration of individuals at these levels as well, though less pronounced than the primary peaks.
- **Range:** The density is concentrated between 0 and 8 years of education, with very little density outside this range, suggesting that the years of education in this dataset primarily fall within this interval.
- **Interpretation:** The multimodal nature of the distribution suggests that educational attainment in this population is not uniform but rather clustered around specific durations, possibly reflecting different stages of schooling or typical educational pathways. The peaks at 0, 3, 5, and 7 might correspond to completion of primary school (around 5 years), middle school (around 8 years, implying a peak around that mark if starting at 0), or other significant educational milestones, although without more context on the education system, precise labelling is difficult. The peak at 0 likely represents individuals with no formal education.

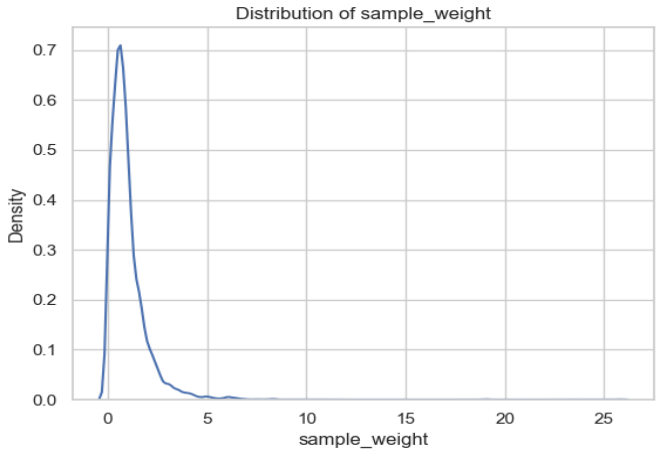


- **Skewed Distribution:** The distribution is clearly skewed to the right. The peak of the density curve is around the mid-twenties (approximately 25-26 years old), indicating that this is the most common age group for mothers in this dataset.
- **Longer Tail to the Right:** The tail of the distribution extends towards older ages, indicating that while less frequent, there are mothers in their thirties and forties present in the dataset. The decline in density is more gradual on the right side of the peak.
- **Range:** The mother's age ranges from the mid-teens to around 50 years old, consistent with the descriptive statistics we saw earlier.
- **Single Peak (Unimodal):** The distribution has a single, clear peak, suggesting one primary mode or central tendency for the mother's age.

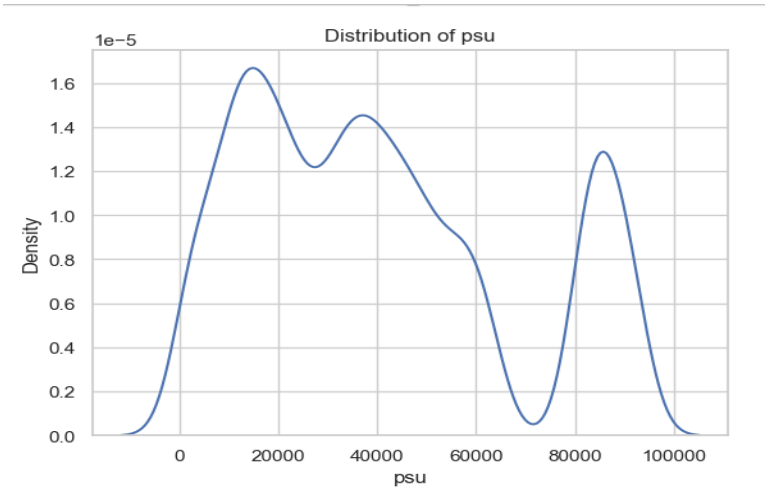


- **Highly Multimodal and Skewed:** The distribution is highly multimodal with distinct peaks at lower integer values, and it is strongly skewed to the right.
- **Prominent Peaks at 1, 2, and 3:** There are very prominent peaks at 1 and 2 children born, with a slightly lower but still significant peak at 3 children. This indicates that these are the most common numbers of children born in this dataset.
- **Decreasing Peaks for Higher Numbers:** As the number of children increases beyond 3, the height of the peaks decreases rapidly. We can see smaller peaks at 4, 5, and so on, but their density is much lower.
- **Long Tail to the Right:** The distribution has a long tail extending towards higher numbers of children born, although the density becomes very low for values above around 6 or 7. This signifies that while larger families exist, they are increasingly rare in this dataset.

