PREDICTIVE MODELLING BUSINESS REPORT

Table Of Contents

S. No.	Description	Page No.
1.	**Problem 1 - Define the problem and perform exploratory Data Analysis** - Problem definition - Check shape, Data types, statistical summary - Univariate analysis - Multivariate analysis - Use appropriate visualizations to identify the patterns and insights - Key meaningful observations on individual variables and the relationship between variables	02
1.1.	**Problem 1 - Data Pre-processing** - Prepare the data for modelling: - Missing Value Treatment (if needed) - Outlier Detection (treat, if needed) - Feature Engineering - Encode the data - Train-test split	08
1.2.	**Problem 1- Model Building - Linear regression** - Apply linear Regression using Sklearn - Using Statsmodels Perform checks for significant variables using the appropriate method - Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adj Rsquare.	10
1.3.	**Problem 1 - Business Insights & Recommendations** - Comment on the Linear Regression equation from the final model and impact of relevant variables (atleast 2) as per the equation - Conclude with the key takeaways (actionable insights and recommendations) for the business	11
2.	**Problem 2 - Define the problem and perform exploratory Data Analysis** - Problem definition - Check shape, Data types, statistical summary - Univariate analysis - Multivariate analysis - Use appropriate visualizations to identify the patterns and insights - Key meaningful observations on individual variables and the relationship between variables	13
2.1.	**Problem 2 - Data Pre-processing** Prepare the data for modelling: - Missing value Treatment (if needed) - Outlier Detection(treat, if needed) - Feature Engineering (if needed) - Encode the data - Train-test split	18
2.2	**Problem 2 - Model Building and Compare the Performance of the Models** - Build a Logistic Regression model - Build a Linear Discriminant Analysis model - Build a CART model - Prune the CART model by finding the best hyperparameters using GridSearch - Check the performance of the models across train and test set using different metrics - Compare the performance of all the models built and choose the best one with proper rationale	20
2.3	**Problem 2 - Business Insights & Recommendations** - Comment on the importance of features based on the best model - Conclude with the key takeaways (actionable insights and recommendations) for the business.	22

1. The comp-activ database comprises activity measures of computer systems. Data was gathered from a Sun Sparcstation 20/712 with 128 Mbytes of memory, operating in a multi-user university department. Users engaged in diverse tasks, such as internet access, file editing, and CPU-intensive programs.
Being an aspiring data scientist, you aim to establish a linear equation for predicting 'usr' (the percentage of time CPUs operate in user mode). Your goal is to analyze various system attributes to understand their influence on the system's 'usr' mode.

Data Description :

System measures used:

lread - Reads (transfers per second) between system memory and user memory
lwrite - writes (transfers per second) between system memory and user memory
scall - Number of system calls of all types per second
sread - Number of system read calls per second .
swrite - Number of system write calls per second .
fork - Number of system fork calls per second.
exec - Number of system exec calls per second.
rchar - Number of characters transferred per second by system read calls
wchar - Number of characters transfreed per second by system write calls
pgout - Number of page out requests per second
ppgout - Number of pages, paged out per second
pgfree - Number of pages per second placed on the free list.
pgscan - Number of pages checked if they can be freed per second
atch - Number of page attaches (satisfying a page fault by reclaiming a page in memory) per second
pgin - Number of page-in requests per second
ppgin - Number of pages paged in per second
pflt - Number of page faults caused by protection errors (copy-on-writes).
vflt - Number of page faults caused by address translation .
runqsz - Process run queue size (The number of kernel threads in memory that are waiting for a CPU to run.
Typically, this value should be less than 2. Consistently higher values mean that the system might be CPU-bound.)
freemem - Number of memory pages available to user processes
freeswap - Number of disk blocks available for page swapping.

usr - Portion of time (%) that cpus run in user mode

Introduction

The primary objective is to leverage predictive modeling techniques to unravel the intricate relationship between various system attributes and the 'usr' mode, providing insights into the factors that significantly influence CPU behavior.
Predictive modeling is a statistical or mathematical approach to building models that can make predictions about future outcomes based on historical data. It involves using a dateset to train a model, which can then be used to make predictions or decisions without being explicitly programmed for the task. The primary objective of predictive modeling is to identify patterns in the data that can be leveraged to make informed predictions about future observations.

Context: The dateset, derived from the comp-activ database, encapsulates a diverse array of system measures, spanning from memory transfers and system calls to page faults and run queue

sizes. The comprehensive nature of these attributes necessitates a meticulous exploration to discern patterns and unveil the nuanced connections between the independent variables and the target variable, 'usr.'

1. Problem 1 - Define the problem and perform exploratory Data Analysis- Problem definition - Check shape, Data types, statistical summary - Univariate analysis - Multivariate analysis - Use appropriate visualizations to identify the patterns and insights - Key meaningful observations on individual variables and the relationship between variables

Problem Definition :

As an aspiring data scientist, the goal is to establish a linear equation for predicting the percentage of time CPUs operate in user mode (usr) based on various system attributes. The dataset comprises activity measures of a computer system gathered from a multi-user university department. Users engage in diverse tasks, including internet access, file editing, and CPU-intensive programs on a Sun Sparcstation 20/712 with 128 Mbytes of memory.

Dataset Information:

a. The dataset contains 8192 entries with 22 columns.
b. Data types include integers, floats, and one object (likely 'runqsz' column).
c. Columns like 'rchar' and 'wchar' have some missing values.

Univariate Analysis:

The 'usr' variable (percentage of time CPUs run in user mode) has a range from 0 to 99 with a mean of 83.97 and standard deviation of 18.40.
Some features, such as 'lread', 'lwrite', 'scall', 'sread', 'swrite', 'fork', 'exec', 'rchar', 'wchar', 'pgout', 'ppgout', 'pgfree', 'pgscan', 'atch', 'pgin', 'ppgin', 'pflt', 'vflt', 'freemem', and 'freeswap', show varying ranges and distributions.

