PREDICTIVE MODELLING

Linear Regression/Logistic Regression/LDA/CART

NOVEMBER 05, 2023

SHUBHAM KUMAR
Batch: G2- SAT 3PM

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.

Solution: -

1. **First 5 rows of the dataset**

	0	1	2	3	4
Wife_age	24.0	45.0	43.0	42.0	36.0
Wife_ education	Primary	Uneducated	Primary	Secondary	Secondary
Husband_education	Secondary	Secondary	Secondary	Primary	Secondary
No_of_children_born	3.0	10.0	7.0	9.0	8.0
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 2.1A
First 5 rows of the dataset

2. **Last 5 rows of the dataset**

	1468	1469	1470	1471	1472
Wife_age	33.0	33.0	39.0	33.0	17.0
Wife_ education	Tertiary	Tertiary	Secondary	Secondary	Secondary
Husband_education	Tertiary	Tertiary	Secondary	Secondary	Secondary
No_of_children_born	NaN	NaN	NaN	NaN	1.0
Wife_religion	Scientology	Scientology	Scientology	Scientology	Scientology
Wife_Working	Yes	No	Yes	Yes	No
Husband_Occupation	2	1	1	2	2
Standard_of_living_index	Very High	Very High	Very High	Low	Very High
Media_exposure	Exposed	Exposed	Exposed	Exposed	Exposed
Contraceptive_method_used	Yes	Yes	Yes	Yes	Yes

Table 2.1B
Last 5 rows of the dataset

3. **Shape of the dataset**

```
Number of rows in a dataset = 1473
Number of columns in a dataset = 10
```

Fig. 2.1A
Number of Rows and Column in a Dataset

4. **Data Types of the variables in a Dataset**

```
<class 'pandas.core.frame.DataFrame'>
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
```

Table 2.1C
Datatypes of Columns

Inferences from Table 2.1C

- Number of columns having float64 datatypes = 2 (Wife_age, No_of_children_born)
- Number of Columns having int64 datatypes = 1 (Husband_Occupation)
- Number Columns having float64 datatypes = 7 (Wife_education, Husband_eductaion, Wife_religion, Wife_working, Stanadard_of_living_Index, Media_exposure, Contraceptive_method_used
- Total Number of Numeric columns = 3
- Total Number of Categorical columns = 7
- Also Wife_age, No_of_children_born has less number of non-null rows which indicates about Null entries in dataset.

5. Descriptive Statistics/Summary of the Dataset

	count	mean	std	min	25%	50%	75%	max
Wife_age	1402.0	32.606277	8.274927	16.0	26.0	32.0	39.0	49.0
No_of_children_born	1452.0	3.254132	2.365212	0.0	1.0	3.0	4.0	16.0
Husband_Occupation	1473.0	2.137814	0.864857	1.0	1.0	2.0	3.0	4.0

Table 2.1D

Descriptive Statistics/Summary of the Dataset

Inferences from Table 2.1D

- Minimum Wife's Age = 16 and Maximum wife's age = 49
- Number of children born has minimum of 0 and maximum of 16 whereas 75% of data points are below 4 which clearly indicates that there are outliers
- Husband Occupation are 1, 2, 3 and 4 where each number represent occupation in random numbers.

6. Let's check categories of the Categorical Variables/Columns

```
WIFE_ EDUCATION 4
Tertiary       577
Secondary      410
Primary        334
Uneducated     152
Name: Wife_ education, dtype: int64

HUSBAND_EDUCATION 4
Tertiary       899
Secondary      352
Primary        178
Uneducated      44
Name: Husband_education, dtype: int64

WIFE_RELIGION 2
Scientology         1253
Non-Scientology      220
Name: Wife_religion, dtype: int64

WIFE_WORKING 2
No      1104
Yes      369
Name: Wife_Working, dtype: int64

STANDARD_OF_LIVING_INDEX 4
Very High     684
High          431
Low           229
Very Low      129
Name: Standard_of_living_index, dtype: int64

MEDIA_EXPOSURE  2
Exposed         1364
Not-Exposed      109
Name: Media_exposure , dtype: int64

CONTRACEPTIVE_METHOD_USED 2
Yes     844
No      629
Name: Contraceptive method used, dtype: int64
```

Fig. 2.1B

Number of Categories for Categorical Variables/Columns

Inferences from Fig 2.1B: -

- Wife Education has 4 ordinal categories namely: - Tertiary, Secondary, Primary, Uneducated.
- Husband Education has 4 ordinal categories namely: - Tertiary, Secondary, Primary, Uneducated.
- Wife Religion has 2 binary categories namely: - Scientology and Non-Scientology
- Wife Working has 2 binary categories namely: - Yes and no
- Living of Standard Index has 4 ordinal categories: - Very High, High, Low, Very Low
- Media Exposure has 2 binary categories: - Exposed and Not-Exposed
- Contraceptive Method Used has 2 binary categories: - Yes and no

7. Check for Duplicates

```
Total Number of Dupliacted rows in a dataset =  80
```

	Wife_age	Wife_ education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_exposu
79	38.0	Tertiary	Tertiary	1.0	Scientology	Yes	1	Very High	Expos
167	26.0	Tertiary	Tertiary	1.0	Scientology	No	1	Very High	Expos
224	47.0	Tertiary	Tertiary	4.0	Scientology	No	1	Very High	Expos
270	30.0	Tertiary	Tertiary	2.0	Scientology	No	1	Very High	Expos
299	26.0	Tertiary	Tertiary	1.0	Scientology	No	1	Very High	Expos
...	...	...	...	...	...	...	...	...	
1367	44.0	Tertiary	Tertiary	5.0	Scientology	Yes	1	Very High	Expos
1387	NaN	Secondary	Tertiary	2.0	Scientology	Yes	2	Very High	Expos
1423	NaN	Tertiary	Tertiary	2.0	Non-Scientology	No	1	Very High	Expos
1440	NaN	Tertiary	Tertiary	1.0	Non-Scientology	Yes	2	Very High	Expos
1447	NaN	Tertiary	Tertiary	2.0	Non-Scientology	Yes	2	Very High	Expos

80 rows × 10 columns

Table 2.1E

Duplicated data's

- We need to drop these duplicated row

```
Number of Duplicated rows in a dataset =  0
```

Wife_age	Wife_ education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_exposure

8. Check for Outliers

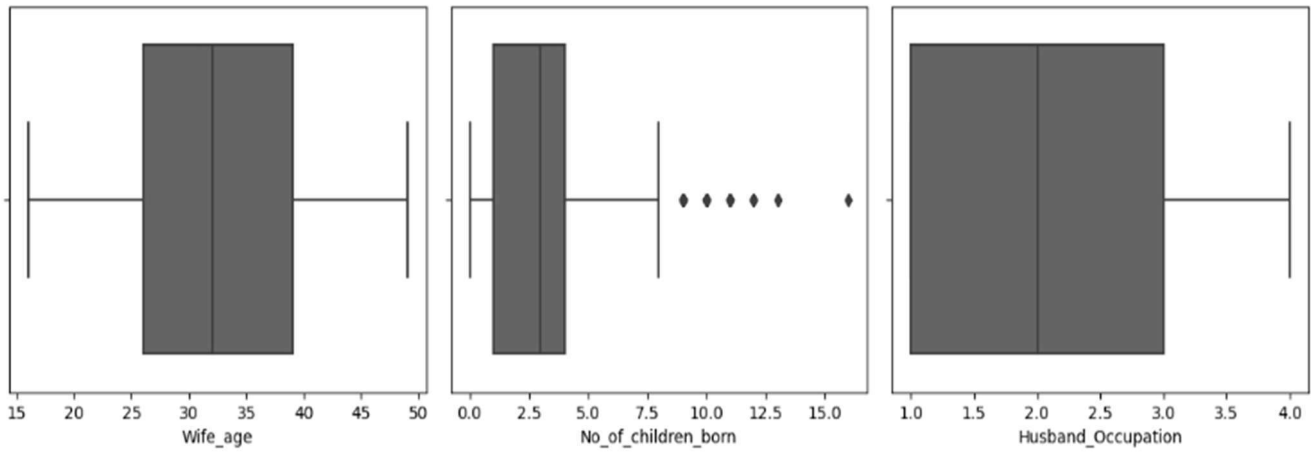


Fig. 2.1C
Boxplot of Numeric Variables

Inferences from Fig. 2.1C: -

- No_of_children_born has outliers present in it which was also evident from descriptive summary of the dataset as max value = 16 whereas 75% of data points where less than or equal 5. We need to treat this outlier
- We had treated the Outlier using IQR method as the dataset is right skewed.

- Wife_age and Husband_Occupation doesn't have any outliers present in it.

9. Check for Null values in a dataset.

Wife_age	67
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	

Table 2.1F
Dataset before treating Null Entries/Values

Wife_age	0
Wife_ education	0
Husband_education	0
No_of_children_born	0
Wife_religion	0
Wife_Working	0
Husband_Occupation	0
Standard_of_living_index	0
Media_exposure	0
Contraceptive_method_used	0
dtype: int64	

Table 2.1G
Dataset after treating Null Entries/Values

Inference from Table 2.1F and Table 2.1G: -

- **We have null values in Wife_age and No_of_children_born**
- **Since these are numeric columns, we have used mean (round off) for imputing these null values in both the columns**

