9/3/2023

Contraceptive Method Prediction – Report

PGP-DSBA

Karthick Raj S

Table of Contents

List of Tables

List of Figures

Problem:

You are a statistician at the Republic of Indonesia Ministry of Health and you are provided with a data of 1473 females collected from a Contraceptive Prevalence Survey. The samples are married women who were either not pregnant or do not know if they were at the time of the survey.

The problem is to predict do/don't they use a contraceptive method of choice based on their demographic and socio-economic characteristics.

Dataset for Problem 2: Contraceptive_method_dataset.xlsx

Data Dictionary:

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

Classification (LOG REG, LDA, CART)

2.1 Data Ingestion: Read the dataset. Do the descriptive statistics and do null value condition check, check for duplicates and outliers and write an inference on it. Perform Univariate and Bivariate Analysis and Multivariate Analysis.

The dataset that is given for analysis is ontraceptive Prevalence Survey.

Shape:

```
The shape of the dataset is (1473, 10)
```

There are 1473 Rows and 10 columns in the dataset.

First Five (Head):

The First Five rows of the dataset (The rows and columns has been transposed for easier view). Refer jupyter workings for the output.

	0	1	2	3	4
Wife_age	24	45	43	42	36
Wife_ education	Primary	Uneducated	Primary	Secondary	Secondary
Husband_education	Secondary	Secondary	Secondary	Primary	Secondary
No_of_children_born	3	10	7	9	8
Wife_religion	Scientology	Scientology	Scientology	Scientology	Scientology
Wife_Working	No	No	No	No	No
Husband_Occupation	2	3	3	3	3
Standard_of_living_index	High	Very High	Very High	High	Low
Media_exposure	Exposed	Exposed	Exposed	Exposed	Exposed
Contraceptive_method_used	No	No	No	No	No

Table 1 First five rows for dataset

Info:

The info of the dataset is

```
RangeIndex: 1473 entries, 0 to 1472
Data columns (total 10 columns):
 #   Column                      Non-Null Count  Dtype
---  ------                      --------------  -----
 0   Wife_age                    1402 non-null   float64
 1   Wife_ education             1473 non-null   object
 2   Husband_education           1473 non-null   object
 3   No_of_children_born         1452 non-null   float64
 4   Wife_religion               1473 non-null   object
 5   Wife_Working                1473 non-null   object
 6   Husband_Occupation          1473 non-null   int64
 7   Standard_of_living_index    1473 non-null   object
 8   Media_exposure              1473 non-null   object
 9   Contraceptive_method_used   1473 non-null   object
dtypes: float64(2), int64(1), object(7)
memory usage: 115.2+ KB
```

There are 10 variables out of which 3 are numerical. There are null values in 'wife_age' and 'No_of_children_born'

Five Point Summary:

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Wife_age	1402.0	NaN	NaN	NaN	32.606277	8.274927	16.0	26.0	32.0	39.0	49.0
Wife_ education	1473	4	Tertiary	577	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Husband_education	1473	4	Tertiary	899	NaN	NaN	NaN	NaN	NaN	NaN	NaN
No_of_children_born	1452.0	NaN	NaN	NaN	3.254132	2.365212	0.0	1.0	3.0	4.0	16.0
Wife_religion	1473	2	Scientology	1253	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Wife_Working	1473	2	No	1104	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Husband_Occupation	1473.0	NaN	NaN	NaN	2.137814	0.864857	1.0	1.0	2.0	3.0	4.0
Standard_of_living_index	1473	4	Very High	684	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Media_exposure	1473	2	Exposed	1364	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Contraceptive_method_used	1473	2	Yes	844	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Table 2 Five point summary for dataset

The Wife age is between the range of 16 to 49 with 33 as mean.

Half of the women in the dataset has below 3 childrens.

Most of the women has tertiary education with very high standard of living, scientology and exposed to media.

EDA:

Univariate:

```
WIFE_ EDUCATION
Tertiary      577
Secondary     410
Primary       334
Uneducated    152
Name: Wife_ education, dtype: int64
**********************************************************************
HUSBAND_EDUCATION
Tertiary      899
Secondary     352
Primary       178
Uneducated     44
Name: Husband_education, dtype: int64
**********************************************************************
WIFE_RELIGION
Scientology        1253
Non-Scientology     220
Name: Wife_religion, dtype: int64
**********************************************************************
WIFE_WORKING
No     1104
Yes     369
Name: Wife_Working, dtype: int64
**********************************************************************
STANDARD_OF_LIVING_INDEX
Very High    684
High         431
Low          229
Very Low     129
Name: Standard_of_living_index, dtype: int64
**********************************************************************
MEDIA_EXPOSURE
Exposed        1364
Not-Exposed     109
Name: Media_exposure , dtype: int64
**********************************************************************
CONTRACEPTIVE_METHOD_USED
Yes    844
No     629
Name: Contraceptive_method_used, dtype: int64
**********************************************************************
```