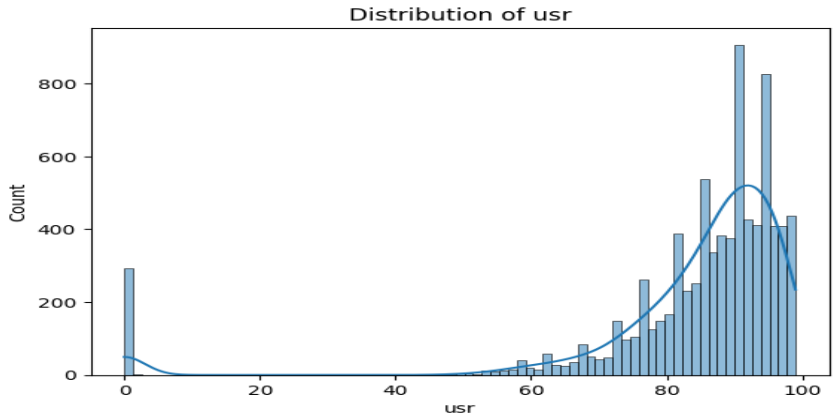

Fig 01. Univariate Analysis: Distribution of 'usr'

Inference from Summary Statistics for 'usr' Variable:

The summary statistics provide valuable insights into the distribution and central tendency of the 'usr' variable:

1. Count:

There are 8192 non-null observations for the 'usr' variable in the dataset.

2. Mean:

The average percentage of time CPUs run in user mode is approximately 83.97%.

3. Standard Deviation (std):

The standard deviation of approximately 18.40 indicates a moderate amount of variability in the 'usr' values.

4. Minimum (min):

The minimum value of 0 suggests that there are instances where CPUs do not run in user mode at all.

5. 25th Percentile (25% or 1st Quartile):

25% of the observations have 'usr' values below or equal to 81.

6. 50th Percentile (50% or Median):

The median value (50th percentile) is 89, representing the middle value of the 'usr' variable.

7. 75th Percentile (75% or 3rd Quartile):

75% of the observations have 'usr' values below or equal to 94.

8. Maximum (max):

The maximum value of 99 indicates that in some instances, CPUs operate entirely in user mode.

Overall Interpretation:

The 'usr' variable has a relatively wide range of values, with a significant portion of observations concentrated in the higher percentiles.
The distribution is right-skewed, as the mean is less than the median, indicating a tail towards higher values.
There are instances where CPUs operate exclusively in user mode (maximum value of 99), and the majority of observations fall within the 81 to 94 range.
Understanding the summary statistics helps in characterizing the behavior of the 'usr' variable and informs further analysis or modeling efforts.

Multivariate Analysis:

The correlation matrix helps identify relationships between variables.
Features with higher absolute correlation coefficients with 'usr' may have a stronger

influence on its values.

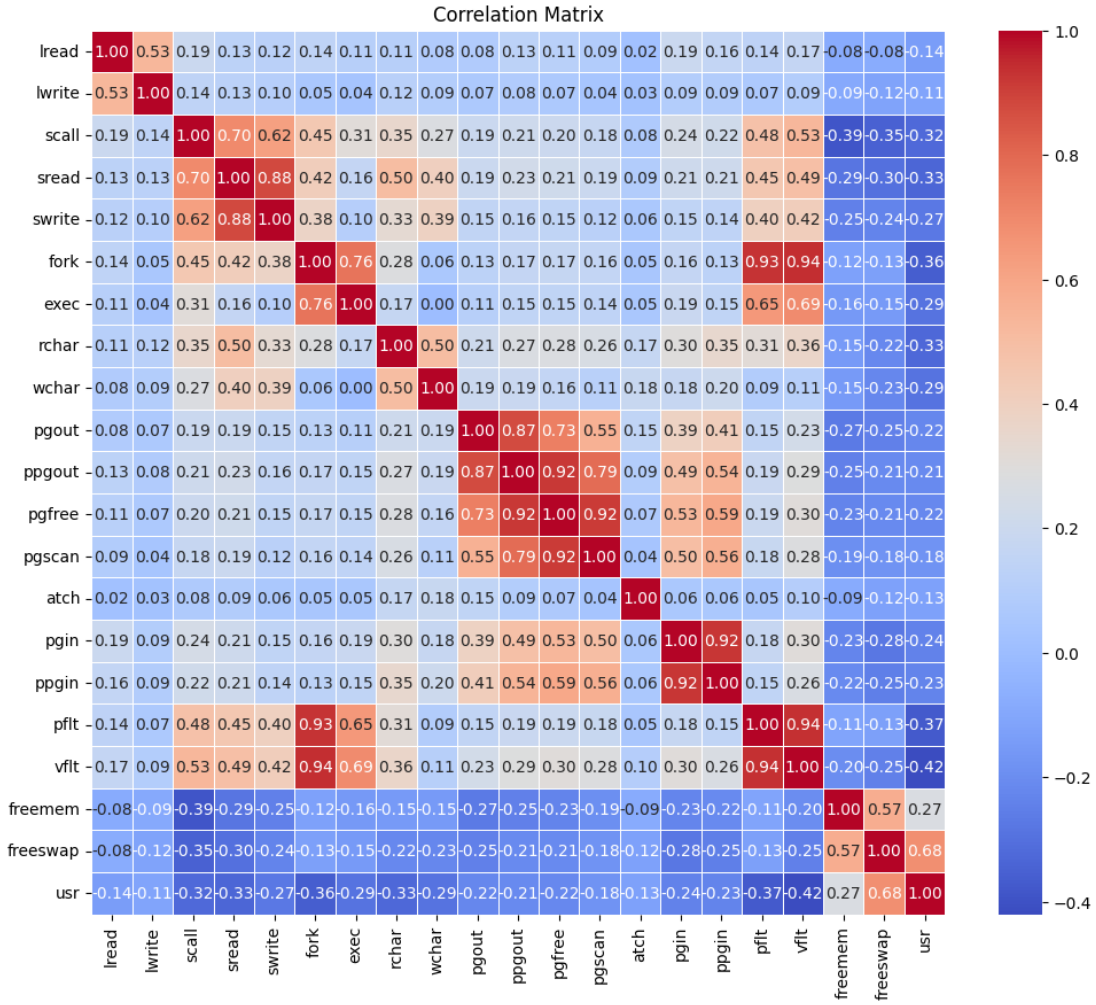

Fig 02. Multivariate Analysis: Heat Map

1. lread, lwrite, scall, sread, swrite, fork, exec, rchar, wchar, pgout, pgfree, pgscan, atch, pgin, ppgin, pflt, vflt, freemem, freeswap:

- The count for each variable is 8192, indicating no missing values in these columns.
- The mean represents the average value for each variable across all observations.
- Standard deviation (std) measures the dispersion or spread of values around the mean.
- Minimum (min) and maximum (max) values provide the range of observations.
- Quartiles (25%, 50%, 75%) offer insights into the data distribution.