- **Concentration at Lower Values:** The majority of the density is concentrated at the very beginning of the distribution, indicating that most individuals in this dataset have a small number of children.

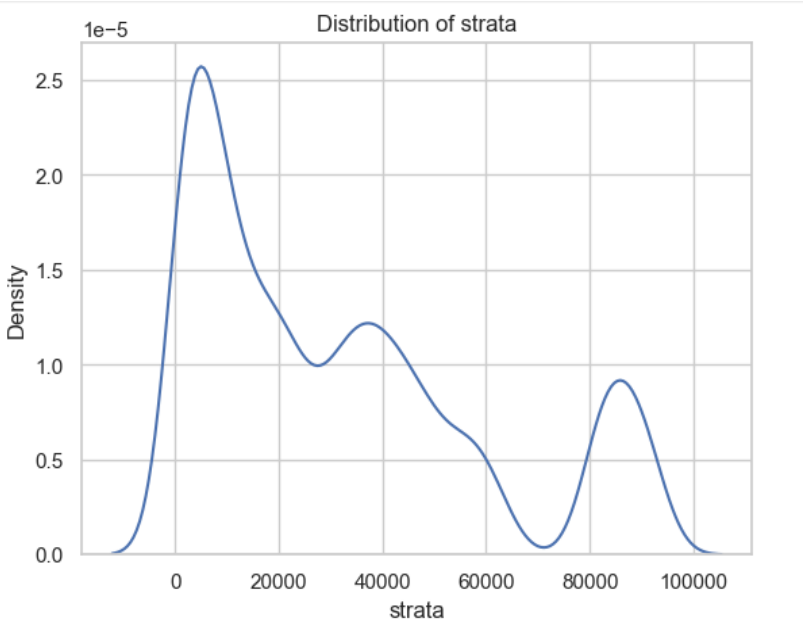


- **Strong Right Skew:** The distribution is highly skewed to the right. There is a very high density of sample weights concentrated at the lower end of the scale, close to zero.
- **Rapid Decay:** The density drops off very rapidly as the sample weight increases. This indicates that most observations in the dataset have relatively low sample weights.
- **Long Tail:** There is a long tail extending towards higher values of sample weight, indicating that while infrequent, some observations have considerably larger weights. These higher weights are necessary in survey analysis to ensure the sample accurately represents the population, often by giving more influence to underrepresented groups.
- **Single Peak:** The distribution has a single, prominent peak near zero, further emphasizing the concentration of lower sample weights.



- **Multimodal Distribution:** The plot displays a multimodal distribution with at least three distinct peaks, indicating that certain PSU values are more frequent in the dataset than others.
- **Peaks at Different PSU Values:** The peaks appear to be located roughly around PSU values of 20,000, 40,000, and 85,000. The peak around 20,000 seems to be the highest, suggesting that PSU has the highest density of observations.
- **Spread Across a Range:** The distribution of PSU values spans a considerable range, from near 0 to around 100,000.

- **Interpretation:** The presence of multiple peaks suggests that the sampling process likely involved selecting primary sampling units from different strata or geographical areas, and these selected units have varying numbers of observations. The different peaks might correspond to different clusters or regions within the surveyed population. The varying heights of the peaks indicate that some primary sampling units were sampled more frequently or contain a larger number of respondents than others.