Boxplot

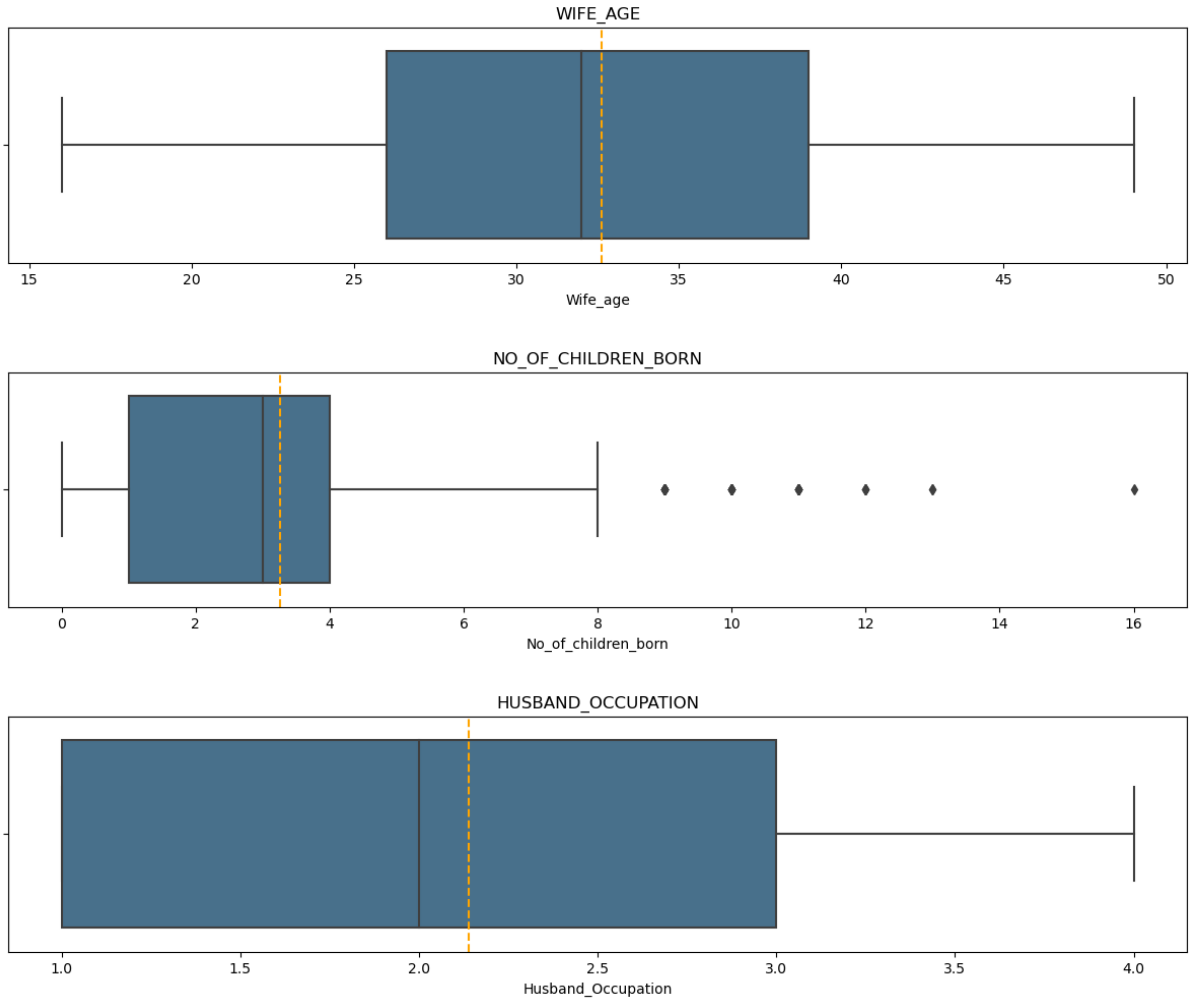
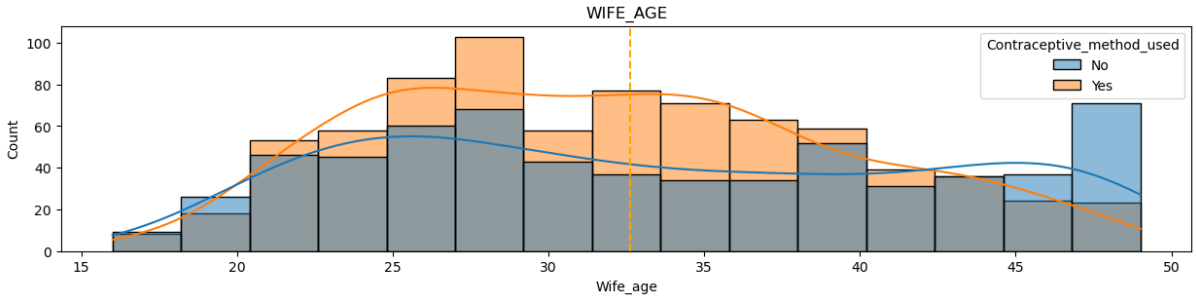


Figure A Boxplot

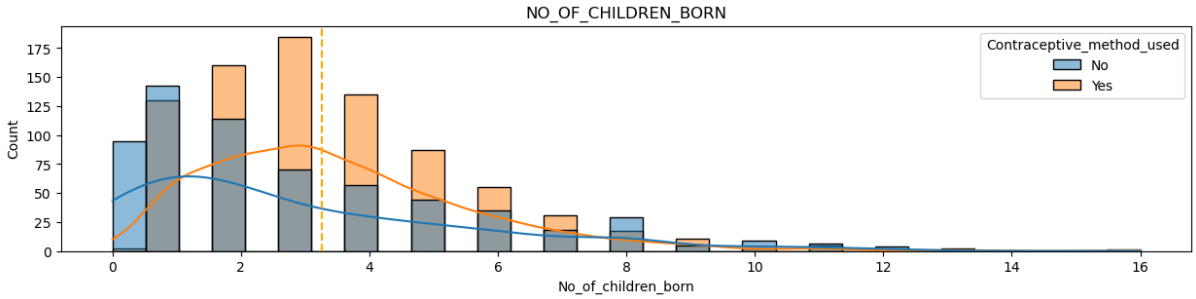
The outlier in the number of children can be acceptable as problem statement is about Contraceptive method used. The number of children can be an efficient variable without treating outliers.