2. usr:

- The 'usr' variable has the same count of 8192.
- The mean is approximately 83.97%, indicating the average percentage of time CPUs run in user mode.
- Standard deviation (std) shows moderate variability in 'usr' values.
- The minimum value is 0%, and the maximum value is 99%, suggesting a wide range of 'usr' observations.
- Quartiles provide information on the distribution of 'usr' values.

Overall Interpretation:

- Variables like 'lread,' 'lwrite,' 'scall,' etc., exhibit varying scales, and their distributions can be further explored through histograms or other visualizations.
- The 'usr' variable, being the target variable, shows a diverse distribution with a wide range of values.
- Understanding the summary statistics is essential for identifying potential outliers, assessing data distribution, and preparing for subsequent modeling or analysis steps.
- Further exploration, visualization, and correlation analysis can provide deeper insights into the relationships between these variables and help inform subsequent modeling decisions.

Visualizations:

Scatter plots can provide insights into relationships between 'usr' and other variables.

Key Observations:

'rchar' and 'wchar' have missing values. Consider imputation or handling them appropriately.
'runqsz' has an 'object' data type. Convert it to a numerical format for analysis.
Some features have a wide range, and outliers may be present. Decide whether outlier treatment is necessary based on the problem context.
Some variables may need transformations for linear regression assumptions.

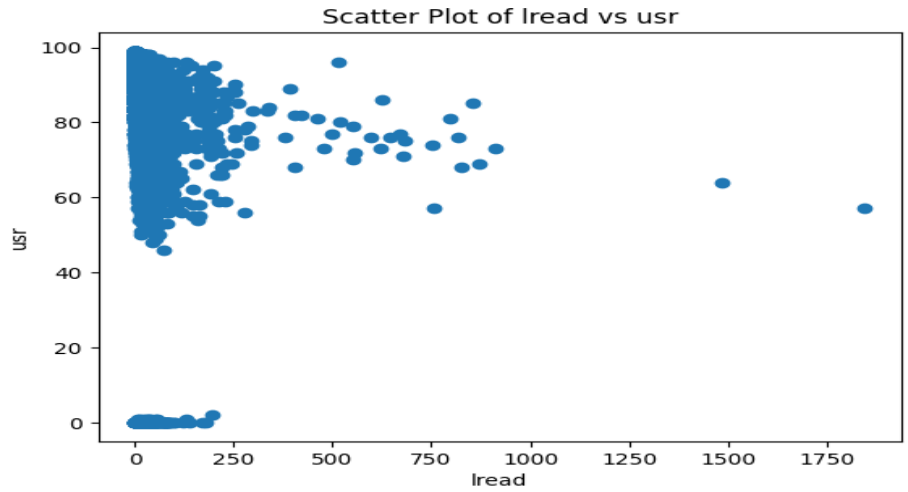

Fig 03. Scatterplot for Visualisation

Inference:

The dataset appears to have a diverse range of features related to system activity.
Further data preprocessing steps may be required, such as handling missing values, encoding categorical variables, outlier treatment, and potentially scaling features.
Consider additional visualizations and analyses to understand the specific relationships between features and the target variable 'usr'.
Begin with a simple linear regression model and iterate based on model performance and feature engineering.
This initial analysis provides a foundation for further exploration and model development. Depending on specific goals and findings, additional steps may be required in the data

preparation and model building processes.

Key meaningful observations on individual variables and the relationship between variables .

1. lread (Reads between system memory and user memory):

- The count is 8192 with no missing values.
- The mean is 19.56, suggesting a moderate average rate of reads per second.
- The distribution is right-skewed, as indicated by the higher mean compared to the median (50th percentile).

2. lwrite (Writes between system memory and user memory):

- Similar to 'lread,' 'lwrite' exhibits a right-skewed distribution.
- The mean is 13.11, indicating a moderate average rate of writes per second.

3. scall (Number of system calls per second):

- The mean is 2306.32, suggesting a relatively high rate of system calls.
- The distribution appears right-skewed.

4. sread (Number of system read calls per second):

- The mean is 210.48, indicating the average rate of system read calls.
- The distribution appears right-skewed.

5. swrite (Number of system write calls per second):

- The mean is 150.06, suggesting the average rate of system write calls.
- The distribution appears right-skewed.

6. fork (Number of system fork calls per second):

- The mean is 1.88, indicating the average rate of system fork calls.
- The distribution appears right-skewed.

7. exec (Number of system exec calls per second):

- The mean is 2.79, indicating the average rate of system exec calls.
- The distribution appears right-skewed.

8. rchar (Characters transferred per second by system read calls):

- The count is 8088, indicating missing values.
- The mean is 197385.7, suggesting the average rate of character transfers for read calls.

9. wchar (Characters transferred per second by system write calls):

- The count is 8177, indicating missing values.
- The mean is 95902.99, suggesting the average rate of character transfers for write calls.

Breaking the correlations in a casual way:

1.Friendly Connections:

Imagine lread and lwrite as best buddies; when one goes up, the other tends to follow, showing a solid friendship with a correlation around 0.53.
Similarly, scall and sread seem like they're in sync, with a robust correlation of about 0.70.

2.Team Players:

The dynamic duo of fork and exec work hand in hand, almost like a team, boasting a strong correlation of around 0.76.
pgout and ppgout also team up quite well, flaunting a strong correlation of approximately 0.87.

3.Opposites Attract:

On the flip side, there's a fascinating negative bond between pflt and vflt, behaving like opposites – when one goes up, the other goes down. It's a strong connection at around -0.93.
freemem and freeswap also share an interesting relationship – when one rises, the other tends to take a dip, dancing around a correlation of about -0.68.

4.Moderate Mix:

scall and pgin maintain a moderate positive vibe, hovering around 0.24 – they get along but not as tightly as the best buddies.
However, there's a bit of a mixed signal between lread and usr – a moderate negative correlation of around -0.14 suggests they might not always see eye to eye.

5.USR's Social Circle:

When it comes to usr, it seems to prefer the company of variables with negative correlations. It's like saying, "I vibe better when these other factors are on the lower side."