 # Univariate Analysis of Numeric Variables: -

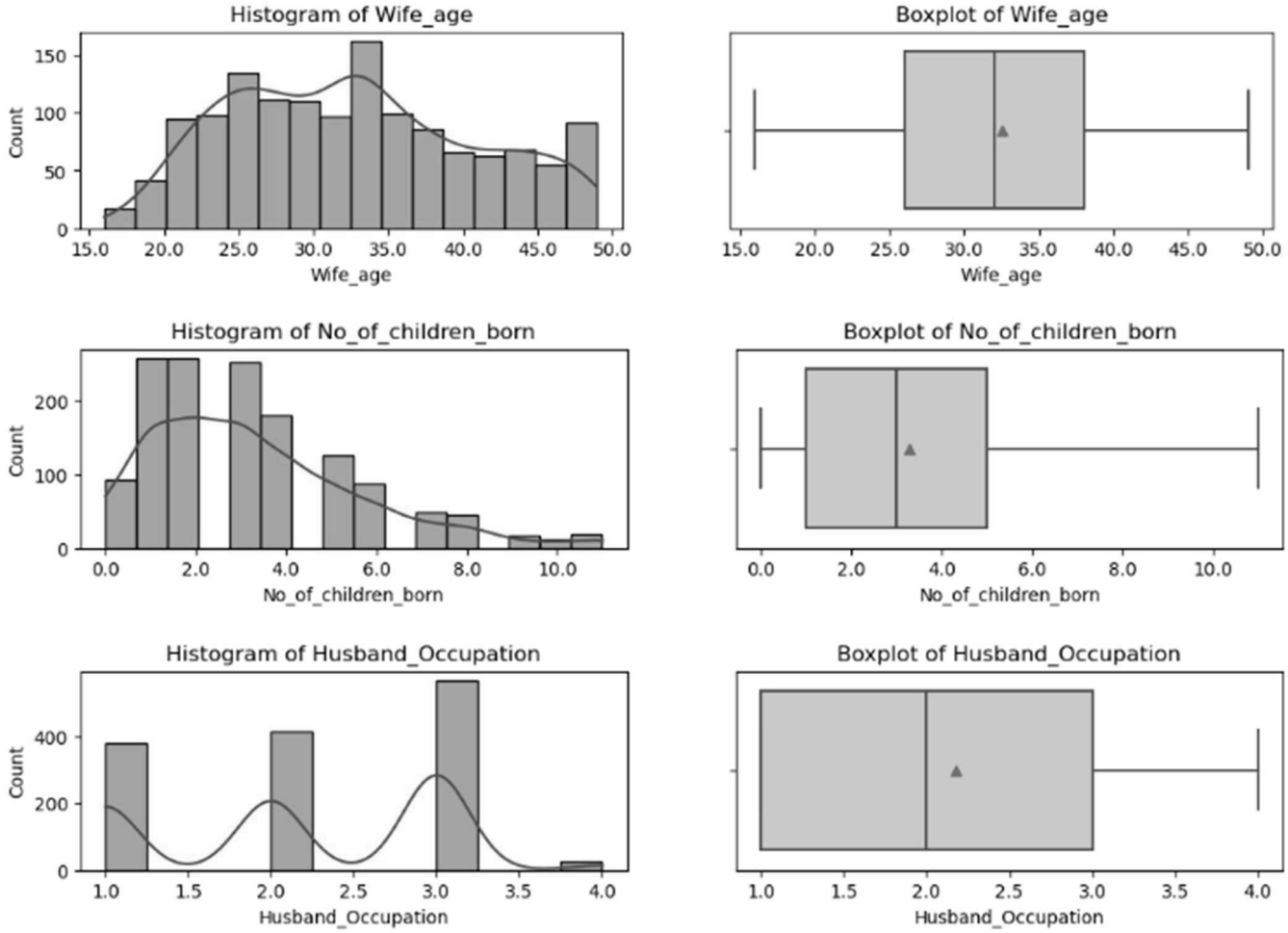


Fig. 2.1D

Boxplot and Histogram of Numeric Variables

Inference from Fig. 2.1D: -

- Wife_age and No_of_children_born are right skewed
- Maximum number of wife age are between 25 to 35
- Maximum number of children born are either 1 or 2
- Most number of Husband's occupation of type 3
- Also Husband's occupation is of categorical types

 # Univariate Analysis of Categorical Variables

`

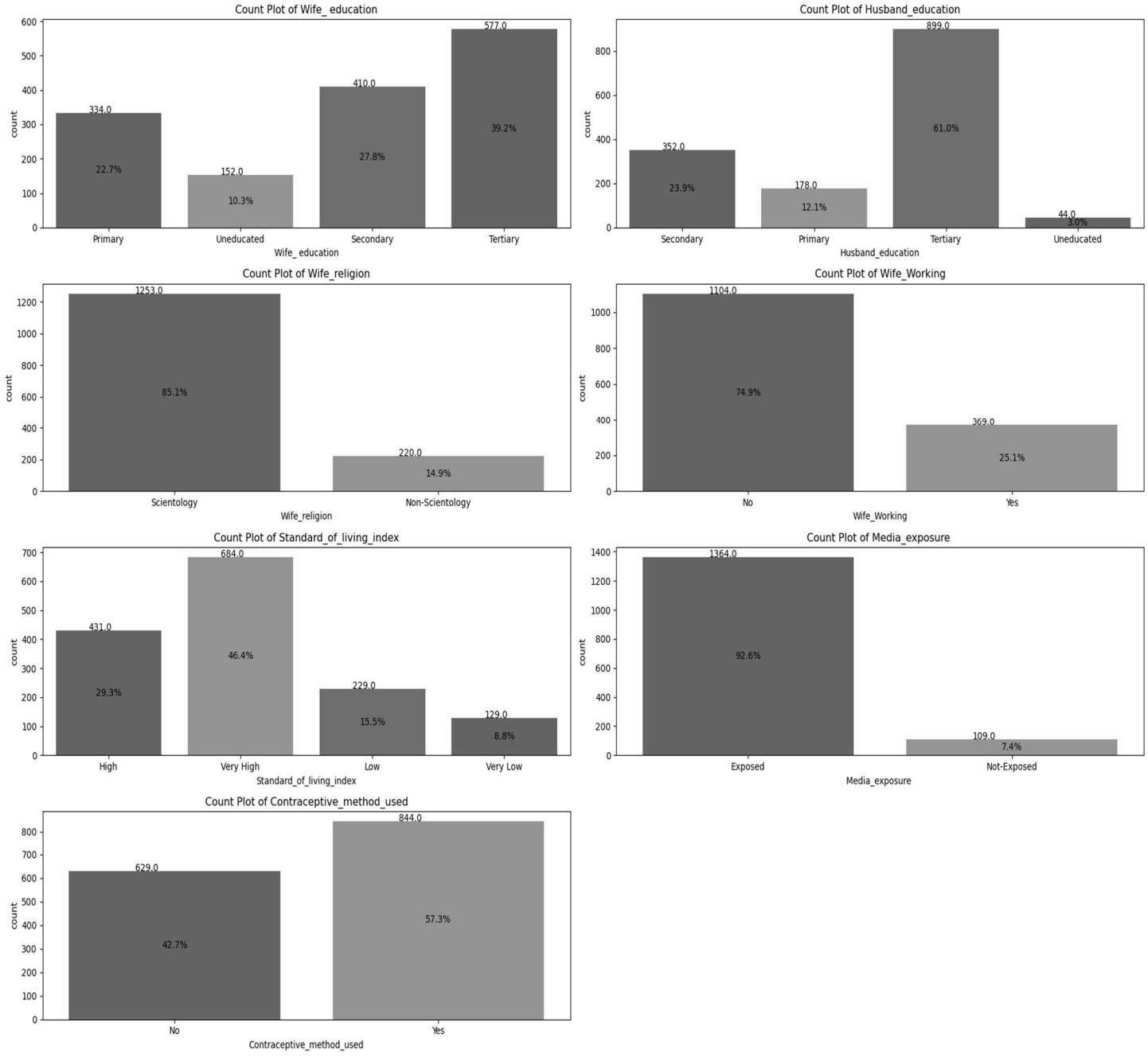


Fig. 2.1E

Count Plot of Categorical Variables

Inferences from Fig. 2.1E

- In Wife_education, most of the wife are educated upto Tertiary i.e., 577 or 39.2 % and very few are uneducated i.e., 152 or 10.3%
- In Husband_education, most of the wife are educated up to Tertiary i.e., 899 or 61 % and very few are uneducated i.e., 44 or 3%

- In Standard_of_living_Index, most family are having very high standard of living i.e., 684 or 46.4 % and very few are having very low i.e., 129 or 8.8%
- In Media_Exposure, most family are exposed to media i.e., 1364 or 92.6%
- In Contraceptive_method_used, most wife have used contraceptive method i.e., 844 or 57.3%

 # Let's perform Bi-Variate Analysis

Comparing each Independent Variables with Target/Predictor Variables

1. **Wife_age and Contraceptive_method_used**

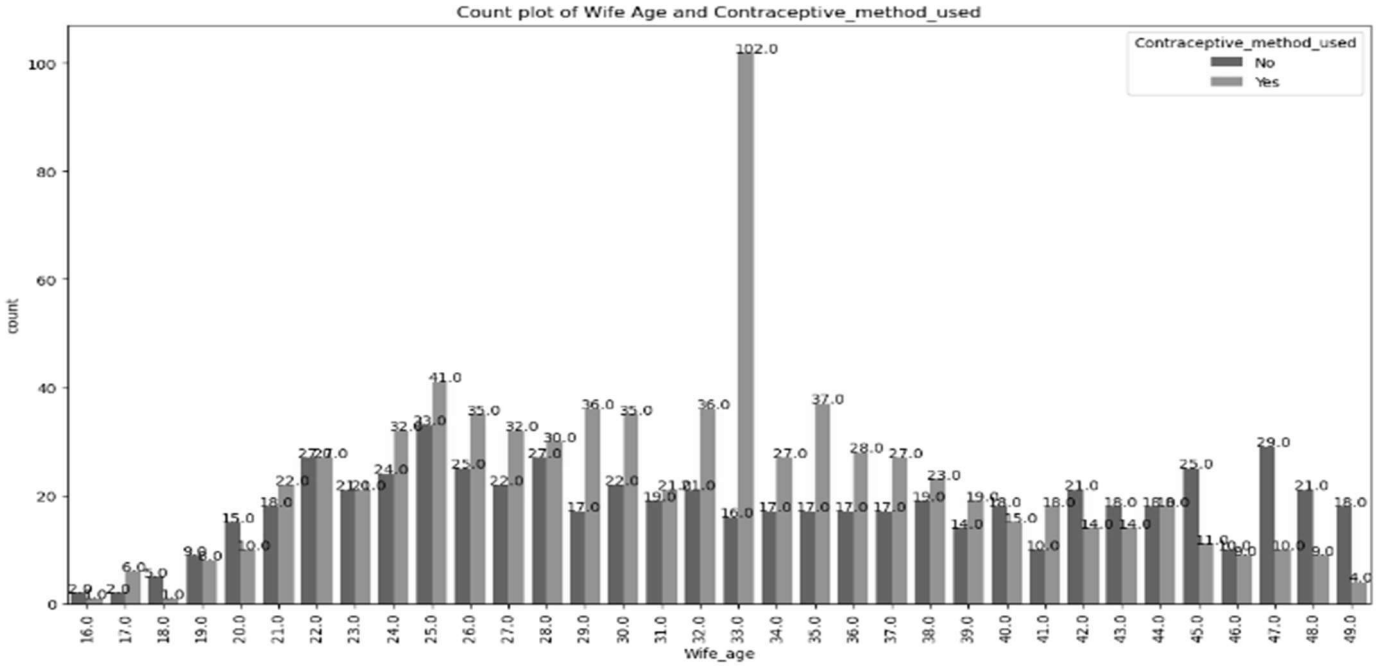


Fig. 2.1F

Count plot between Wife_age and Contraceptive_method_used

Inferences: -

- Maximum percentage of women who used Contraceptive_method_used are of Age 33
- Maximum number of woman who doesn't used Contraceptive_method_ are of age 47