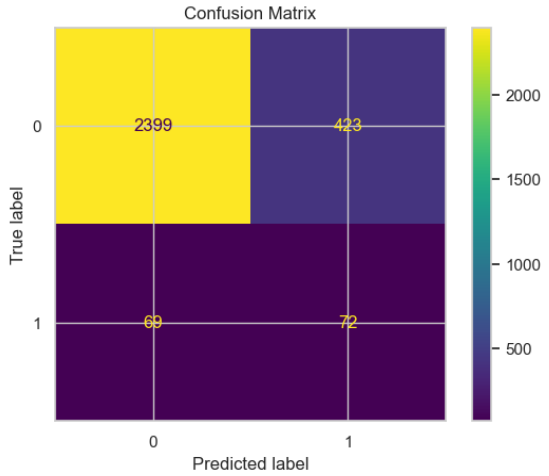
- **Multimodal Distribution:** Similar to the "psu" variable, the distribution of "strata" also shows multiple peaks, indicating that certain strata levels are more prevalent in the dataset.
- **Prominent Peaks:** There appear to be notable peaks around strata values of approximately 10,000, 40,000, and 85,000. The peak around 10,000 seems to have the highest density of observations.
- **Spread Across a Range:** The strata values are distributed across a wide range, from near 0 to around 100,000.
- **Relationship with PSU:** Given the strong positive correlation observed between "psu" and "strata" earlier, it's not surprising to see a multimodal distribution for "strata" as well. The peaks in the "strata" distribution might correspond to specific groupings or categories used for stratification during the survey design.
- **Interpretation:** The varying densities across different strata levels suggest that the number of sampled units or respondents within each stratum is not uniform. Some strata were likely designed to include a larger segment of the population or were oversampled for specific analytical purposes. The specific values of the peaks and their relative heights would be determined by the stratification scheme used in the survey.

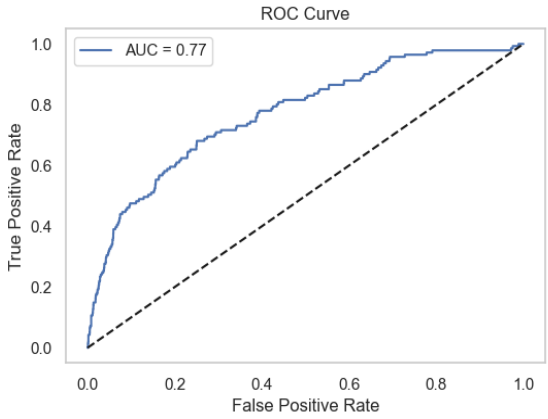
9. Box Plot:

- **Median:** The horizontal line within the box represents the median number of years of education for married individuals. It appears to be around 3 years.
- **Interquartile Range (IQR):** The box itself spans the interquartile range, containing the middle 50% of the data. The bottom of the box is around 2 years, and the top is around 5 years. This means that 50% of married individuals in this dataset have between 2 and 5 years of education.
- **Whiskers:** The lines extending from the box (the whiskers) show the spread of the rest of the data, excluding outliers. The bottom whisker extends down to 0 years, and the top whisker extends up to 8 years. This indicates the range of years of education for the majority of married individuals.
- **Outliers:** There are no individual points plotted outside the whiskers, which suggests there are no extreme outliers in the years of education for married individuals in this dataset.
- **Overall Distribution:** The box plot suggests that the distribution of years of education for married individuals is somewhat concentrated between 2 and 5 years, with the median slightly closer to the lower end of this range. The full range of education levels observed for married individuals is from 0 to 8 years.

Result of Model performance

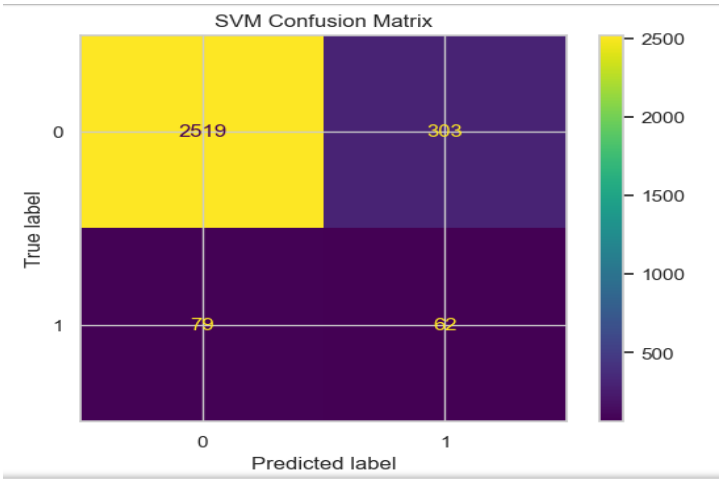
1. Logistic regression:

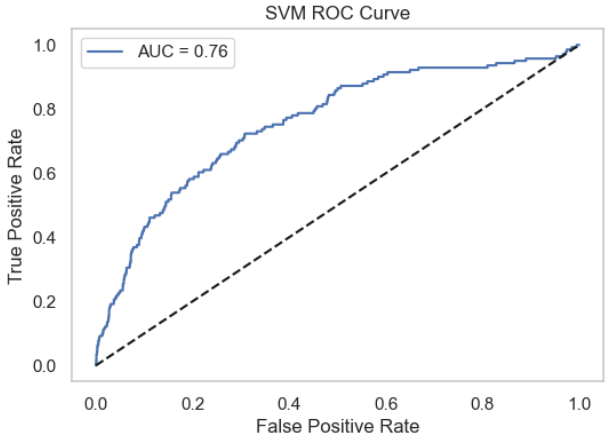




- **Confusion Matrix Breakdown:**
  True Negatives (TN) = 2399
  False Positives (FP) = 423
  False Negatives (FN) = 69
  True Positives (TP) = 72
- **Metric Interpretation:**
  Accuracy: 83.4% That is, the model accurately predicted the outcome in 83.4% of instances. But: Accuracy is deceptive in imbalanced datasets (which this appears to be).
  Precision: 14.5% Of all instances the model predicted as "1" (i.e., positive), only 14.5% actually were positive. Low precision indicates a lot of false positives.
  Recall: 51.1% The model found approximately half of the actual positive instances. This is moderate and indicates the model can identify some positive instances but misses a lot.
  F1 Score: 22.6% This is precision vs. recall balance. The low F1 shows the model is not able to classify the minority (positive) class well.
  AUC Score: 0.77 This is good. This shows how well the model classifies between classes in general. 0.77 means that the model has good discrimination power.

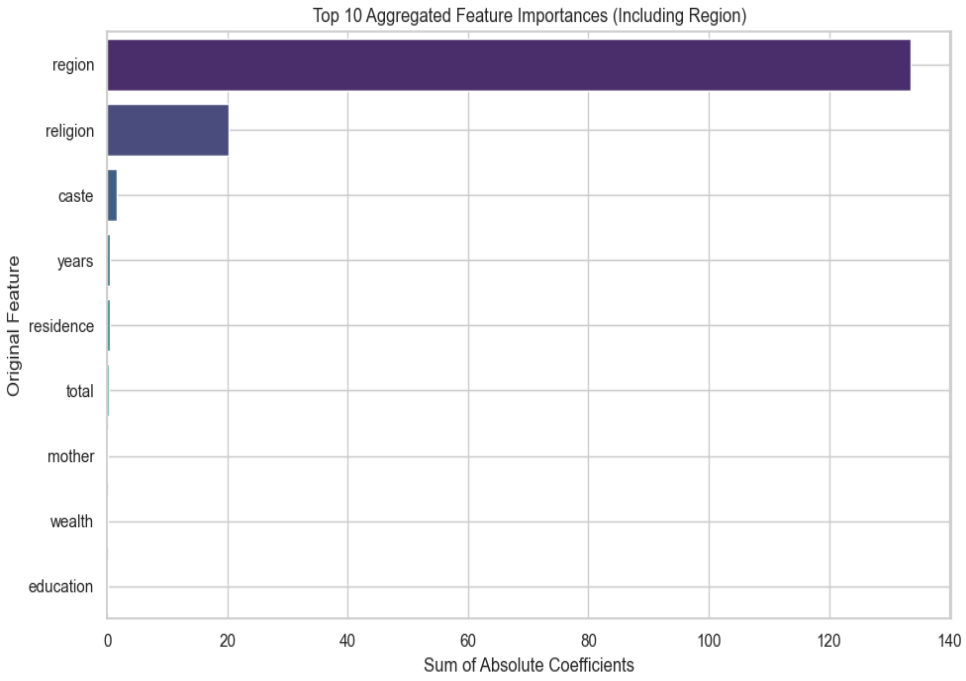
2. SVM Model:





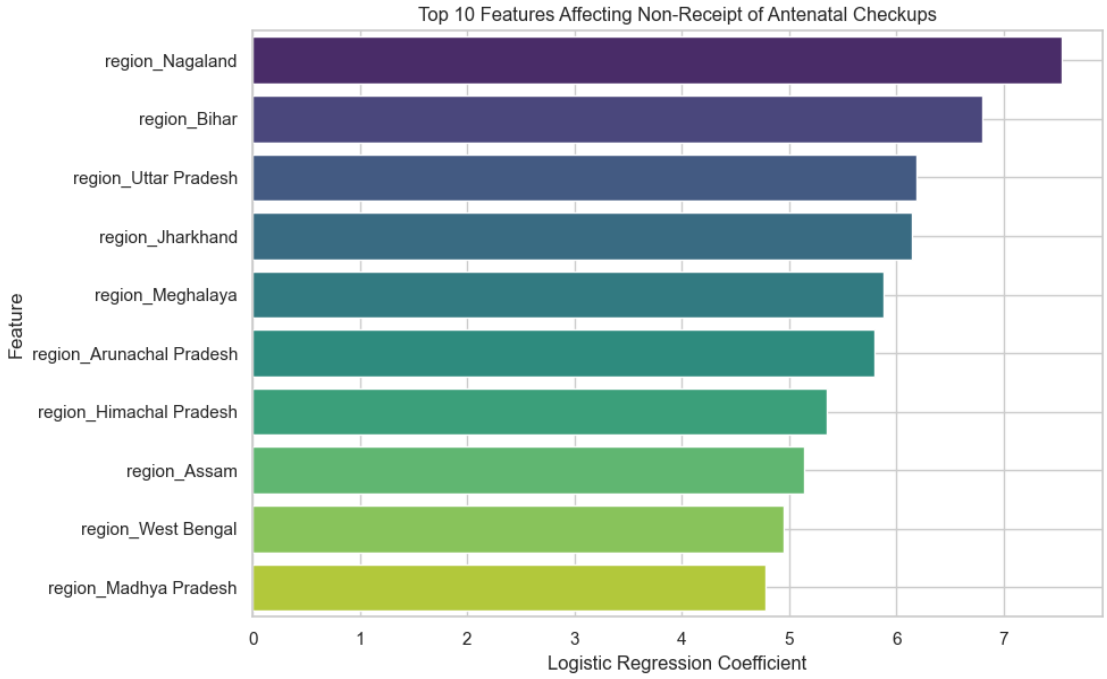
- **Confusion Matrix Breakdown:**
  True Negatives (TN) = 2519: The model correctly predicted class 0.
  False Positives (FP) = 303: Predicted class 1 but it was actually class 0.
  False Negatives (FN) = 79: Predicted class 0 but it was actually class 1.
  True Positives (TP) = 62: Correctly predicted class 1.
- **Metric Interpretation:**
  Accuracy: 87.1% The model predicted the correct class in 87.1% of cases. This high score reflects good general performance but may be skewed by the imbalance in class labels.
- **Precision**: 16.98% Only 17% of the predicted positives were actually correct. The model still makes a fair number of false positive errors.
- **Recall: 43.97%** The model was able to identify around 44% of the actual positive cases. Slightly lower than logistic regression but still captures some of the minority class.
- **F1 Score: 24.5%** Slightly better than logistic regression. Indicates moderate balance between precision and recall, but still far from ideal.
- **AUC Score: 0.76** The model has a 76% probability of ranking a randomly chosen positive instance higher than a negative one. This is a good level of discrimination.

FEATURE IMPORTANCE



Rank	Feature	Sum of Coefficients	Interpretation
1	Region	133.65	By far the most influential factor. Regional disparities dominate the likelihood of antenatal care non-receipt. This reinforces earlier findings that geography is the strongest determinant of maternal health access.
2	Religion	20.26	Religious background has a noticeable influence, possibly reflecting cultural or community-level health behaviors or beliefs.
3	Caste	1.66	Caste plays a smaller role, but still reflects social stratification and possible discrimination or marginalization in health service delivery.
4	Years (likely age-related)	0.45	Minimal effect; maternal age seems to have limited influence on checkup utilization.
5	Residence (urban/rural)	0.42	Surprisingly small, suggesting regional variation within urban/rural categories may overshadow general residence classification.
6	Total children ever born	0.37	A minor influence—more children may slightly reduce the likelihood of seeking ANC due to experience, fatigue, or resource constraints.
7	Mother-related variables (e.g., literacy)	0.16	Limited direct impact in the model, but may interact with other variables.