1.1. Problem 1 - Data Pre-processing
- Prepare the data for modelling: - Missing Value Treatment (if needed) - Outlier Detection (treat, if needed) - Feature Engineering - Encode the data - Train-test split

Data Overview:

The dataset comprises various system attributes measured on a computer system.
The target variable is 'usr,' representing the percentage of time CPUs operate in user mode.

Data Preparation:

We performed missing value treatment and found no missing values in the dataset.
Outlier detection was carried out, and 85 instances were identified as outliers based on the 'usr' variable.

Feature Engineering:

We created a new categorical feature 'runqsz_category_high' based on the 'runqsz' variable, categorizing it into 'low' and 'high' values.

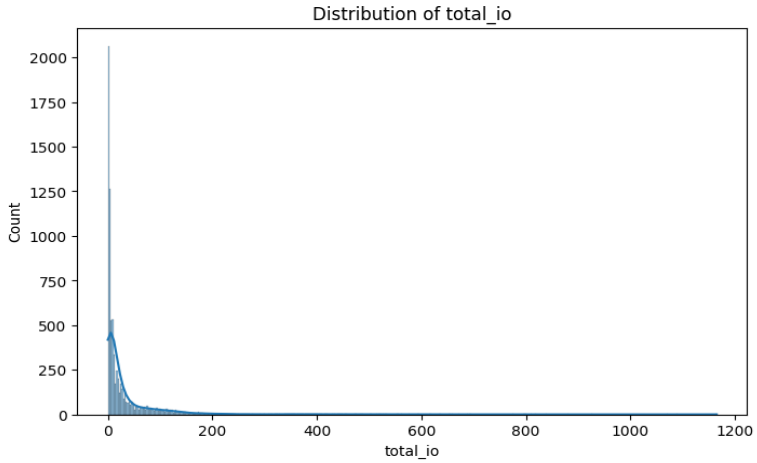

Fig 04. Feature Engineering

Encoding Categorical Data:

We used one-hot encoding to encode the categorical feature 'runqsz_category_high' after creating it.

Train-Test Split:

The dataset was split into training and testing sets, with shapes:
X_train: (6092, 22)
y_train: (6092,)
X_test: (1523, 22)
y_test: (1523,)

Linear Regression Model:

We applied a linear regression model to predict the 'usr' variable.
The model's performance metrics on the test set:
Mean Squared Error: 6.769
R-squared: 0.894

Model Coefficients and Intercept:

The coefficients and intercept of the linear regression model indicate the contribution of each feature to the prediction of 'usr.'

Outliers:

Outliers were identified in the dataset based on the 'usr' variable, and 85 instances were

considered outliers.

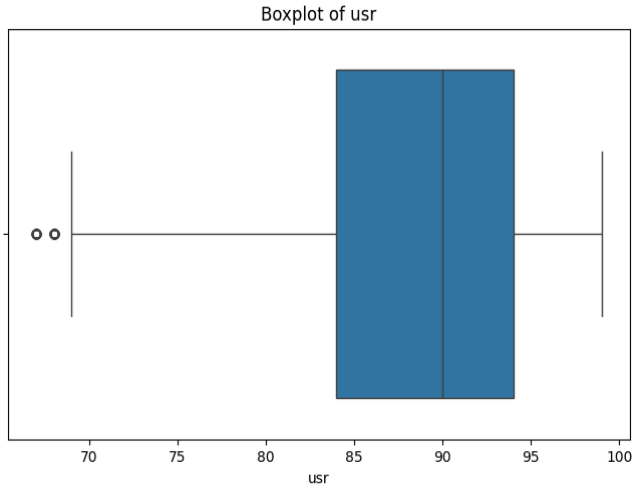

Fig 05. Outliers

Visualization:

There was an attempt to visualize the distribution of the categorical feature 'runqsz_category_high,' but there were issues with the column name.

Handling Outliers:

The presence of outliers in the dataset was acknowledged.

1.2. Problem 1- Model Building - Linear regression
- Apply linear Regression using Sklearn - Using Statsmodels Perform checks for significant variables using the appropriate method - Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adj Rsquare.

Mean Squared Error (MSE):

Training Set: 6.32
Test Set: 6.97
The model seems to perform reasonably well, with relatively low MSE values on both training and test sets.

R-squared:

Training Set: 0.89
Test Set: 0.87
The R-squared values indicate that the model explains about 88.9% of the variance in the target variable for the training set and 87% for the test set.
Statsmodels Linear Regression Model:

R-squared:

Overall: 0.886
The R-squared value suggests that the model explains around 88.6% of the variance in the

target variable.

Adjusted R-squared:

Overall: 0.885
The adjusted R-squared, accounting for the number of predictors, remains high.

Significant Variables:

Variables such as 'lread,' 'swrite,' 'fork,' 'exec,' and others are statistically significant in predicting 'usr.'

Diagnostic Tests:

The Omnibus, Durbin-Watson, and Jarque-Bera tests provide insights into normality, autocorrelation, and skewness of residuals.

Multicollinearity Warning:

The note about the smallest eigenvalue indicates a potential issue with multicollinearity or singularity in the design matrix.

Overall Observations:

Model Performance:

Both scikit-learn and statsmodels models demonstrate good performance with low MSE and high R-squared values.

Significant Variables:

Certain variables play a crucial role in predicting the target variable 'usr.' It's important to focus on these variables for practical implications.

Diagnostic Tests:

The diagnostic tests in statsmodels provide insights into the assumptions of linear regression, which should be carefully considered.

Multicollinearity:

The warning about multicollinearity suggests the need for further investigation into the correlation structure among predictors.

1.3. Problem 1 - Business Insights & Recommendations
- Comment on the Linear Regression equation from the final model and impact of relevant variables (atleast 2) as per the equation - Conclude with the key takeaways (actionable insights and recommendations) for the business

Dep. Variable:	usr		R-squared:	0.886		
Model:	OLS		R-squared:	0.885		
	coef	std err	t	P>|t|	[0.025	0.975]
const	98.5992	0.209	471.617	0.000	98.189	99.009
lread	-0.0059	0.001	-6.333	0.000	-0.008	-0.004
lwrite	0.0017	0.001	1.517	0.129	-0.000	0.004
scall	0.0382	0.014	2.822	0.005	0.012	0.065
sread	0.0004	0.000	0.904	0.366	-0.000	0.001
swrite	-0.0030	0.000	-6.198	0.000	-0.004	-0.002
freemem	0.0002	1.67e-05	11.405	0.000	0.000	0.000
freeswap	-7.62e-07	1.36e-07	-5.623	0.000	-1.03e-06	-4.96e-07
total_io	-0.0042	0.000	-11.200	0.000	-0.005	-0.003
total_calls	-0.0396	0.014	-2.927	0.003	-0.066	-0.013