2. **Wife_education and Contraceptive_method_used**

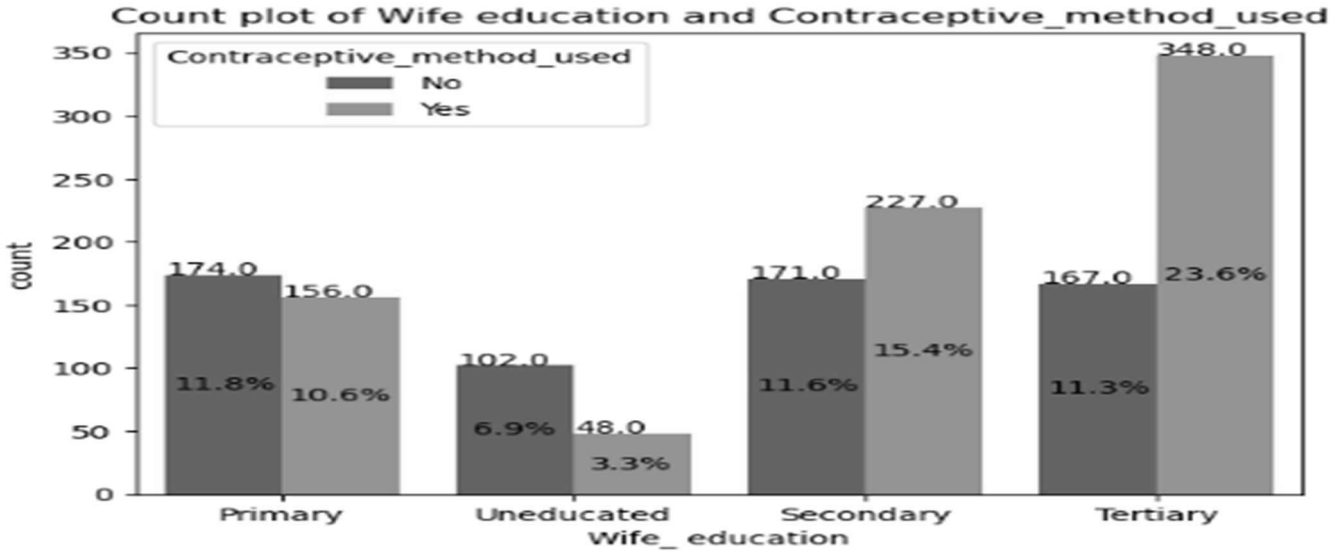


Fig. 2.1G

Count plot between Wife Education and Contraceptive_method_used

Inferences from above Fig. 2.1G

- Most number of wife who used contraceptive method have education level of Tertiary i.e., 348 or 23.6%
- Most number of wife who doesn't used contraceptive method have education level of Primary i.e., 174 or 11.8%

3. **Husband_education and Contraceptive_method_used**

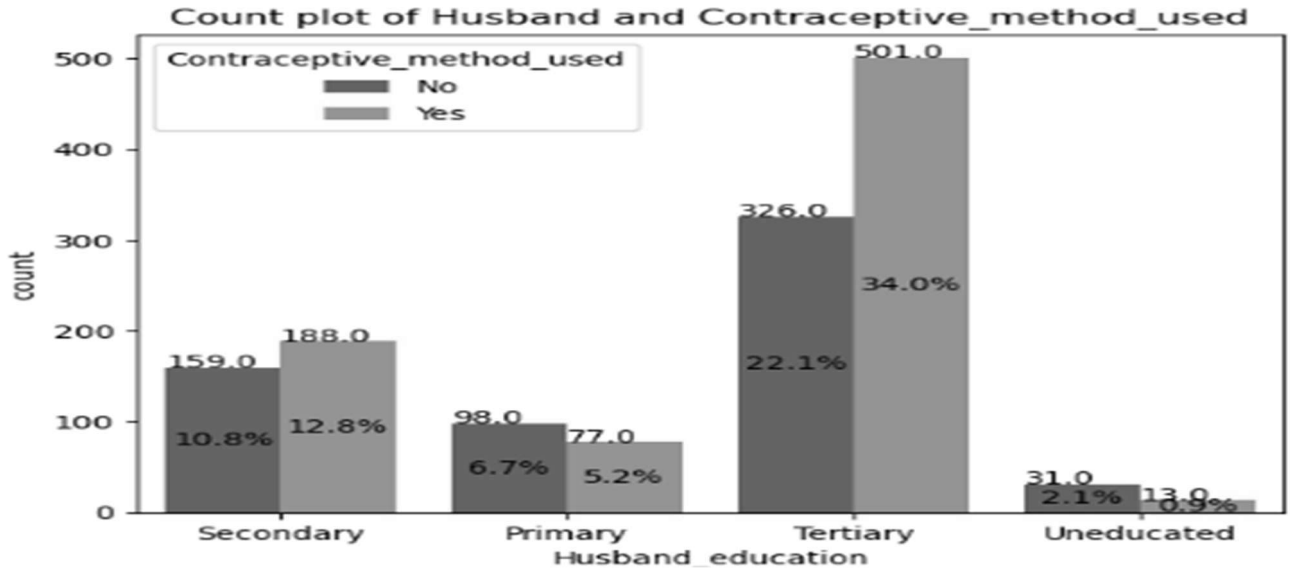


Fig. 2.1H

Count plot between Husband Education and Contraceptive_method_used

Inferences from above Fig. 2.1H

- Most number of husband whose wife used contraceptive method have education level of Tertiary i.e., 501 or 34 %
- Most number of Husband whose wife doesn't used contraceptive method have education level of Uneducated i.e., 31 or 2.1 %

4. **Husband_education and Contraceptive_method_used**

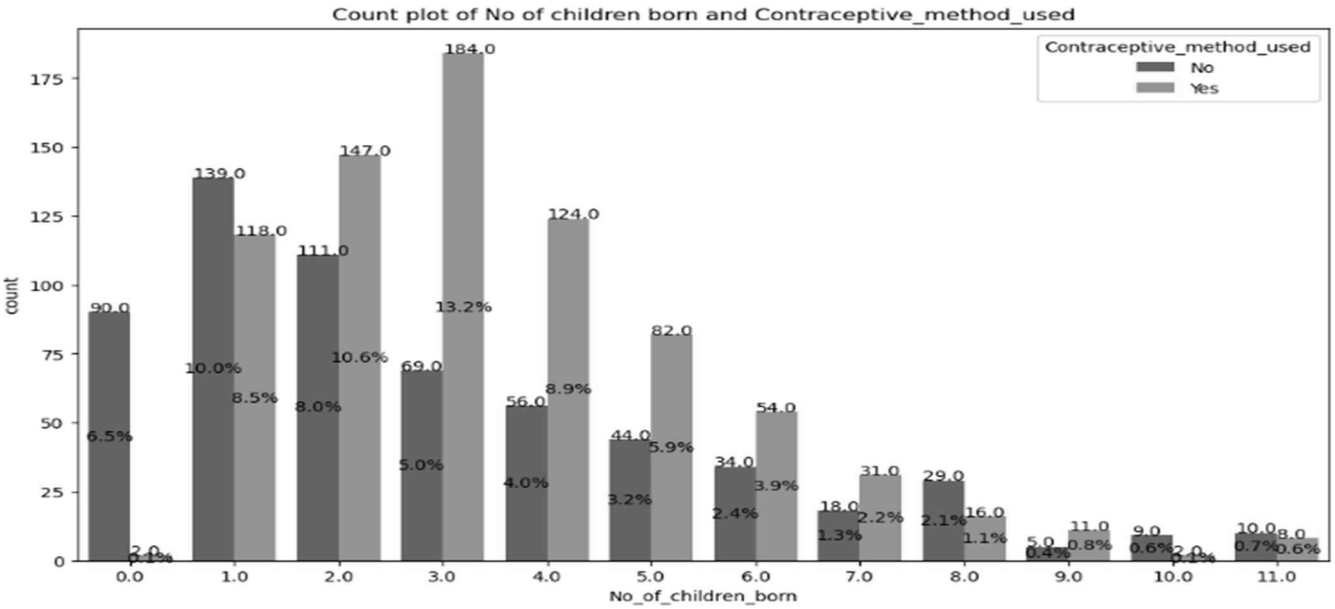


Fig. 2.1I

Count plot between No_of_Children_born and Contraceptive_method_used

Inferences from above Fig. 2.1I

- Most number of Children born even after using contraceptive method = 3 i.e., 13.2%
- Most number of Children born even without using contraceptive method = 1 i.e., 10%

5. **Wife_Religion and Contraceptive_method_used**

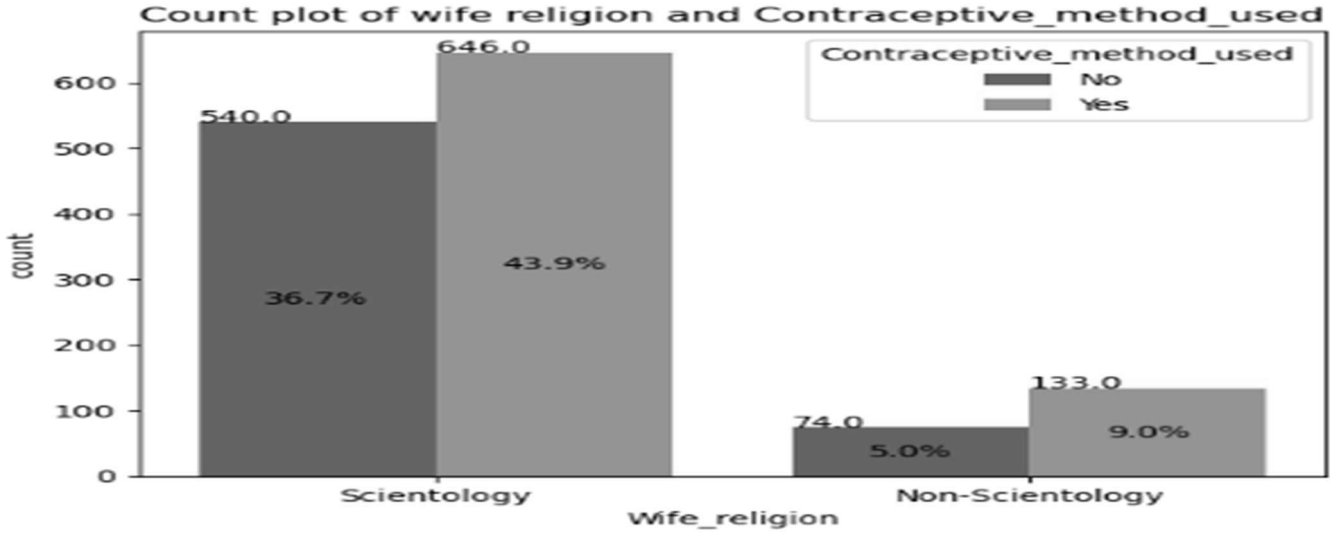


Fig. 2.1J

Count plot between Wife_religion and Contraceptive_method_used

Inferences from above Fig. 2.1J

- For Wife's whose religion be Scientology or Non-Scientology have preferred contraceptive method

6. **Wife_working and Contraceptive_method_used**

Fig. 2.1K

Count plot between Wife_working and Contraceptive_method_used

Inferences from above Fig. 2.1K

- For Wife's who are working or not working both have preferred contraceptive method

7. **Husband_Occupation and Contraceptive_method_used**

Fig. 2.1L

Count plot between Husband_Occupation and Contraceptive_method_used

Inferences from above Fig. 2.1L

- Wife's of Husband whose occupation is of Type 3 has used max. Contraceptive method i.e., 21.5 %
- Wife's of Husband whose occupation is of Type 3 has not used Contraceptive method i.e., 17.2 %