Bivariate:

Histogram



The women not using contraceptive methods are high in older age women.



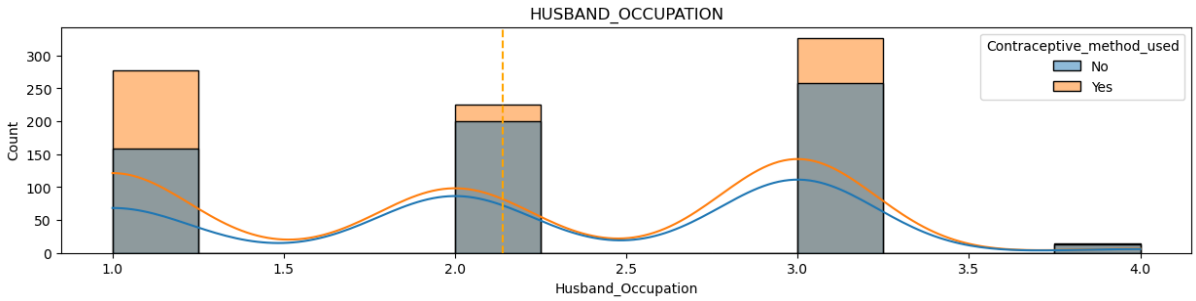


Figure B Histogram

There is any clear difference with women using and not using contraceptive method.

`Skewness in Wife_age = 0.27561045371388626`

`Skewness in No_of_children_born = 1.1000491486638886`

`Skewness in Husband_Occupation = -0.15477540428760683`

The Skewness shows the outliers direction. There is any outlier in wife age and husband occupation as their skewness are near to zero.

Correlation Heatmap

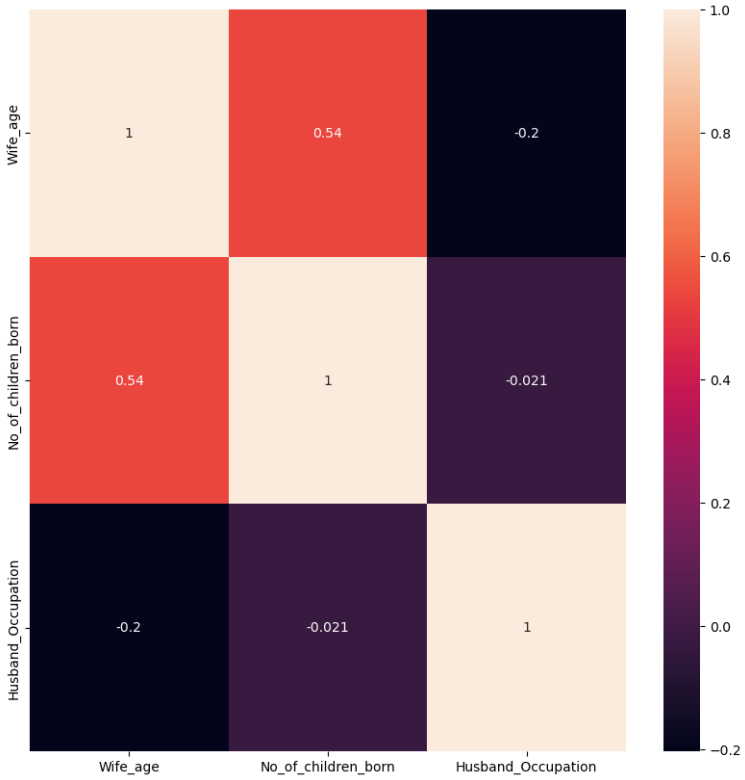


Figure C Correlation Heatmap

Wife age and no of children born has high correlation. It makes sense as children increases with increase in age.

Multivariate:

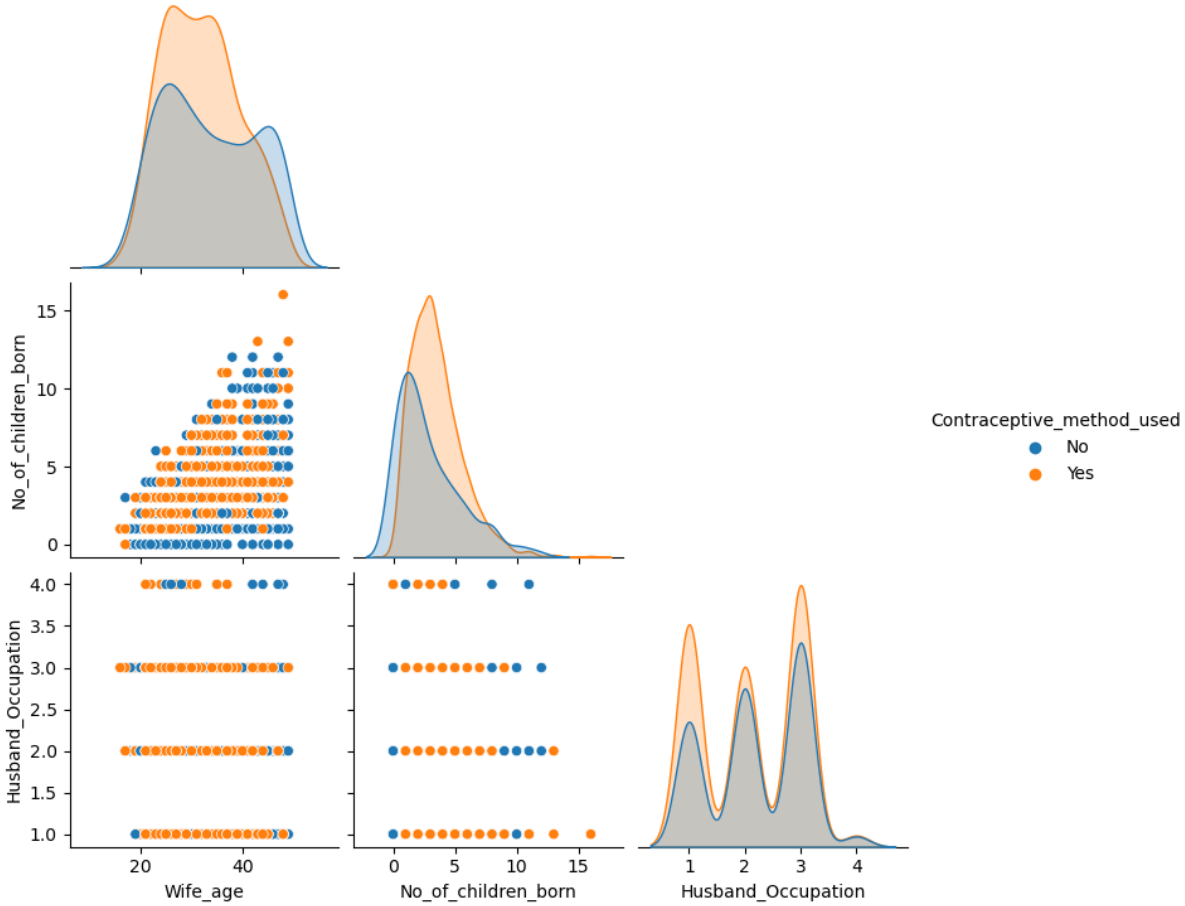


Figure D Pairplot

Among the three, wife age and no of children born can be an important variable in segregating the women using/not using contraceptive method.

Null values and Duplicates:

```
Wife_age                  71
No_of_children_born       21
```

There are 71 and 21 null values in wife age and no of children being born.

There are 80 duplicates in the dataset.

As we don't have any id column, the duplicates can be of two different individuals. In this case, duplicates are removed as they are very few.

The shape after treating duplicates (1393, 10)

The Null values in wife age and no of children born are imputed with median.

2.2 Do not scale the data. Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and LDA (linear discriminant analysis) and CART.

The Data is encoded with labels.

The Dataset after labelling is (Refer jupyter notebook for more clear view)

	Wife_age	Wife_education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_e
0	24.0	1	2	3.0	0	0	2	2	
1	45.0	0	2	10.0	0	0	3	3	
2	43.0	1	2	7.0	0	0	3	3	
3	42.0	2	1	9.0	0	0	3	2	
4	36.0	2	2	8.0	0	0	3	1	

Table 3 Dataset after labelling

The Percentage of 1s and 0s in target variable is

```
%0s 0.5592246949030869
%1s 0.44077530509691315
```

The dataset is split into 70% Train and 30% Test data.

The training data is fitted into the model for linear regression. The model 1 has all the variables.

The Logistic Regression, LDA and CART are fitted with the training data and hyper tuning has been done to get the best optimised results in all these three. (Refer jupyter notebook for workings)

Optimised logistic model:

LogisticRegression(max_iter=10000, n_jobs=-1, solver='newton-cg')

CART:

The classification report before pruning for training and test data shows overfitting issue.

	precision	recall	f1-score	support
0	0.97	1.00	0.99	553
1	1.00	0.96	0.98	422
accuracy			0.98	975
macro avg	0.99	0.98	0.98	975
weighted avg	0.98	0.98	0.98	975

Table 4 Classification report for Training model(before optimising)

	precision	recall	f1-score	support
0	0.63	0.68	0.65	226
1	0.58	0.53	0.56	192
accuracy			0.61	418
macro avg	0.61	0.60	0.60	418
weighted avg	0.61	0.61	0.61	418

Table 5 Classification report for Testing model(before optimising)

The Training data has high accuracy, recall and precision due to overfitting.

Optimised Decision Tree:
DecisionTreeClassifier(max_depth=75,min_impurity_decrease=0.0051,min_samples_leaf=8, min_samples_split=3)

In the decision tree, the train dataset has been overfitted with the model. After optimising with hyper tuning, it has been pruned to solve the overfitting problem.