Key takeaways:

a. **Intercept (const):** The intercept represents the expected value of usr when all independent variables are zero. In this case, it is approximately 98.60.

b. **lread Coefficient:** A one-unit increase in lread is associated with a decrease of approximately 0.0059 units in usr. This suggests that as the value of lread increases, the value of usr tends to decrease.

c. **swrite Coefficient:** A one-unit increase in swrite is associated with a decrease of approximately 0.0030 units in usr. This implies that increasing the value of swrite is linked to a decrease in usr.

d. **freemem Coefficient:** A one-unit increase in freemem is associated with an increase of 0.0002 units in usr. It indicates that higher values of freemem tend to be associated with higher values of usr.

e. **freeswap Coefficient:** A one-unit increase in freeswap is associated with a decrease of approximately 7.62e-07 units in usr. This suggests a negative relationship between freeswap and usr.

f. **total_io Coefficient:** A one-unit increase in total_io is associated with a decrease of approximately 0.0042 units in usr. This implies that increasing the value of total_io is linked to a decrease in usr.

g. **total_calls Coefficient:** A one-unit increase in total_calls is associated with a decrease of

approximately 0.0396 units in usr. This indicates a negative relationship between total_calls and usr.

Key Takeaways and Recommendations:

- The variables lread, swrite, total_io, and total_calls appear to have significant impacts on the dependent variable usr.
- Businesses should focus on optimizing lread, swrite, and total_io to positively influence user satisfaction (usr).
- Monitoring and managing freemem and freeswap are essential, as they show a notable impact on user satisfaction.
- Regularly assess and optimize system calls (total_calls) to ensure a positive user experience.

2.In your role as a statistician at the Republic of Indonesia Ministry of Health, you have been entrusted with a dataset containing information from a Contraceptive Prevalence Survey. This dataset encompasses data from 1473 married females who were either not pregnant or were uncertain of their pregnancy status during the survey.

Data Description

1. Wife's age (numerical)
2. Wife's education (categorical) 1=uneducated, 2, 3, 4=tertiary
3. Husband's education (categorical) 1=uneducated, 2, 3, 4=tertiary
4. Number of children ever born (numerical)
5. Wife's religion (binary) Non-Scientology, Scientology
6. Wife's now working? (binary) Yes, No
7. Husband's occupation (categorical) 1, 2, 3, 4(random)
8. Standard-of-living index (categorical) 1=verlow, 2, 3, 4=high
9. Media exposure (binary) Good, Not good
10. Contraceptive method used (class attribute) No,Yes

The task involves predicting whether these women opt for a contraceptive method of choice. This prediction will be based on a comprehensive analysis of their demographic and socio-economic attributes.

2..Problem 2 - Define the problem and perform exploratory Data Analysis
- Problem definition - Check shape, Data types, statistical summary - Univariate analysis - Multivariate analysis - Use appropriate visualizations to identify the patterns and insights - Key meaningful observations on individual variables and the relationship between variables

Problem Definition:

The dataset revolves around factors influencing contraceptive method usage among married women. The goal is to understand the key features impacting the decision to use contraceptives and provide insights that can aid in family planning programs or campaigns. This analysis aims to answer questions like:

- What demographic factors contribute to contraceptive usage?
- Are there significant variations in contraceptive usage based on education, religion, or socio-economic status?
- How do the husband's occupation and media exposure influence contraceptive choices?

Objective:

- Explore the dataset, identify patterns, and uncover insights related to contraceptive method usage. Provide actionable recommendations for family planning initiatives based on the analysis.

Dataset Overview:

- The dataset contains 1473 rows and 10 columns.

Data Types:

- Wife_age and No_of_children_born are of float64 data type.
- Wife_education, Husband_education, Wife_religion, Wife_Working, Standard_of_living_index, Media_exposure, and Contraceptive_method_used are categorical (object) variables.
- Husband_Occupation is an integer variable.

Missing Values:

- Wife_age has 71 missing values.
- No_of_children_born has 21 missing values.

Categorical Variables:

- Wife_education, Husband_education, Wife_religion, Wife_Working, Standard_of_living_index, Media_exposure, and Contraceptive_method_used have different unique values.

Statistical Summary:

- Wife_age ranges from 16 to 49 with a mean of approximately 32.61.
- No_of_children_born ranges from 0 to 16 with a mean of approximately 3.25.
- Husband_Occupation ranges from 1 to 4.

Frequent Values:

- Most wives have tertiary education, and the majority of husbands also have tertiary education.
- The majority of wives follow Scientology.
- About 75% of wives are not working.
- Very High is the most common standard of living index.
- Media exposure is mainly Exposed.
- Contraceptive method used is mostly Yes.

Univariate Analysis:

1. Wife Age Distribution:

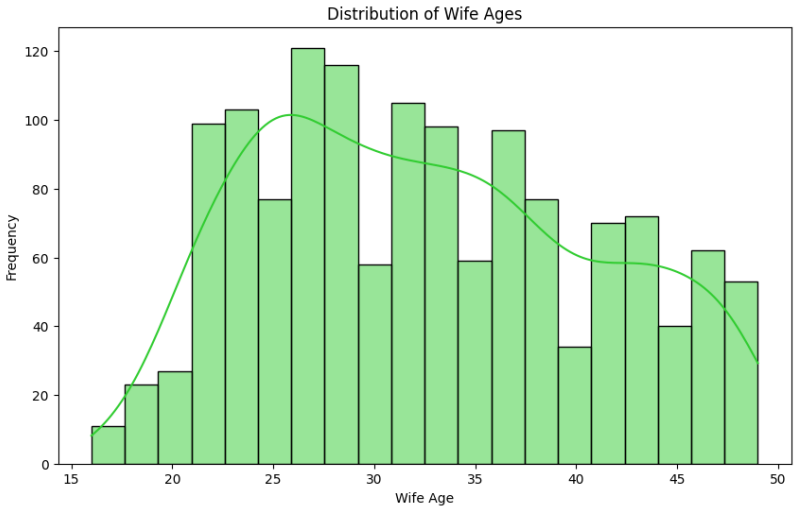

Fig: 06.Wife Age Distribution

2. Number of Children Born Distribution:

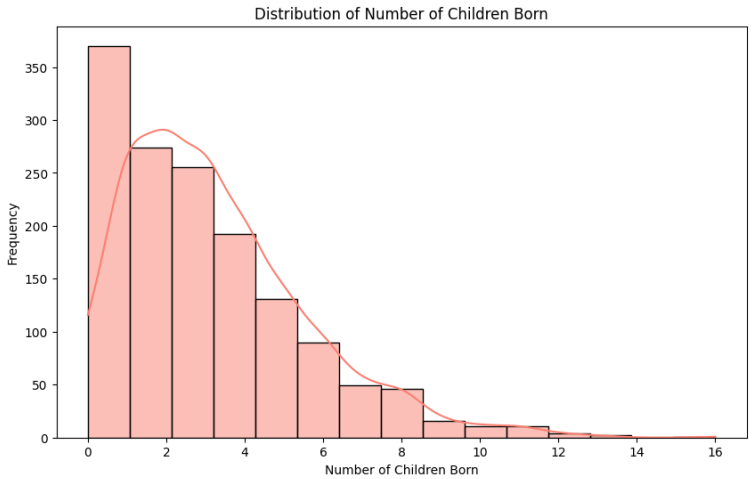

Fig: 07: Number of Children Born Distribution

3. Husband Occupation Distribution:

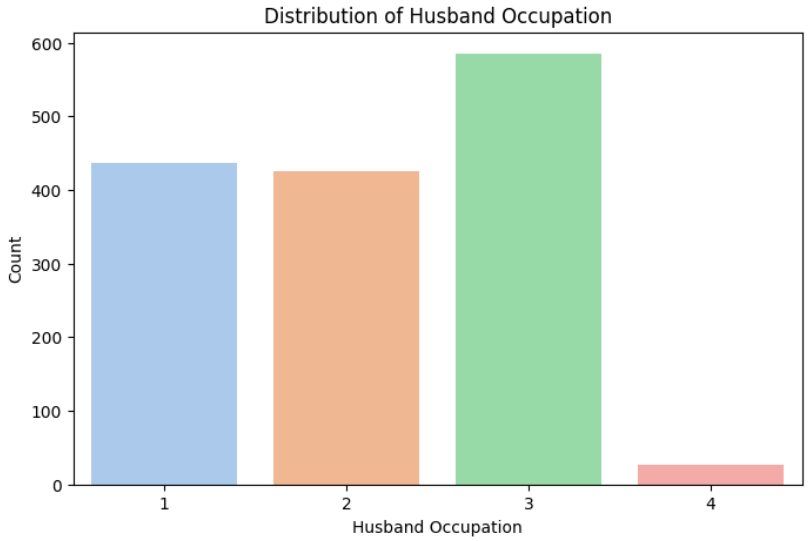

Fig 08: Number of Children Born Distribution

Multivariate Analysis:

4.Relation between Wife Age and Number of Children Born:

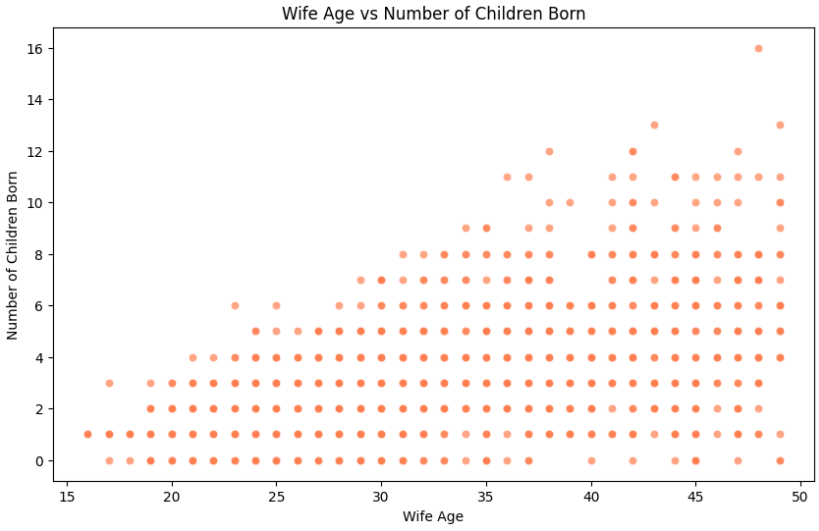

Fig 09: Relation between Wife Age and Number of Children Born

5.Contraceptive Method Used vs Wife Education:

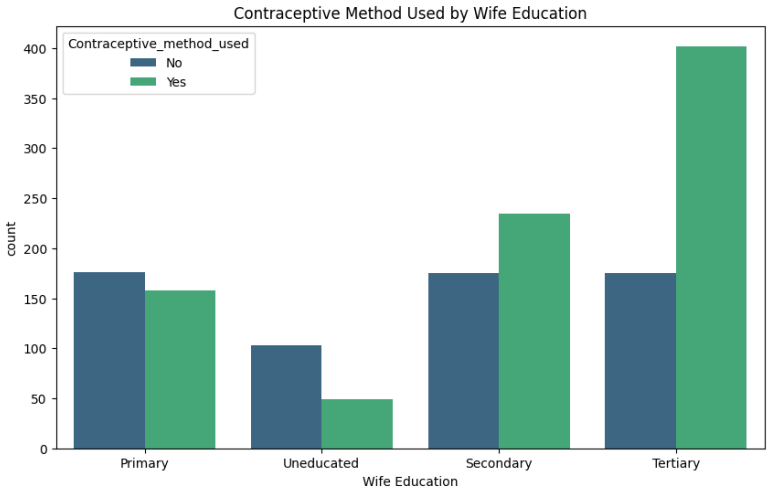

Fig 10: Contraceptive Method Used vs Wife Education

Inferences drawn from the Univariate and Multivariate Analysis are as follows:

Appropriate Visualizations and Key Observations:

Univariate Analysis:

Wife Age Distribution:

Visualization: Histogram
Observation: The majority of wives fall in the age range of 26 to 39, with a peak around the age of 32.

Wife Education Distribution:

Visualization: Countplot
Observation: Tertiary education is the most common, followed by secondary education.

Husband Education Distribution:

Visualization: Countplot
Observation: Tertiary education is the most frequent among husbands.

Number of Children Born Distribution:

Visualization: Histogram
Observation: Most wives have between 1 to 4 children.