8. **Standard_of_living_index and Contraceptive_method_used**

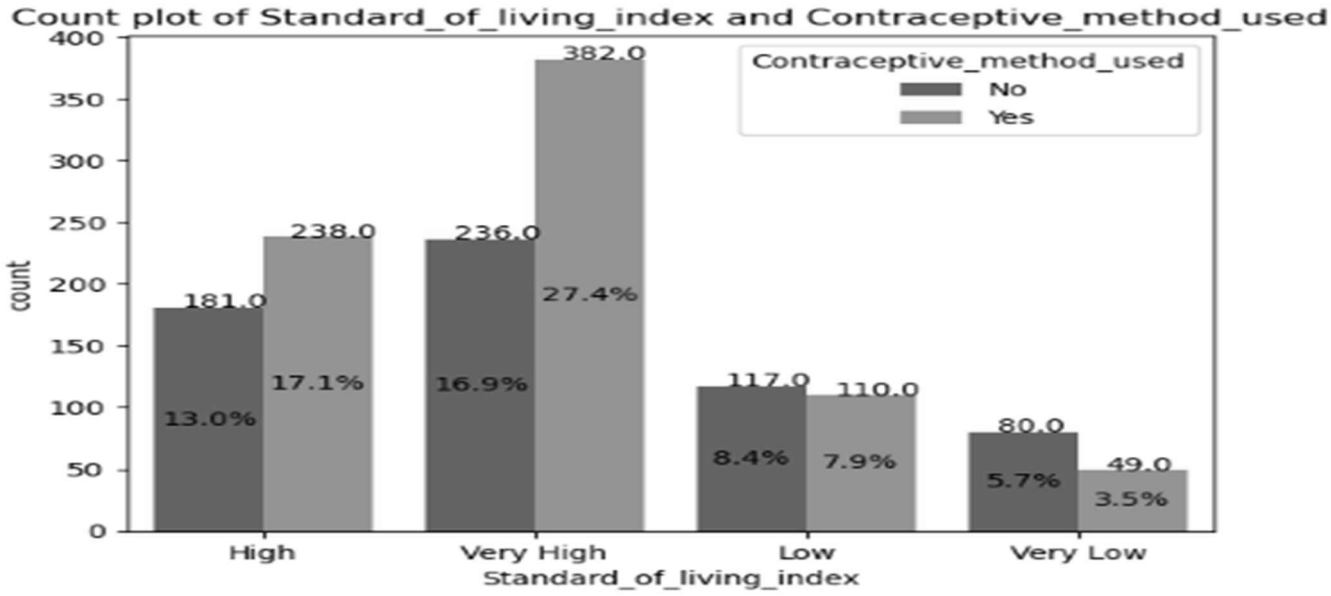


Fig. 2.1M

Count plot between Standard_of_living_index and Contraceptive_method_used

Inferences from above Fig. 2.1M

- Family having standard of living index very high has maximum used contraceptive i.e., 382 or 27.4 %
- Family having standard of living index very low has minimum number of wife's who used contraceptive i.e., 80 or 5.7%

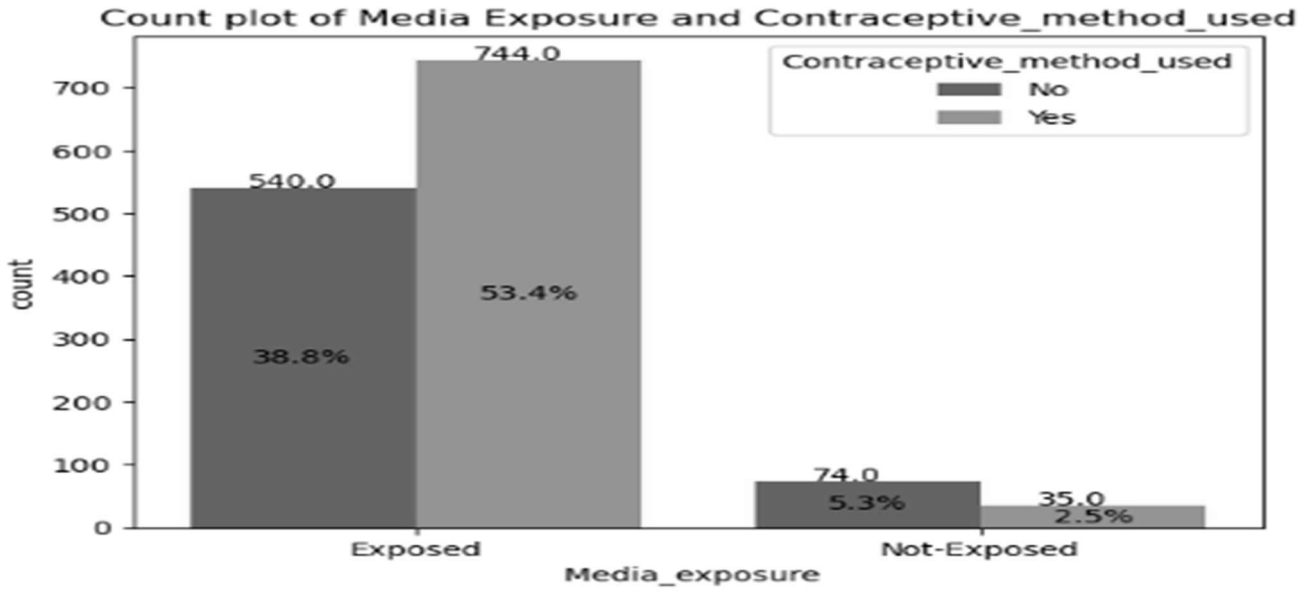


Fig. 2.1M

Count plot between Standard_of_living_index and Contraceptive_method_used

Inferences from above Fig. 2.1M

- Family who are not exposed to media have mostly not used contraceptive method
- Family who are exposed to media have mostly used contraceptive method

 # Mutli-Variate Analysis

Pairplot (pairwise relationship between continuous variables)

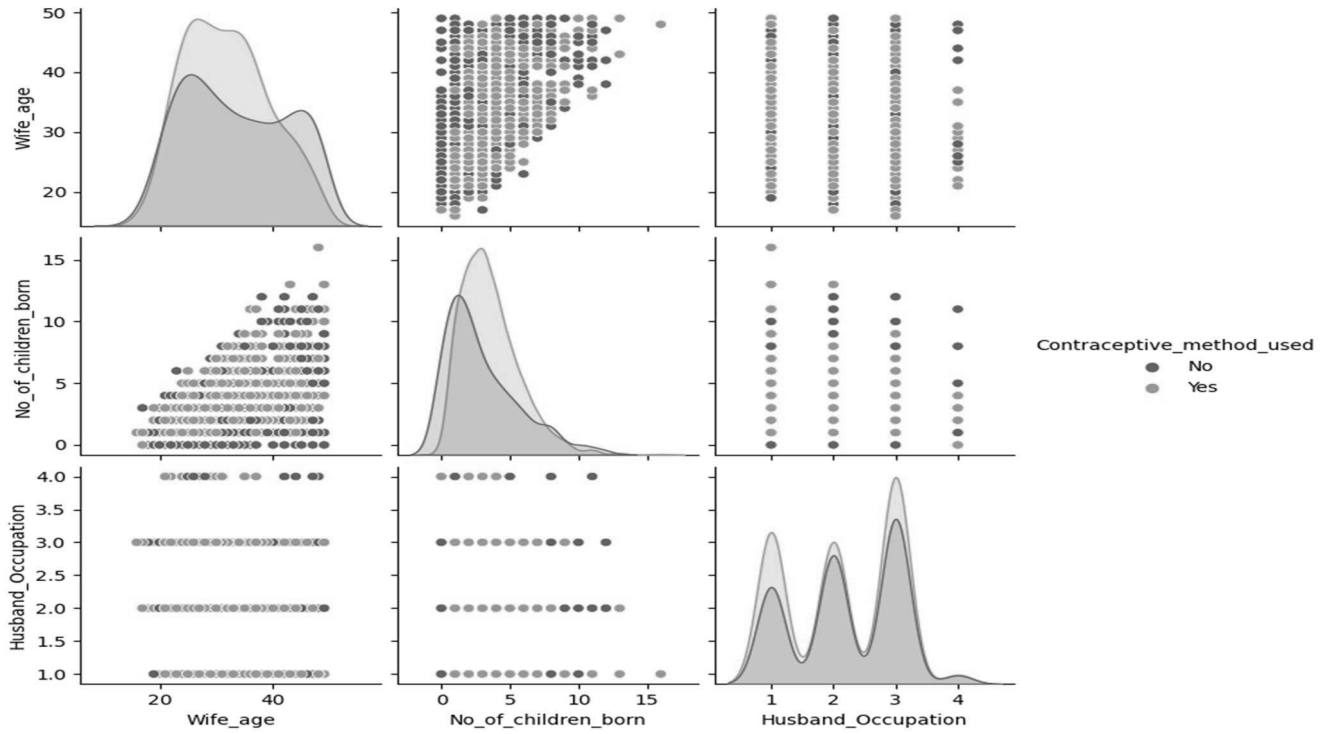


Fig. 2.1N

PairPlot (pairwise relationship between continuous variables)

 # Heatmap of Numeric variables

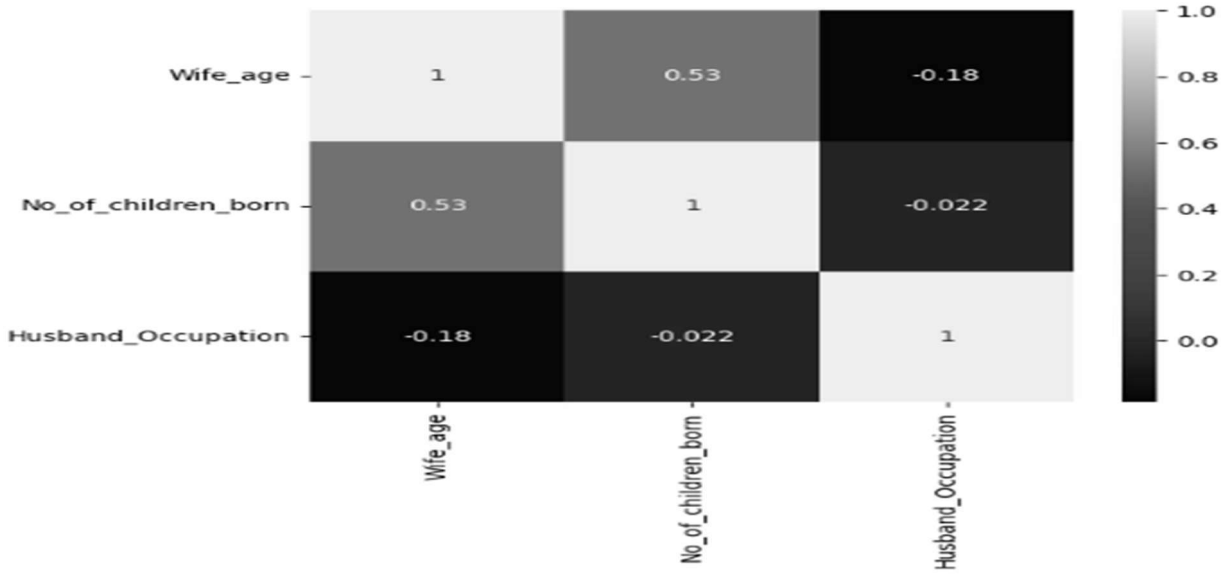


Fig 2.1O

Heatmap of Numeric variables

Inferences from Fig 2.1O

- There is a slight strong relationship between Wife_age and No_of_children_born i.e., 0.53
- There is no correlation between husband Occupation and wife age or No of children born