Final Decision Tree:

The below decision tree has been pruned after optimising the decision tree.

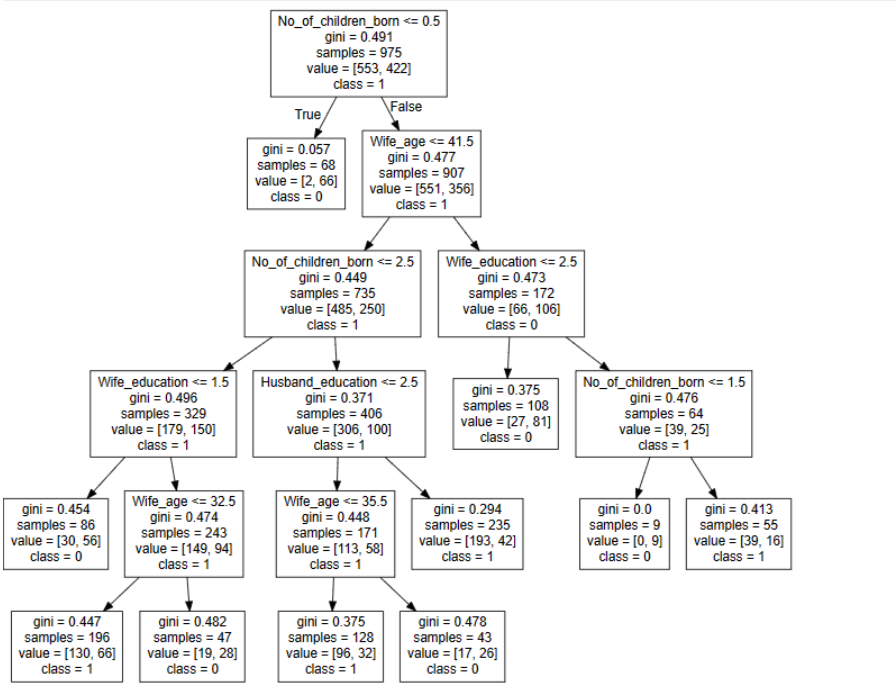


Figure E Decision Tree

From the decision tree, the important feature for classification are the four variables which is shown in the below feature importance plot.

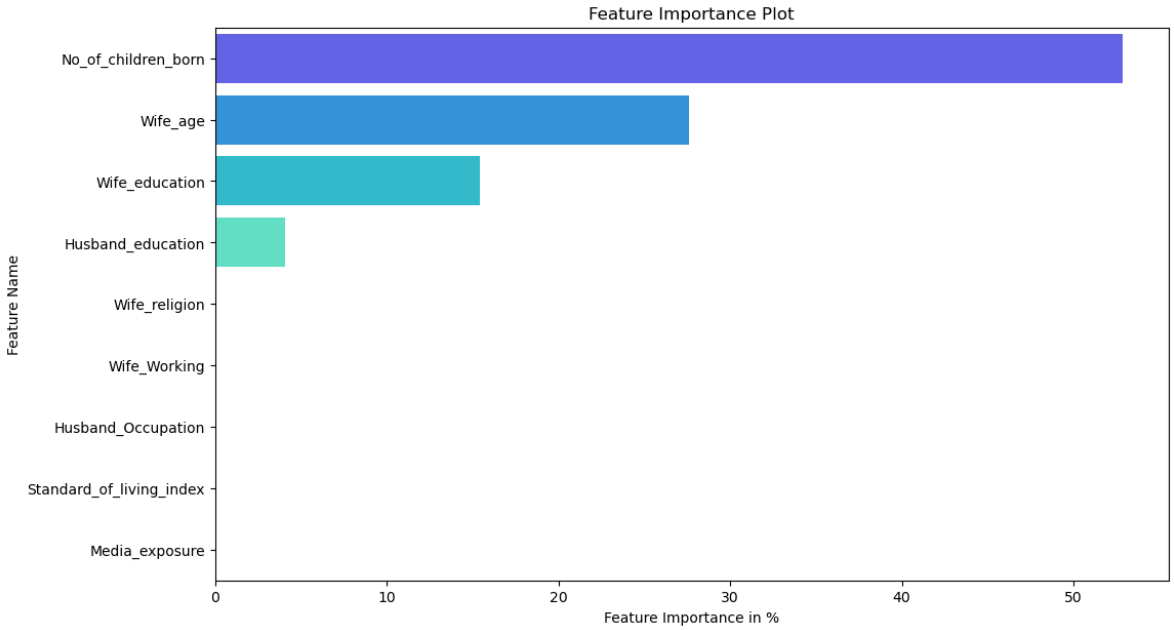


Figure F Feature importance Plot

The Chart shows the important feature for classification.

	Imp
No_of_children_born	**0.528979**
Wife_age	**0.275917**
Wife_education	**0.154272**
Husband_education	**0.040832**
Wife_religion	0.000000
Wife_Working	0.000000
Husband_Occupation	0.000000
Standard_of_living_index	0.000000
Media_exposure	0.000000

The Final Model can be obtained from these three model by comparing the classification report, confusion matrix and ROC curve.

2.3 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.

The classification that is of interest is women who are not using Contraceptive method.

The AUC, ROC curve, Confusion matrix and Classification report for the final model of Log Reg, LDA and CART has been compared below.