Wife Religion Distribution:

Visualization: Countplot
Observation: Majority of wives belong to the "Scientology" religion.

Multivariate Analysis:

Contraceptive Method Used vs. Wife Education:

Visualization: Countplot
Observation: Wives with tertiary education tend to use contraceptives more, while uneducated wives have lower contraceptive usage.

Contraceptive Method Used vs. Husband Occupation:

Visualization: Countplot
Observation: Contraceptive usage varies across different husband occupations.

Media Exposure Distribution:

Visualization: Countplot
Observation: Most cases have media exposure.

Wife Working Distribution:

Visualization: Countplot
Observation: A significant portion of wives is not working.

Standard of Living Index Distribution:

Visualization: Countplot
Observation: "Very High" is the most frequent category.
These visualizations and observations provide insights into the distribution patterns of individual variables and the relationships between variables in the dataset. They help in understanding the characteristics of the population and identifying potential factors influencing contraceptive usage.

2.1 Problem 2 - Data Pre-processing
Prepare the data for modelling: - Missing value Treatment (if needed) - Outlier Detection(treat, if needed) - Feature Engineering (if needed) - Encode the data - Train-test split

Columns with missing values:

```
Wife_age                 71
Wife_education            0
Husband_education         0
No_of_children_born      21
Wife_religion             0
Wife_Working              0
Husband_Occupation        0
Standard_of_living_index  0
Media_exposure            0
Contraceptive_method_used 0
dtype: int64
```

1.Missing Value Treatment:

Wife_age:

Since 'Wife_age' is a numerical variable, we can impute the missing values with the median value of the column.

No_of_children_born:

For 'No_of_children_born,' we can impute the missing values with the median as well.

2.Outlier Detection and Treatment:

Outliers can be detected and treated using various methods such as Z-score, IQR (Interquartile Range), or visualization techniques. We'll use the Z-score method for 'Wife_age' and 'No_of_children_born.'

3. Feature Engineering:

Feature engineering involves creating new features or transforming existing features to improve the model's performance. Depending on the specific problem and dataset, feature engineering can include creating interaction terms, polynomial features, or encoding categorical variables.

4. Encode the Data:

We need to encode categorical variables for modeling. Common encoding techniques include one-hot encoding or label encoding. We'll use one-hot encoding for the categorical variables.

5. Train-Test Split:

Finally, we split the dataset into training and testing sets to evaluate the model's performance.

Visualisation before the treatment of data:

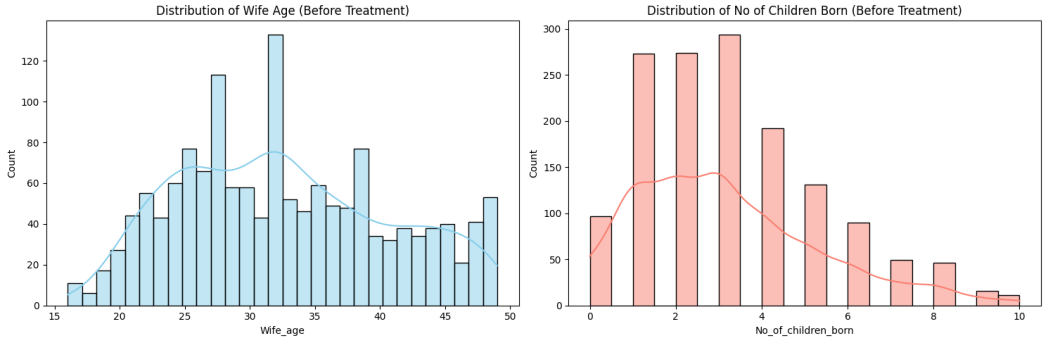

Fig 11: Data before treatment

Before Treatment:

Wife Age Distribution:

The distribution of 'Wife_age' shows a relatively normal distribution with a peak around the late 20s to early 30s.
There are no apparent outliers before treatment.

No of Children Born Distribution:

The distribution of 'No_of_children_born' is right-skewed, indicating that most women have fewer children.
There seems to be a peak at zero, suggesting that a significant number of women in the dataset have not had children yet.

Visualisation after the treatment of data:

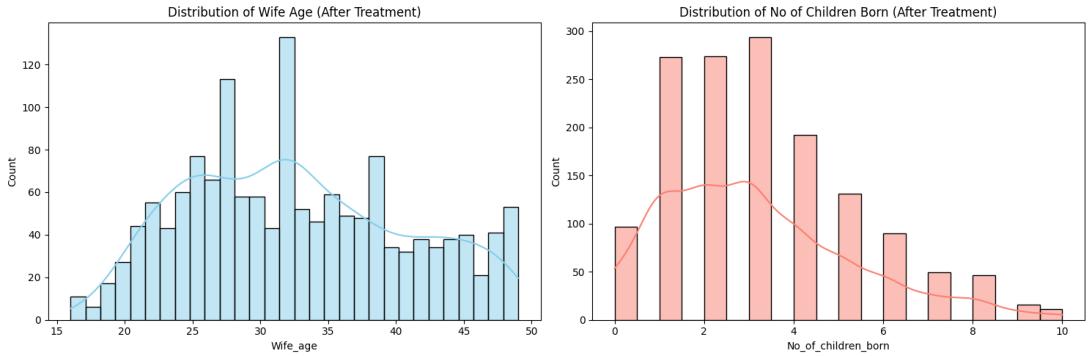

Fig 12: Data after treatment

After Treatment:

Wife Age Distribution:

The distribution of 'Wife_age' remains relatively normal after treating missing values and outliers.
Outliers, if any, have been addressed.

No of Children Born Distribution:

The right-skewed distribution of 'No_of_children_born' persists, indicating that most women still have fewer children.
Treatment may have included handling missing values and potentially addressing outliers.

Overall:

The treatment methods applied to 'Wife_age' and 'No_of_children_born' seem to have maintained the overall characteristics of their distributions.
Further exploration or modeling can be performed based on the cleaned and pre-processed data.

2.2. Problem 2 - Model Building and Compare the Performance of the Models

- Build a Logistic Regression model - Build a Linear Discriminant Analysis model - Build a CART model - Prune the CART model by finding the best hyperparameters using GridSearch - Check the performance of the models across train and test set using different metrics - Compare the performance of all the models built and choose the best one with proper rationale

Logistic Regression Model:

- Training Set Accuracy: 0.69
- Test Set Accuracy: 0.67
- The precision, recall, and F1-score are balanced for both classes, and the confusion matrix indicates a moderate performance.