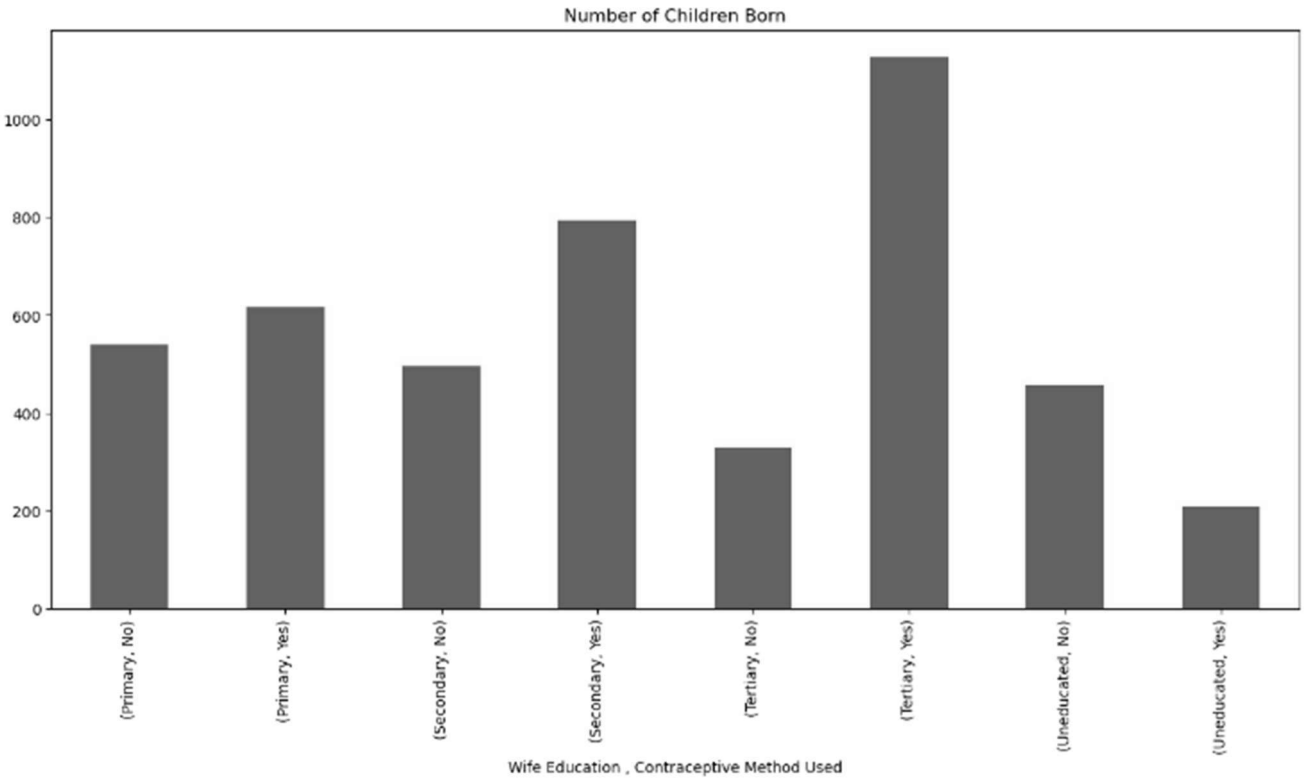


Fig 2.1P

Count plot between Number of children born/Wife Education/ Contraceptive Method used

Inferences from Fig 2.1P

- Most number of children were born from wife having tertiary education and Contraceptive Method used

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.

Solution: -

Encoding of Categorical data in order to perform Classification: -

First we have done the Encoding for Ordinal categories

1. **For Wife Age columns: -**

```
categories in Wife Education
 Tertiary       515
Secondary       398
Primary         330
Uneducated      150
Name: wife_education, dtype: int64
Ordinal categories after Encoding
 4     515
3      398
2      330
1      150
Name: wife_education, dtype: int64
```

2. **Husband Education**

```
categories in Husband Education
 Tertiary       827
Secondary       347
Primary         175
Uneducated       44
Name: Husband_education, dtype: int64

 Ordinal categories after Encoding
 4     827
3      347
2      175
1       44
Name: Husband_education, dtype: int64
```

3. **Standard of Living Index**

```
categories in Standard of Living Index
 Very High     618
High           419
Low            227
Very Low       129
Name: standard_of_living, dtype: int64

 Ordinal categories after Encoding
 4     618
3      419
2      227
1      129
Name: standard_of_living, dtype: int64
```

Then we have done the Encoding for Binary categories using pd.get_dummies: -

1. Wife Religion: - Scientology = 1, Non-Scientology = 0
2. Wife Working: - Yes = 1, No = 0
3. Media Exposure: - Exposed = 1, Not Exposed = 0

Also we have dropped the first dummy for each column.

Let's look into Dependent Variable Column i.e., Contraceptive Method Used: -

```
Number of each categories in Contraceptive Method Used : -
 Yes    779
No      614
Name: Contraceptive_method_used, dtype: int64
Encoded categories: -
 1    779
0     614
Name: Contraceptive_method_used, dtype: int64
```

Fig. 2.2 A

Before and After Encoding of Target/Dependent column

Inference from above Fig 2.2A: -

- There are 779 (Yes) and 614 (No) in Contraceptive Method Used column
- We encoded 1 to Yes and 0 to No
- Yes (1) = 779 & No (0) = 614

	0	1	2	3	4
Wife_age	24.0	45.0	43.0	42.0	36.0
wife_education	2.0	1.0	2.0	3.0	3.0
Husband_education	3.0	3.0	3.0	2.0	3.0
No_of_children_born	3.0	10.0	7.0	9.0	8.0
Husband_Occupation	2.0	3.0	3.0	3.0	3.0
standard_of_living	3.0	4.0	4.0	3.0	2.0
Contraceptive_method_used	1.0	1.0	1.0	1.0	1.0
Wife_religion_Scientology	1.0	1.0	1.0	1.0	1.0
Wife_Working_Yes	0.0	0.0	0.0	0.0	0.0
media_exposure_Not-Exposed	0.0	0.0	0.0	0.0	0.0

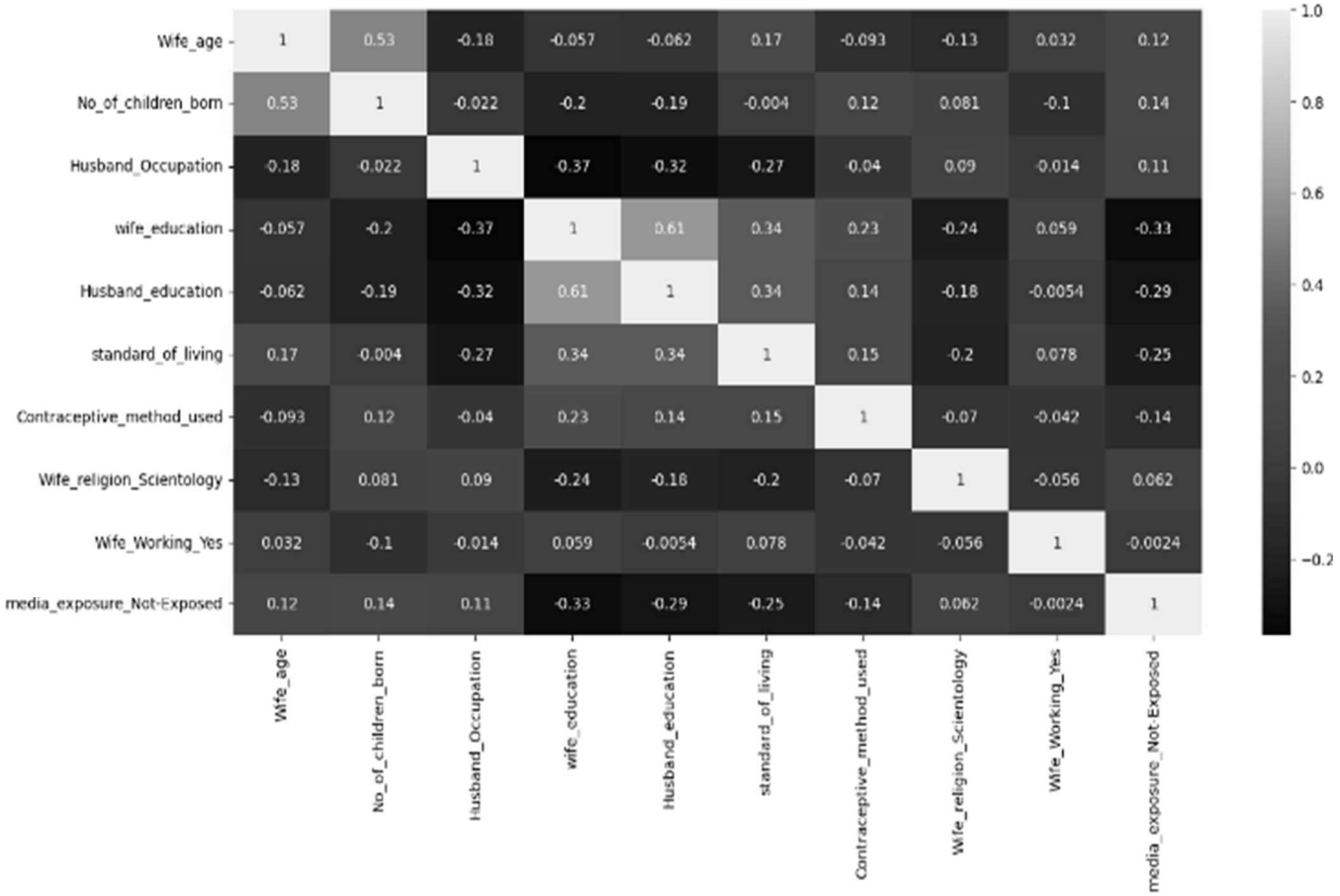
	1468	1469	1470	1471	1472
Wife_age	33.0	33.0	39.0	33.0	17.0
No_of_children_born	3.0	3.0	3.0	3.0	1.0
Husband_Occupation	2.0	1.0	1.0	2.0	2.0
wife_education	4.0	4.0	3.0	3.0	3.0
Husband_education	4.0	4.0	3.0	3.0	3.0
standard_of_living	4.0	4.0	4.0	2.0	4.0
Contraceptive_method_used	1.0	1.0	1.0	1.0	1.0
Wife_religion_Scientology	1.0	1.0	1.0	1.0	1.0
Wife_Working_Yes	1.0	0.0	1.0	1.0	0.0
media_exposure_Not-Exposed	0.0	0.0	0.0	0.0	0.0

Table 2.2A

First and Last 5 Rows after Encoding

Lets split our data set into test and train in ratio of 30:70 i.e., Test data = 30%, Train data = 70%
After splitting the dataset let perform Classification Model 1 by 1

 # Checking for correlation after encoding of Categorical variables



Heat Map (After encoding of categorical columns)

Inference from above Fig: -

- Husband education and wife education are slightly correlated
- Wife Age and Number of Children born are also slightly correlated

Classification Using Logistic Regression

First we have created a simple Logistic Regression model and calculated the Confusion Matrix and Classification Report