Logistic Regression:

Train Performance vs Test Performance:

	TRAIN	TEST
AUC:	0.7252041856997162	0.654901825221239

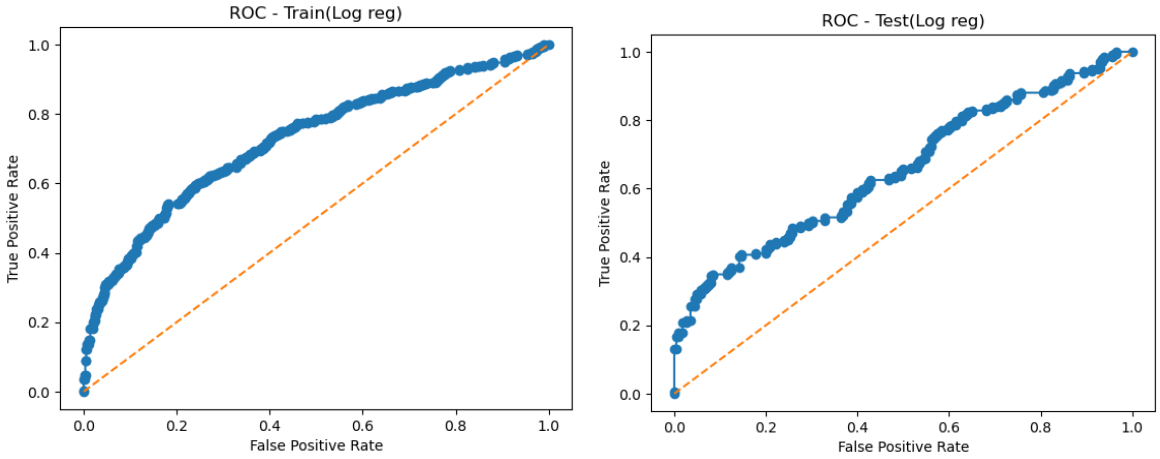


Figure G ROC Curve for Train and Test (Log Reg)

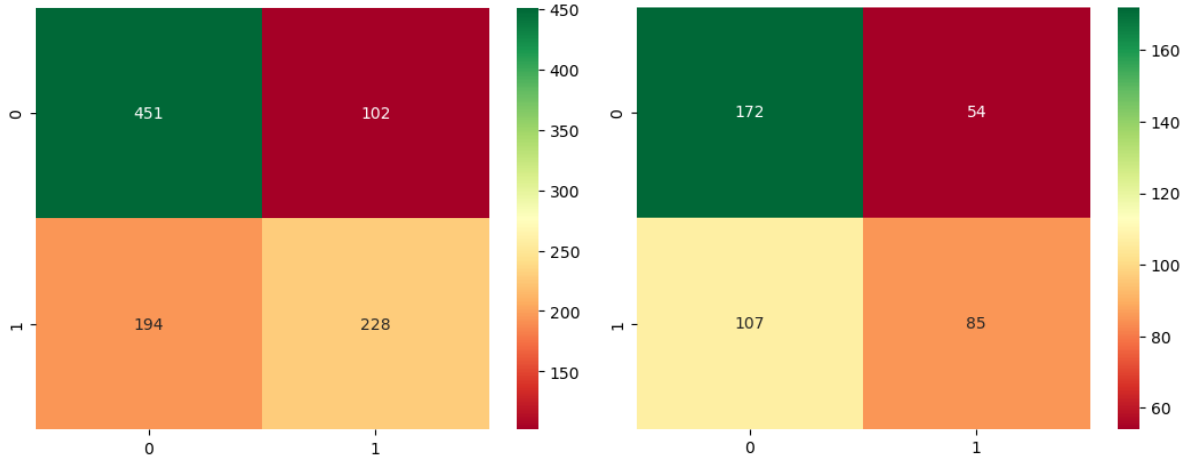


Figure H Confusion Matrix for train and Test (Log Reg)

	precision	recall	f1-score	support
0	0.70	0.82	0.75	553
1	0.69	0.54	0.61	422
accuracy			0.70	975
macro avg	0.70	0.68	0.68	975
weighted avg	0.70	0.70	0.69	975

	precision	recall	f1-score	support
0	0.62	0.76	0.68	226
1	0.61	0.44	0.51	192
accuracy			0.61	418
macro avg	0.61	0.60	0.60	418
weighted avg	0.61	0.61	0.60	418

Table 6 Classification Report for Train and Test Data (Log Reg)

The logistic regression performs well in train but for the testing data is not enough, the accuracy and f1-score for label 1 drops by 10%.

ROC curve and AUC confirms the same that Log Reg is not performing well with test data.