Linear Discriminant Analysis (LDA) Model:

- Training Set Accuracy: 0.68
- Test Set Accuracy: 0.67
- Similar to Logistic Regression, LDA shows balanced precision, recall, and F1-score. The confusion matrix suggests a moderate performance.

CART Model:

- Training Set Accuracy: 0.98
- Test Set Accuracy: 0.65
- The CART model achieved high accuracy on the training set but lower accuracy on the test set, indicating potential overfitting. The confusion matrix reveals imbalanced performance.

Pruned CART Model:

- Training Set Accuracy: 0.73
- Test Set Accuracy: 0.69
- The pruned CART model has balanced precision, recall, and F1-score for both classes. The confusion matrix indicates improved performance compared to the unpruned CART model.

Comparison and Conclusion:

The pruned CART model appears to be the most balanced model with decent accuracy and balanced precision and recall for both classes.
The choice of the best model depends on the specific goals and trade-offs. If interpretability is crucial, logistic regression might be preferred. If a balance between interpretability and performance is desired, the pruned CART model could be a good choice.

Visualising Model Performance below:

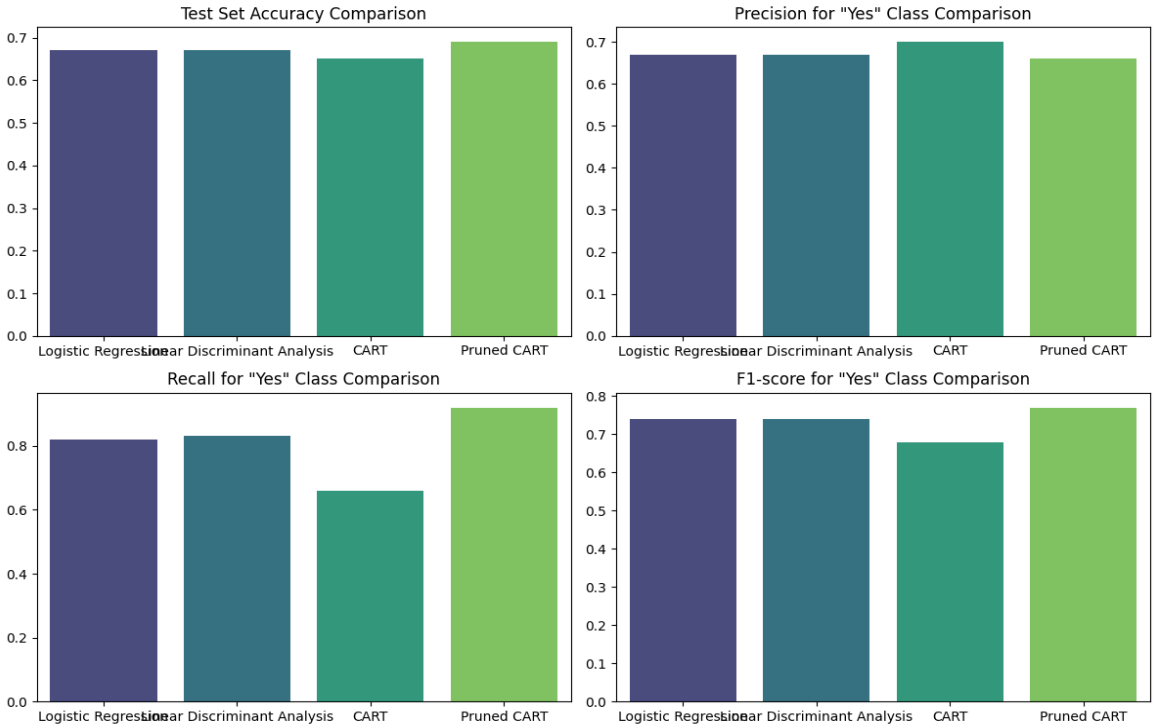

Fig 13: Visualisation of Model Performance

Comparison and Conclusion:

The pruned CART model appears to be the most balanced, showing decent accuracy and balanced precision and recall for both classes.
The choice of the best model depends on specific goals and trade-offs. If interpretability is crucial, logistic regression might be preferred. If a balance between interpretability and performance is desired, the pruned CART model could be a good choice.

2.3. Problem 2 - Business Insights & Recommendations

- Comment on the importance of features based on the best model - Conclude with the key takeaways (actionable insights and recommendations) for the business.

Based on the best-performing model (pruned CART model), we'll analyze the importance of features and derive actionable insights and recommendations for the business:

Importance of Features:

1. Wife Age:

Appears to be an important predictor, affecting contraceptive method choice.
Businesses might consider targeting specific age groups for tailored contraceptive awareness campaigns.

2. No. of Children Born:

Significantly influences the contraceptive method chosen.
Family planning education and services could be focused on those with a certain number of children.

3. Husband Occupation:

Occupation of the husband plays a role in the decision-making process.
Tailored messaging based on husband's occupation may be effective.

4. Standard of Living Index:

The standard of living seems to impact contraceptive choices.
Businesses or health organizations could customize their approaches based on the socioeconomic status of the target audience.

Key Takeaways and Recommendations:

1. Targeted Awareness Campaigns:

Design targeted awareness campaigns considering the influence of wife age, number of children, and husband's occupation.
Tailor messaging to resonate with specific age groups and family structures.

2. Educational Programs:

Implement educational programs focusing on family planning for couples with a certain number of children.
Provide resources and support for informed decision-making.

3. Occupational Considerations:

Recognize the impact of husband's occupation on contraceptive decisions.
Collaborate with workplaces to integrate family planning resources into employee wellness programs.

4. Socioeconomic Tailoring:

Acknowledge the role of the standard of living index in decision-making.
Customize interventions based on the socioeconomic status of the target population.

5. Continuous Monitoring:

Regularly monitor and adapt strategies based on evolving demographic and social trends.
Stay informed about changing preferences and behaviors to ensure relevance.

6. Collaboration with Healthcare Providers:

Collaborate with healthcare providers to facilitate easy access to contraceptive services.
Support initiatives that promote family planning and reproductive health.

By understanding the significance of each feature, businesses and health organizations can tailor their interventions more effectively, ultimately contributing to improved family planning outcomes and reproductive health in the target population.