Classification Report performing simple Logistic Regression on Train Data

	precision	recall	f1-score	support
0	0.66	0.52	0.58	430
1	0.68	0.79	0.73	545
accuracy			0.67	975
macro avg	0.67	0.65	0.65	975
weighted avg	0.67	0.67	0.66	975

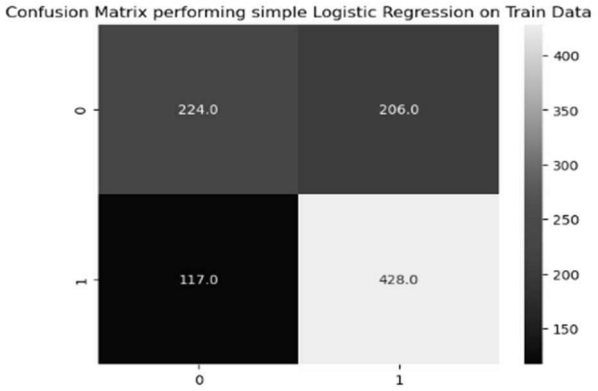


Fig. 2.2B

Classification Report and Confusion matrix with simple Logistic Regression (Train Data)

Inferences from above Fig. 2.2B

- For Train Data out of 545 number of 1's(Yes), 428 were actually predicted as 1's (Yes)
- Also 117 number of 1's (Yes) were predicted as 0's and 207 number of 0's(No) were predicted as 1's(Yes)
- Recall of 1's(Yes) = 0.79 Recall of 0's(No) = 0.52
- Precision of 1's (Yes) = 0.66 Precision of 0's(No) = 0.68
- Accuracy = 0.67
- Model score for train data (without using GridSearchCV) = 66.87

Classification Report performing simple Logistic Regression on Test Data

	precision	recall	f1-score	support
0	0.62	0.46	0.53	184
1	0.65	0.78	0.71	234
accuracy			0.64	418
macro avg	0.64	0.62	0.62	418
weighted avg	0.64	0.64	0.63	418

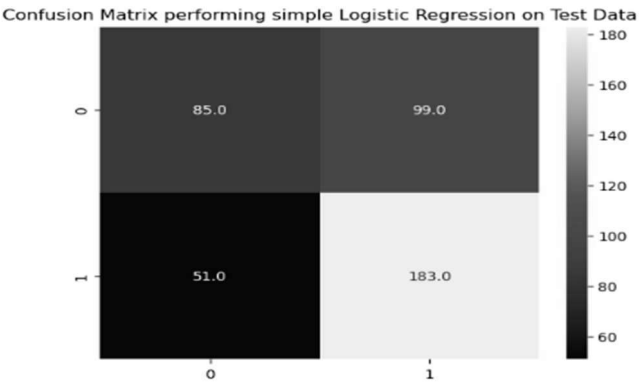


Fig. 2.2C

Classification Report and Confusion matrix with simple Logistic Regression (Test Data)

Inferences from above Fig. 2.2B

- For Test Data out of 234 number of 1's(Yes), 183 were actually predicted as 1's (Yes)
- Also 51 numbers of 1's (Yes) were predicted as 0's and 99 number of 0's(No) were predicted as 1's(Yes)
- Recall of 1's(Yes) = 0.78 Recall of 0's(No) = 0.46
- Precision of 1's (Yes) = 0.65 Precision of 0's(No) = 0.62
- Accuracy = 0.64
- Model score for train data (without using GridSearchCV) = 64.11

Now we will create Logistic Regression model using GridSearchCV parameter and calculated the Confusion Matrix and Classification Report

GridSearchCV helps us to calibrate our model using different parameter such as penalty, solver, tol, etc.
For GridSearchCV () we need to create a dictionary, which has all the parameter in order to calibrate our model.
Let's check our best parameter's for the model after applying GridSearchCV : -

```
The best paarmeter after applying GridSearchCV to our model
 LogisticRegression(max_iter=300, random_state=1, solver='saga')
```

 # Logistic Regression Model Prediction Probabilities on Train Dataset

	0	1	2	3	4	5	6	7	8	9	...	965	966	967	968	969
0-No	0.726992	0.190232	0.190744	0.415129	0.709844	0.478782	0.480257	0.691034	0.162164	0.309134	...	0.527662	0.411215	0.268884	0.678126	0.183435
1-Yes	0.273008	0.809768	0.809256	0.584871	0.290156	0.521218	0.519743	0.308966	0.837836	0.690866	...	0.472338	0.588785	0.731116	0.323874	0.816565

Let's create a Logistic Regression Model using these parameters and calculated the Confusion Matrix and Classification Report: -

 # Logistic Regression Classification Report and Confusion matrix (with GridSearchCV) on Train Data

Logistic Regression Classification Report(with GridSearchCV) on Train Data

	precision	recall	f1-score	support
0	0.66	0.52	0.58	430
1	0.68	0.79	0.73	545
accuracy			0.67	975
macro avg	0.67	0.65	0.65	975
weighted avg	0.67	0.67	0.66	975

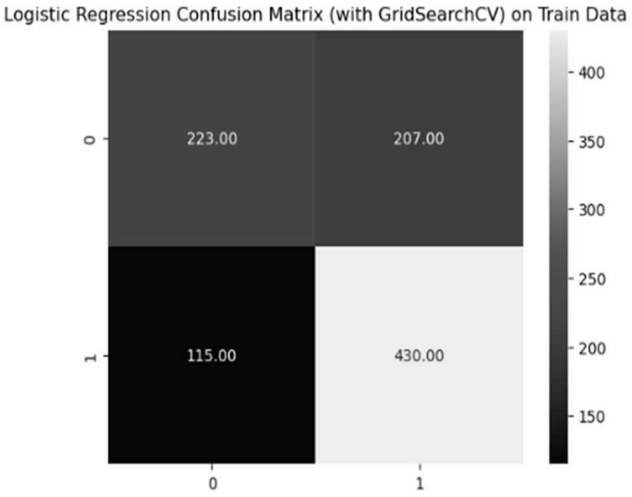


Fig. 2.2D
Logistic Regression Classification Report and Confusion matrix (with GridSearchCV) on Train Data

Inferences from above Fig. 2.2D

- For Train Data out of 545 number of 1's(Yes), 430 were actually predicted as 1's (Yes)
- Also 115 number of 1's (Yes) were predicted as 0's and 207 number of 0's(No) were predicted as 1's(Yes)
- Recall of 1's(Yes) = 0.79 Recall of 0's(No) = 0.52

- Precision of 1's (Yes) = 0.66 Precision of 0's(No) = 0.68
- Accuracy = 0.67

Logistic Regression (with GridSearchCV) Model score for train data = 66.97

 # Logistic Regression Model Prediction Probabilities on Test Dataset

	0	1	2	3	4	5	6	7	8	9	...	408	409	410	411	412
0-No	0.274183	0.596588	0.33523	0.286708	0.235427	0.2778	0.285201	0.285231	0.515282	0.240069	...	0.613029	0.737835	0.174687	0.475656	0.439187	0.3
1-Yes	0.725817	0.403414	0.66477	0.713292	0.764573	0.7222	0.714799	0.714769	0.484718	0.759931	...	0.386971	0.262165	0.825313	0.524344	0.560813	0.6

 # Logistic Regression Classification Report and Confusion matrix (with GridSearchCV) on Test Data

Logistic Regression Classification Report(with GridSearchCV) on Test Data

	precision	recall	f1-score	support
0	0.64	0.46	0.53	184
1	0.65	0.79	0.72	234
accuracy			0.65	418
macro avg	0.64	0.63	0.62	418
weighted avg	0.64	0.65	0.63	418

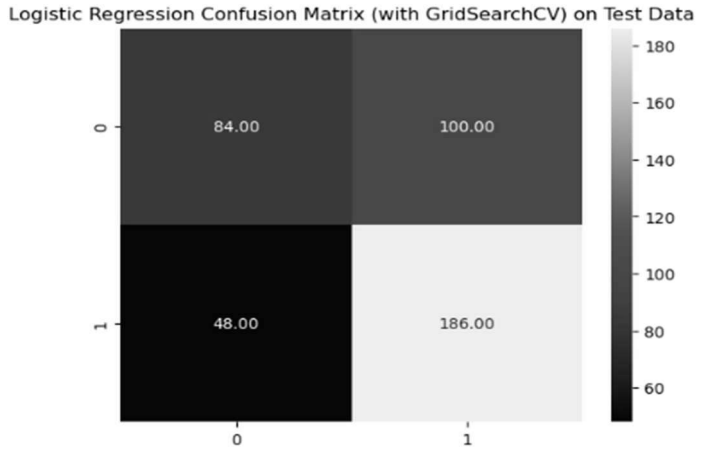


Fig. 2.2E
Logistic Regression Classification Report and Confusion matrix (with GridSearchCV) on Test Data

Inferences from above Fig. 2.2E

- For Test Data out of 234 number of 1's(Yes), 186 were actually predicted as 1's (Yes)
- Also 48 number of 1's (Yes) were predicted as 0's and 100 number of 0's(No) were predicted as 1's(Yes)
- Recall of 1's(Yes) = 0.79 Recall of 0's(No) = 0.46
- Precision of 1's (Yes) = 0.65 Precision of 0's(No) = 0.64
- Accuracy = 0.65

Logistic Regression (with GridSearchCV) Model score for test data = 64.59

 # Lets Calculate ROC AUC Score and Create ROC Curve for Train and Test Data

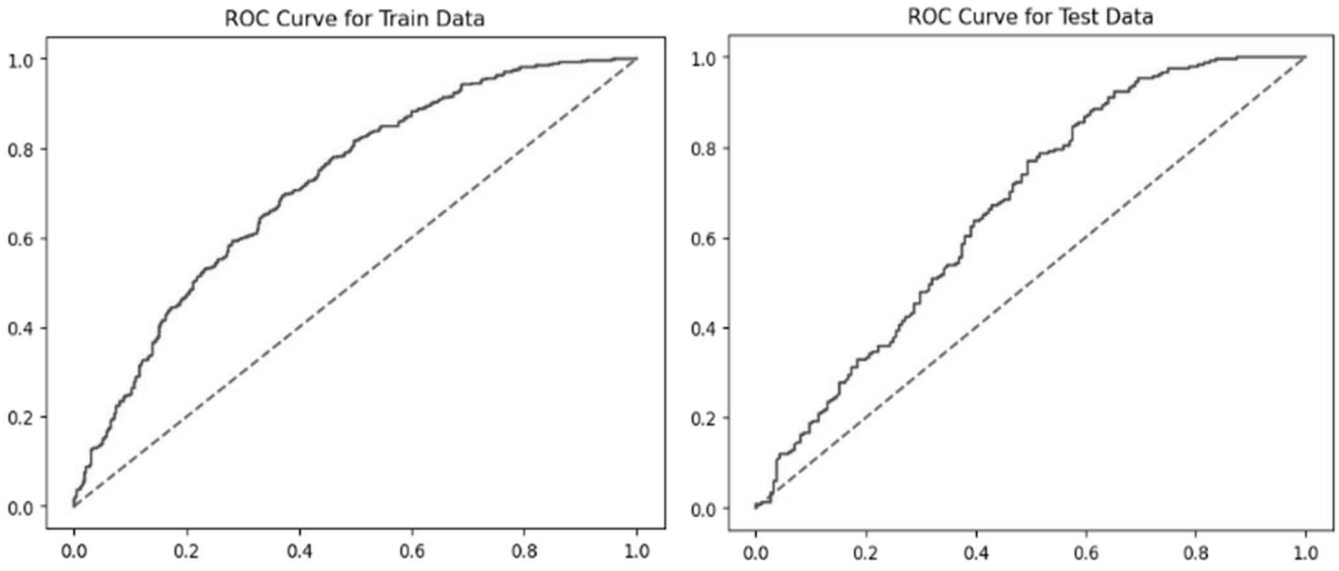


Fig. 2.2F

Logistic Regression Model ROC Curve for train and test data

- Logistic Regression Model ROC AUC Score for Train data = 65.37
- Logistic Regression Model ROC AUC Score for Test data = 62.57