LDA:

Train Performance vs Test Performance:

	TRAIN	TEST
AUC:	0.7236358338404052	0.6563997971976401

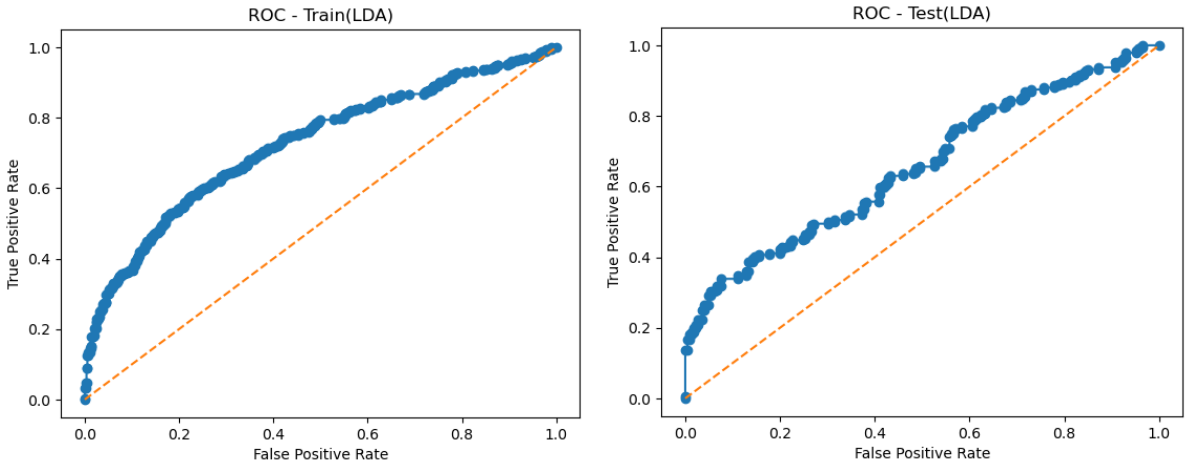


Figure I ROC Curve for Train and Test (LDA)

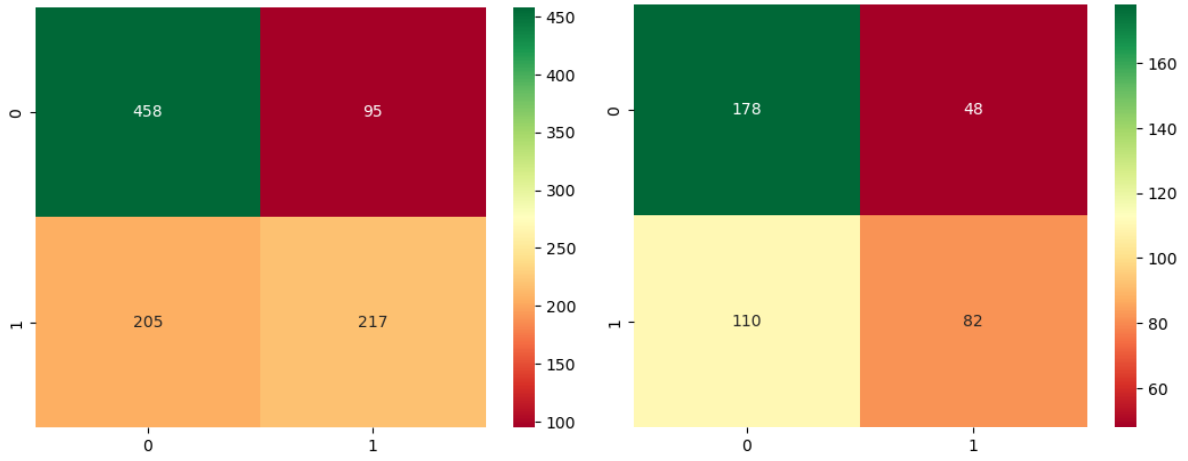


Figure J Confusion matrix for Train and Test (LDA)

	precision	recall	f1-score	support
0	0.69	0.83	0.75	553
1	0.70	0.51	0.59	422
accuracy			0.69	975
macro avg	0.69	0.67	0.67	975
weighted avg	0.69	0.69	0.68	975

	precision	recall	f1-score	support
0	0.62	0.79	0.69	226
1	0.63	0.43	0.51	192
accuracy			0.62	418
macro avg	0.62	0.61	0.60	418
weighted avg	0.62	0.62	0.61	418

Table 7 Classification report for Train and Test (LDA)

The LDA performs well in train but for the testing data is not enough, the accuracy and f1-score for label 1 drops here also like logistic regression.

ROC curve and AUC confirms the same that LDA is not performing well with test data.

CART:

Train Performance vs Test Performance:

	TRAIN	TEST
AUC:	0.7790445051978437	0.7196833517699115

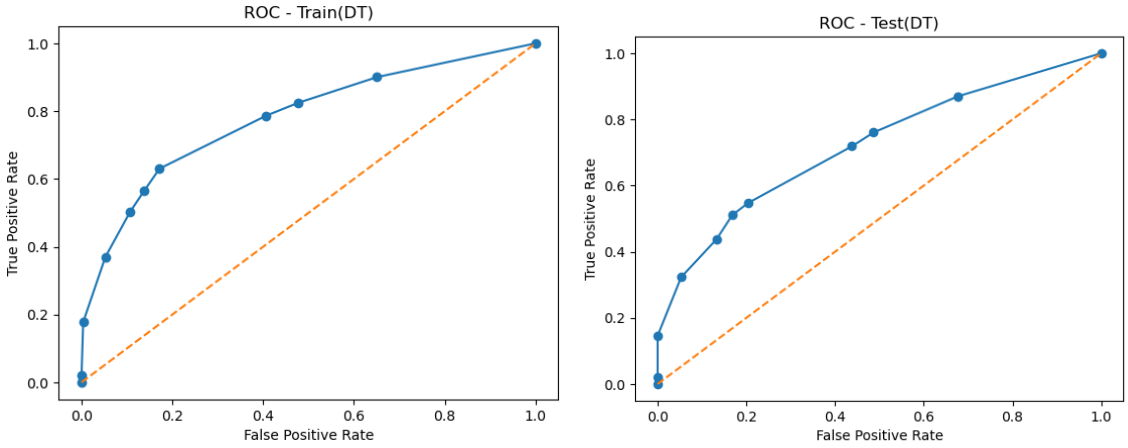


Figure K ROC Curve for Train and Test (CART)