Rank	Feature	Sum of Coefficients	Interpretation
8	Wealth	0.08	Socioeconomic status has much less weight than expected—possibly due to overlap with region or religion effects.
9	Education	0.08	Similarly, low individual impact—may be explained by collinearity with region or wealth.



Rank	Region	Coefficient	Odds Ratio	Interpretation
1	Nagaland	7.54	1890.10	Women from Nagaland are 1890 times more likely to miss antenatal checkups than those from the reference region.
2	Bihar	6.80	901.64	Women from Bihar are 901 times more likely to not receive antenatal care.
3	Uttar Pradesh	6.19	486.00	Very high likelihood of missing antenatal care; significant public health concern.
4	Jharkhand	6.15	467.46	High risk of exclusion from ANC services.
5	Meghalaya	5.88	356.69	Strongly associated with lack of access/utilization of antenatal services.
6	Arunachal Pradesh	5.79	327.96	Indicates substantial barriers in maternal health outreach.

Rank	Region	Coefficient	Odds Ratio	Interpretation
7	Himachal Pradesh	5.35	211.03	More than 200 times likely to miss checkups compared to the reference.
8	Assam	5.14	169.89	Reflects potential systemic issues in ANC coverage.
9	West Bengal	4.95	141.30	Still highly significant contributor to underutilization.
10	Madhya Pradesh	4.78	118.55	Elevated risk of missing care, though comparatively lower than top 5.

CONCLUSION

The current study investigated the socioeconomic determinants of use of antenatal care (ANC) in India based on NFHS-5 data. The findings reinforce that maternal education, wealth status, place of residence, and exposure to the media have considerable impacts on uptake of ANC. Educated females had higher levels of full ANC usage, thus confirming the empowering role of information on health-seeking behaviour. As expected, wealthier households reported increased use of ANC services, implying economic deprivation in poorer sections.

Urban-rural inequalities were evident, with urban women having improved access, mirroring rural healthcare infrastructure gaps. Exposure to mass media also had a positive influence on ANC use, demonstrating its potential in disseminating information and encouraging health services. Age, birth order, and region also played a role in the variations, highlighting the importance of context-specific interventions.

The results show that ANC use is strongly linked to entrenched social and economic disparities. Thus, a multi-dimensional policy response is needed—one that targets female education, economic assistance to poor families, rural health system development, and good health communication through media. Local planning from NFHS data can maximize the coverage and effectiveness of such programs. In summary, addressing inequality and enhancing access in all regions and groups is the key to universal, equitable maternal health care in India.

DISCUSSION

The research study "Socioeconomic Determinants of Antenatal Care Utilization in India: An NFHS-Based Analysis" explores what drives antenatal care (ANC) services among women in India, based on National Family Health Survey (NFHS) data. The research highlights ANC as critical to producing healthy mothers and infants, while recognizing continuing inequities in the use of such services in India. The research methodology follows a systematic procedure, beginning with data pre-processing and cleaning to exploratory data analysis and statistical modelling. Socioeconomic factors such as education level, wealth index, type of residence, and demographic factors like age of the mother, caste, and religion are analysed to identify their effects on ANC usage. The results emphasized the substantial role of education and wealth in ANC use prediction, as well as the higher rates of utilization among more educated and wealthier women. On the contrary, women from rural backgrounds, certain religious or caste groups experience structural and social inequalities in accessing proper antenatal care, showing the existence of inequalities.

The study concludes by proposing targeted interventions for reducing these inequalities and enhancing the outcomes of maternal healthcare in India. Through the identification of determinants of ANC use, the research offers significant information for policymakers to develop effective strategies that foster equitable access to care, in turn supporting the Sustainable Development Goal 3 to ensure healthy lives and well-being for all.