Classification Using Linear Discriminant Analysis (LDA)

Let's create a model using LDA: -

A. Intercept of the LDA Model

```
The Intercept of the LDA model :-  [-0.89472272]
```

B. Coefficient of each Independent variable for LDA Equation: -

```
The coefficient each varaible of the LDA model :- [[-0.072  0.319  0.14   0.518  0.041  0.318 -0.444 -0.171 -0.354]]
```

C. LDA model Equation: -

Equation:

Contraceptive Method Used = [-0.895] +(-0.072) * WIFE_AGE +(0.319) * NO_OF_CHILDREN_BORN +(0.14) * HUSBAND_OCCUPATION +(0.518) * WIFE_EDUCATION +(0.041) * HUSBAND_EDUCATION +(0.318) * STANDARD_OF_LIVING +(-0.444) * WIFE_RELIGION_SCIENTOLOGY +(-0.171) * WIFE_WORKING_YES +(-0.354)* MEDIA_EXPOSURE_NOT-EXPOSED

 # LDA Model Prediction Probabilities on Test Dataset

	0	1	2	3	4	5	6	7	8	9	...	965	966	967	968	969	
0-No	0.689844	0.179351	0.183432	0.387694	0.71348	0.456307	0.472451	0.692721	0.159148	0.304567	...	0.514344	0.39511	0.262263	0.691146	0.181185	0.5
1-Yes	0.310356	0.820649	0.816568	0.612306	0.28652	0.543693	0.527549	0.307279	0.840852	0.695433	...	0.485656	0.60489	0.737737	0.308854	0.818815	0.4

2 rows × 975 columns

 # LDA Confusion Matrix and Classification Report on train data

LDA Classification Report on train data

	precision	recall	f1-score	support
0	0.67	0.51	0.58	430
1	0.68	0.81	0.73	545
accuracy			0.67	975
macro avg	0.67	0.66	0.66	975
weighted avg	0.67	0.67	0.67	975

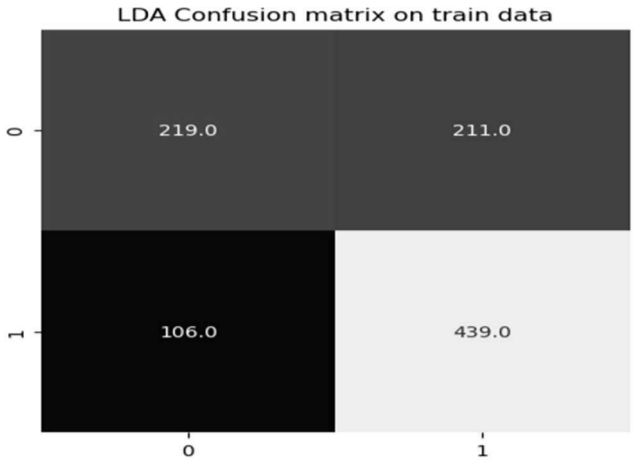


Fig. 2.2G
LDA Classification Report and Confusion matrix (Train Data)

Inferences from above Fig. 2.2G

- For Train Data out of 545 number of 1's(Yes), 439 were actually predicted as 1's (Yes)
- Also 106 number of 1's (Yes) were predicted as 0's and 211 number of 0's(No) were predicted as 1's(Yes)
- Recall of 1's(Yes) = 0.81 Recall of 0's(No) = 0.51
- Precision of 1's (Yes) = 0.68 Precision of 0's(No) = 0.67
- Accuracy = 0.67

 # LDA Model score for train data = 67.48

 # LDA Model Prediction Probabilities on Test Dataset

	0	1	2	3	4	5	6	7	8	9	...	408	409	410	411	412
0-No	0.280734	0.60973	0.333497	0.288619	0.234016	0.272349	0.287621	0.277177	0.482781	0.235902	...	0.601639	0.751523	0.17943	0.473966	0.427266	0.
1-Yes	0.719266	0.39027	0.666503	0.713381	0.765984	0.727651	0.712379	0.722823	0.517219	0.764098	...	0.398361	0.248477	0.82057	0.526034	0.572734	0.

2 rows × 418 columns

 # LDA Confusion Matrix and Classification Report on Test Dataset

LDA Classification Report on test data

	precision	recall	f1-score	support
0	0.64	0.44	0.52	184
1	0.65	0.80	0.72	234
accuracy			0.64	418
macro avg	0.64	0.62	0.62	418
weighted avg	0.64	0.64	0.63	418

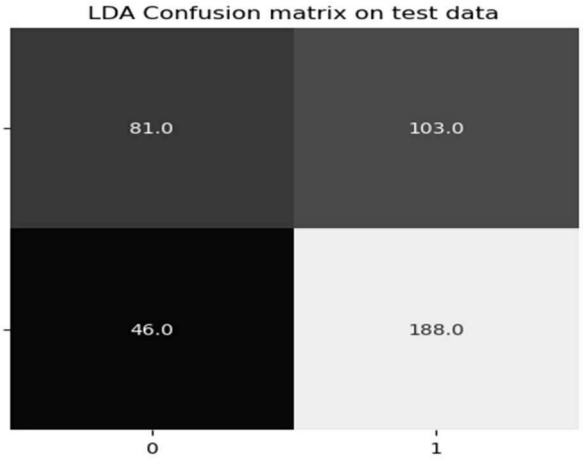


Fig. 2.2H
LDA Classification Report and Confusion matrix (Test Data)

Inferences from above Fig. 2.2H

- For Test Data out of 234 number of 1's(Yes), 188 were actually predicted as 1's (Yes)
- Also 46 number of 1's (Yes) were predicted as 0's and 103 number of 0's(No) were predicted as 1's(Yes)
- Recall of 1's(Yes) = 0.80 Recall of 0's(No) = 0.44
- Precision of 1's (Yes) = 0.65 Precision of 0's(No) = 0.64
- Accuracy = 0.64

LDA Model score for test data = 64.35

 # Lets Calculate LDA Model ROC AUC Score and Create ROC Curve for Train and Test Data

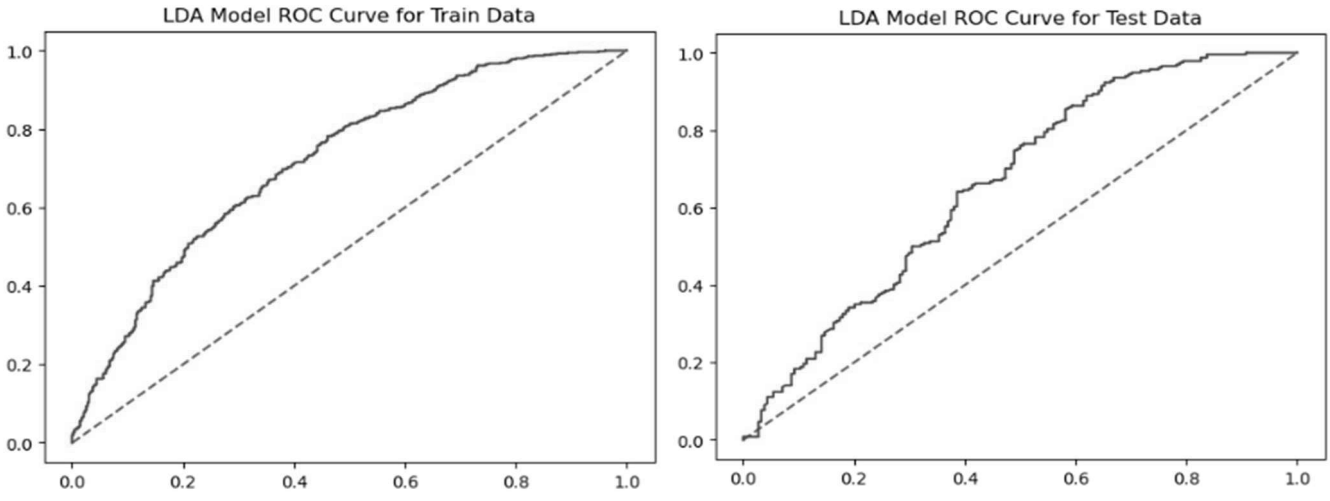


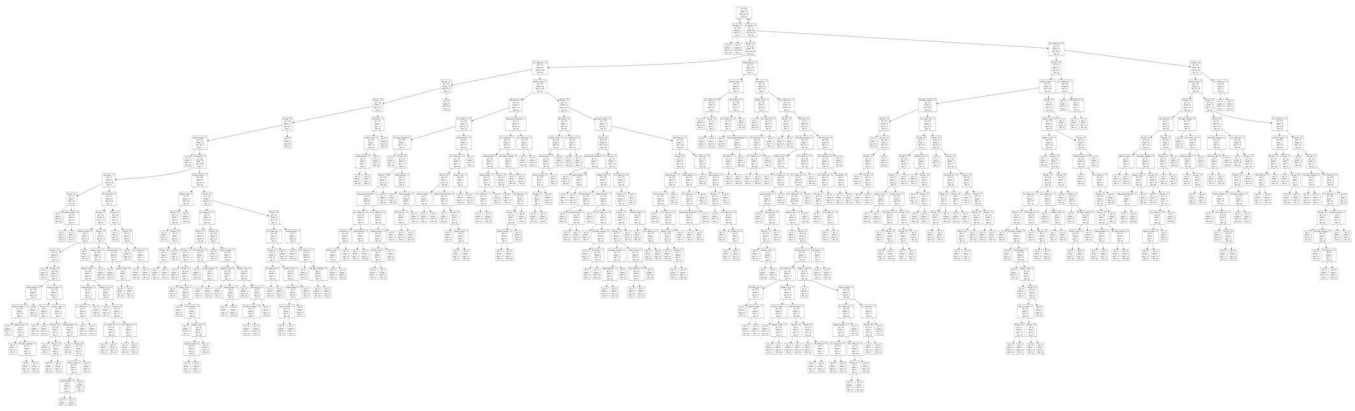
Fig. 2.2I
LDA Model ROC Curve for train and test data

- LDA Model ROC AUC Score for Train Data = 65.740
- LDA Model ROC AUC Score for Test Data = 62.18

Classification Using CART (Decision Tree)