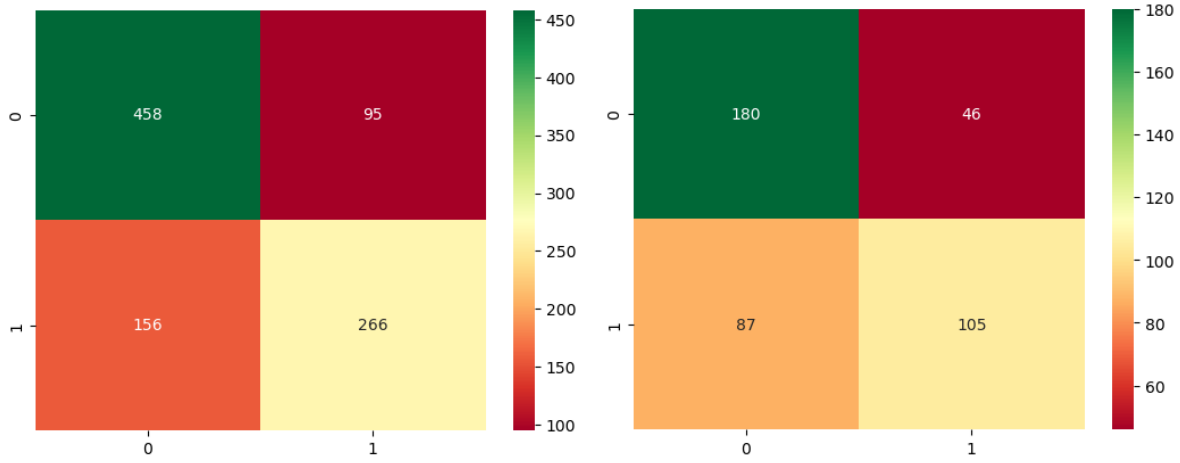


Figure L Confusion matrix for Train and Test (CART)

	precision	recall	f1-score	support
0	0.75	0.83	0.78	553
1	0.74	0.63	0.68	422
accuracy			0.74	975
macro avg	0.74	0.73	0.73	975
weighted avg	0.74	0.74	0.74	975

	precision	recall	f1-score	support
0	0.67	0.80	0.73	226
1	0.70	0.55	0.61	192
accuracy			0.68	418
macro avg	0.68	0.67	0.67	418
weighted avg	0.68	0.68	0.68	418

Table 8 Classification report for Train and Test (CART)

Among the three, CART has less drop in train and test f1-score and accuracy.

AUC Score is more than 0.70 for both train and test only for CART.

The Recall for other models is poor when compared to CART. The error of predicting a women not using contraceptive method as using contraceptive method is more important than the inverse case. CART is the best model for prediction.

2.4 Inference: Basis on these predictions, what are the insights and recommendations.

Step 1: Performing Descriptive Statistics and EDA.

From this, the insights from EDA are

- The Wife age is between the range of 16 to 49 with 33 as mean.
- Most of the women has tertiary education with very high standard of living, scientology and exposed to media.
- As the age increases, number of children also increases.
- The women not using contraceptive method is more above the age of 45 compared to women using contraceptive method.
- The mean number of children born is 3.
- The women using/not using contraceptive method are overlapped and there is no clear visual separation between two categories.

Step 2: Null values and Duplicates treatment.

- The Null values are imputed with median.
- There are duplicates for this dataset which are removed.

Step 3: Encoding the dataset**.**

- The Categorical variables are labelled.

Step 4: Data Split 70:30

- The Data has been split into Train and Test.
- The Training Data is used for training the model and the test is used for validation.

Step 5: Building a Model – Logistic Regression.

- The Train data is fitted into the model and the hyper tuning is done for optimised model.
- The Log Reg performs well in train but for the testing data is not enough, the accuracy and f1-score for label 1 drops.

Step 6: Building a Model - LDA

- The LDA performs well in train but for the testing data is not enough, the accuracy and f1-score for label 1 drops.

Step 7: Building a Model – CART.

- The Train data is fitted into the model and the hyper tuning is done for optimised model.
- The Decision tree is pruned for solving overfitting.
- AUC Score is more than 0.70 for both train and test only for CART.

- Among the three, CART is the best model for prediction.

Step 8: Business Insights**.**

- The CART Model has four important factors as
  - ✓ No_of_children_born
  - ✓ Wife_age
  - ✓ Wife_education
  - ✓ Husband_education
- The prediction of women not using contraceptive method as using contraceptive method is more important than the inverse case.
- If the prediction of the women using contraceptive method as not using can be ignored but it also should not be high as the cost for resources in creating awareness will go to waste.
- The Main insight for the model is to predict who are not using so the awareness can be focused on those segments.