REFERENCES

1. Government of India. (2021). National Family Health Survey (NFHS-5), 2019-21 International Institute for Population Sciences. https://mohfw.gov.in/sites/default/files/NFHS-5_Phase-II_0.pdf
2. World Health Organization. (2016). WHO recommendations on antenatal care for a positive pregnancy experience. World Health Organization. https://www.who.int/publications/i/item/9789241549912
3. National Health Mission. (2020). Guidelines for antenatal care and institutional deliveries in India. Ministry of Health and Family Welfare, Government of India. https://nhm.gov.in/images/pdf/programmes/maternal-health/guidelines/sba_guidelines_for_skilled_attendance_at_birth.pdf
4. Mishra, P. S., Veerapandian, K., & Choudhary, P. K. (2021). Impact of socio-economic inequity in access to maternal health benefits in India: Evidence from Janani Suraksha Yojana using NFHS data. *Plos one*, *16*(3), e0247935.8. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0247935
5. Girotra, S., Malik, M., Roy, S., & Basu, S. (2023). Utilization and determinants of adequate quality antenatal care services in India: evidence from the National Family Health Survey (NFHS-5)(2019-21). *BMC Pregnancy and Childbirth*, *23*(1), 800. https://link.springer.com/article/10.1186/s12884-023-06117-z
6. Singh Sardar, S., Bhattacharya, S., & Mandal, M. (2025). Exploring the Factors Influencing Antenatal Care Utilization in India: A Study on Socioeconomic and Caste Disparities, with Logistic Regression Analysis and Outlier Detection Using NFHS-5 Data. *Global Social Welfare*, 1-9.13. Chaudhary, R., & Mohanty, S. K (2020). Regional disparities in maternal healthcare in India: An NFHS analysis. PLOS ONE, 15(6), e0233429. https://link.springer.com/article/10.1007/s40609-025-00375-9
7. UNICEF. (2015). UNICEF Annual Report 2015. New York. Retrieved from https://www.unicef.org/reports/unicef-annual-report-2015
8. UNICEF Antenatal care : https://data.unicef.org/topic/maternal-health/antenatal-care/
9. ANTENATAL CARE RESOURCES: https://www.unicef.org.uk/babyfriendly/baby-friendly-resources/antenatal-care-resources/
10. Sheikhtaheri, A., Zarkesh, M. R., Moradi, R., & Kermani, F. (2021). Prediction of neonatal deaths in NICUs: development and validation of machine learning models. BMC Medical Informatics and Decision Making, 21(1), 131. https://doi.org/10.1186/s12911-021-01497-8
11. Sullivan, E. (2022). Understanding from machine learning models. *The British Journal for the Philosophy of Science*. https://www.journals.uchicago.edu/doi/abs/10.1093/bjps/axz035?journalCode=bjps
12. Varoquaux, G., & Colliot, O. (2023). Evaluating machine learning models and their diagnostic value. *Machine learning for brain disorders*, 601-630. https://link.springer.com/protocol/10.1007/978-1-0716-3195-9_20
13. Bishop, C. M. (2013). Model-based machine learning. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *371*(1984), 20120222. https://royalsocietypublishing.org/doi/full/10.1098/rsta.2012.0222

14. SDG Goals: https://sdgs.un.org/goals
15. Salazar, D. A., Vélez, J. I., & Salazar, J. C. (2012). Comparison between SVM and logistic regression: Which one is better to discriminate?. *Revista Colombiana de Estadística*, *35*(2), 223-237. https://www.redalyc.org/pdf/899/89923144003.pdf
16. Musa, A. B. (2013). Comparative study on classification performance between support vector machine and logistic regression. *International Journal of Machine Learning and Cybernetics*, *4*, 13-24. https://link.springer.com/article/10.1007/s13042-012-0068-x
17. Chang, Y. C. I. (2003). Boosting SVM classifiers with logistic regression. *See www. stat. sinica. edu. tw/library/c_tec_rep/2003-03. pdf*. https://d1wqtxts1xzle7.cloudfront.net/54268253
18. Villar, J., Ba'aqeel, H., Piaggio, G., Lumbiganon, P., Belizán, J. M., Farnot, U., ... & Berendes, H. (2001). WHO antenatal care randomised trial for the evaluation of a new model of routine antenatal care. *The lancet*, *357*(9268), 1551-1564. https://www.thelancet.com/journals/lancet/article/PIIS014067360004722X/fulltext