Let's create a model using CART: -

A. Decision Tree (using graphic wiz)



B. Feature Importance on performing DecisionTreeClassifier with Default Parameters: -

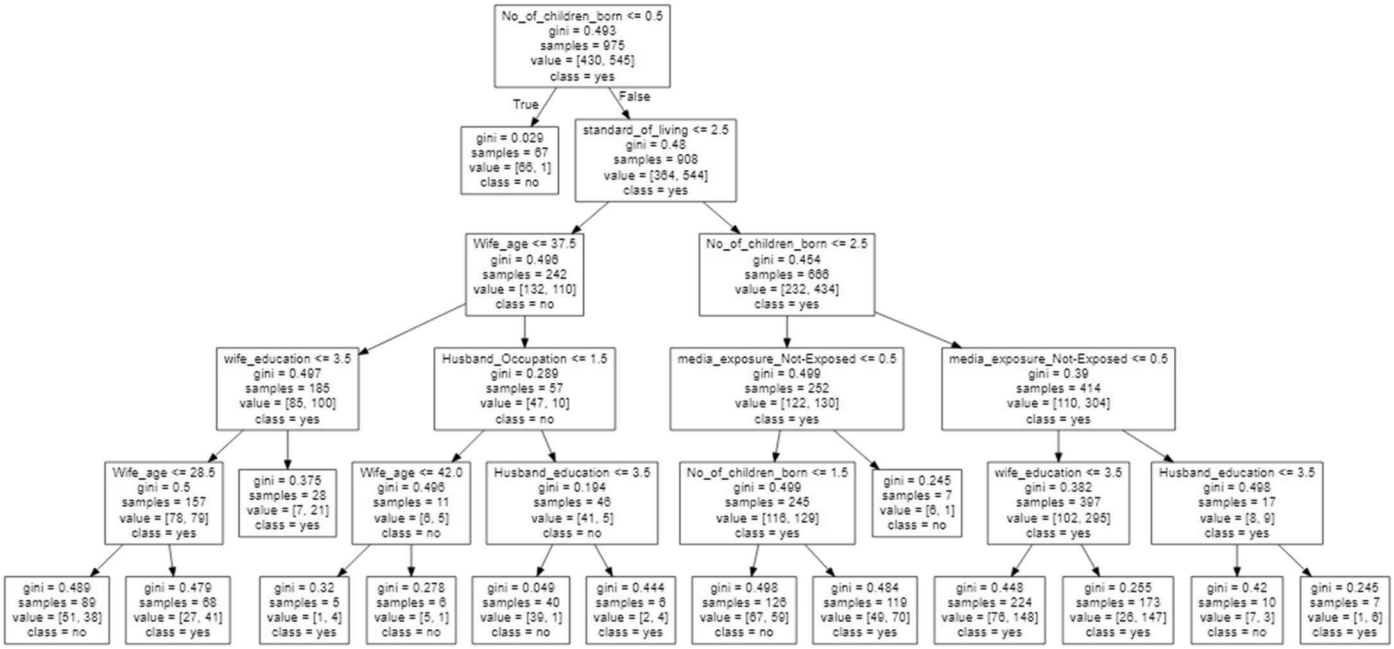
	Imp
Wife_age	0.320963
No_of_children_born	0.241714
standard_of_living	0.088392
wife_education	0.088194
Husband_education	0.082927
Husband_Occupation	0.079430
Wife_Working_Yes	0.043602
Wife_religion_Scientology	0.040408
media_exposure_Not-Exposed	0.014370

Let's apply GridSearchCV () in order to find best parameter for the model

A. Best parameters after applying GridSearchCV: -

```
{'ccp_alpha': 0.001, 'criterion': 'gini', 'max_depth': 5, 'max_features': 'log2', 'min_samples_leaf': 5}
```

B. Decision Tree with best parameters (using graphic wiz)



C. Feature Importance after performing DecisionTreeClassifier with best parameters: -

	Imp
No_of_children_born	0.531256
Wife_age	0.145065
standard_of_living	0.123564
wife_education	0.088411
Husband_education	0.061393
media_exposure_Not-Exposed	0.031285
Husband_Occupation	0.019026
Wife_religion_Scientology	0.000000
Wife_Working_Yes	0.000000

 # CART Model Prediction Probabilities on Train Dataset

	0	1	2	3	4	5	6	7	8	9	...	965	966	967	968	969
0-No	0.985075	0.150289	0.339286	0.411765	0.975	0.333333	0.339286	0.573034	0.985075	0.531746	...	0.397059	0.150289	0.150289	0.985075	0.339286	0.5
1-Yes	0.014925	0.849711	0.660714	0.588235	0.025	0.666667	0.660714	0.426966	0.014925	0.468254	...	0.602941	0.849711	0.849711	0.014925	0.660714	0.4

2 rows × 975 columns

 # CART Confusion Matrix and Classification Report on train data

CART Model Classification Report on train data

	precision	recall	f1-score	support
0	0.70	0.56	0.62	430
1	0.70	0.81	0.75	545
accuracy			0.70	975
macro avg	0.70	0.68	0.69	975
weighted avg	0.70	0.70	0.69	975

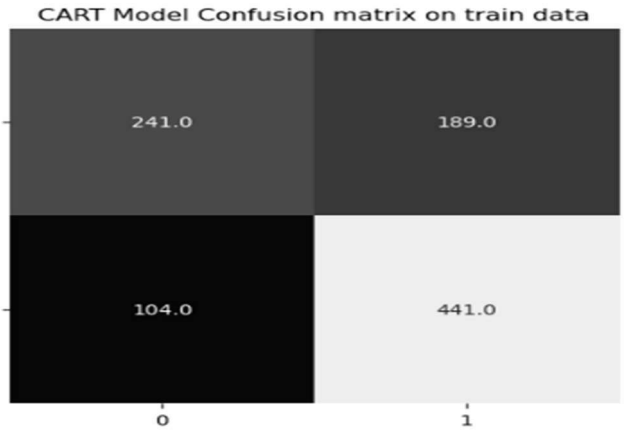


Fig 2.2J

CART Classification Report and Confusion matrix (Train Data)

Inferences from above Fig. 2.2J

- For Train Data out of 545 number of 1's(Yes), 441 were actually predicted as 1's (Yes)
- Also 104 number of 1's (Yes) were predicted as 0's and 189 number of 0's(No) were predicted as 1's(Yes)
- Recall of 1's(Yes) = 0.81 Recall of 0's(No) = 0.56
- Precision of 1's (Yes) = 0.70 Precision of 0's(No) = 0.70
- Accuracy = 0.70

 # CART Model score for train data = 69.95

 # CART Model Prediction Probabilities on Test Dataset

	0	1	2	3	4	5	6	7	8	9	...	408	409	410	411	412
0-No	0.531746	0.985075	0.411765	0.411765	0.150289	0.339286	0.339286	0.150289	0.411765	0.150289	...	0.339286	0.397059	0.150289	0.339286	0.531746
1-Yes	0.468254	0.014925	0.588235	0.588235	0.849711	0.660714	0.660714	0.849711	0.588235	0.849711	...	0.660714	0.602941	0.849711	0.660714	0.468254

2 rows × 418 columns

 # CART Confusion Matrix and Classification Report on test data

CART Model Classification Report on test data

	precision	recall	f1-score	support
0	0.66	0.47	0.55	184
1	0.66	0.81	0.73	234
accuracy			0.66	418
macro avg	0.66	0.64	0.64	418
weighted avg	0.66	0.66	0.65	418

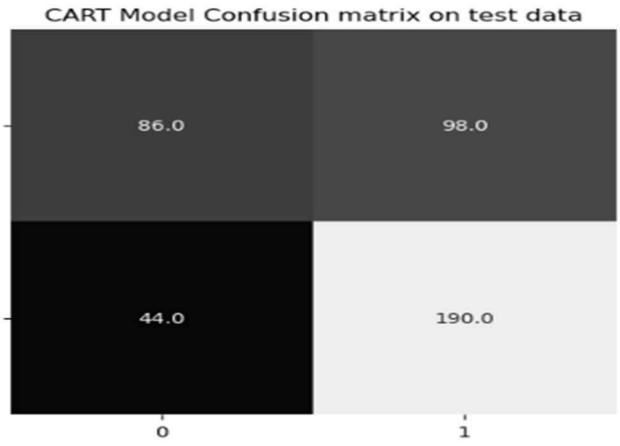


Fig 2.2K

CART Classification Report and Confusion matrix (Test Data)

Inferences from above Fig. 2.2K

- For Test Data out of 234 number of 1's(Yes), 190 were actually predicted as 1's (Yes)
- Also 44 numbers of 1's (Yes) were predicted as 0's and 98 number of 0's(No) were predicted as 1's(Yes)
- Recall of 1's(Yes) = 0.81 Recall of 0's(No) = 0.47
- Precision of 1's (Yes) = 0.66 Precision of 0's(No) = 0.66
- Accuracy = 0.66

 # CART Model score for test data = 66.028

 # Lets Calculate CART Model ROC AUC Score and Create ROC Curve for Train and Test Data

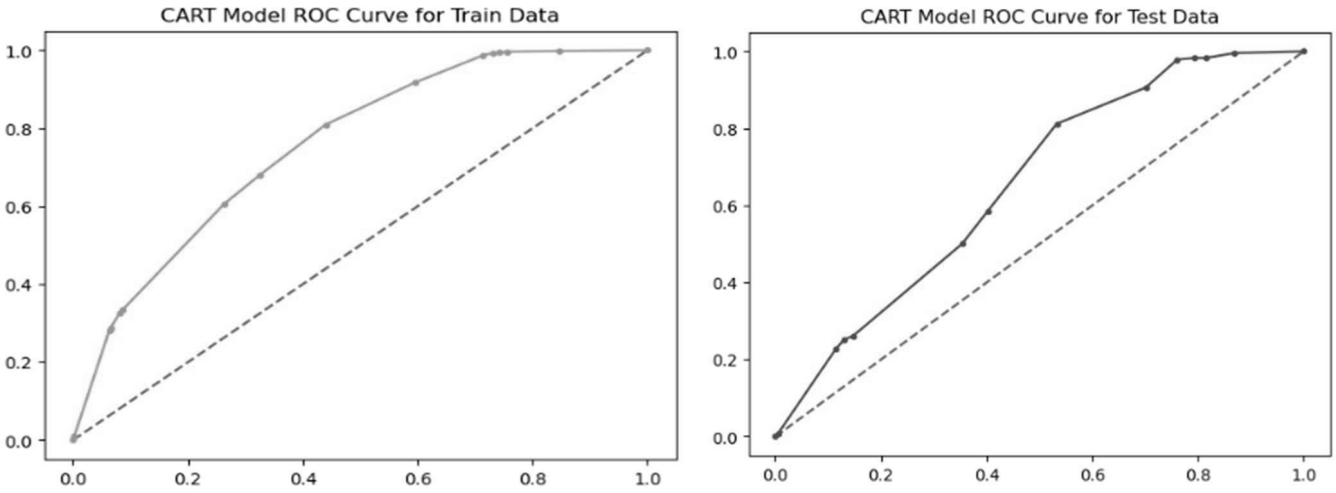


Fig. 2.2L

CART Model ROC Curve for train and test data

- CART Model ROC AUC Score for Train Data = 68.48
- CART Model ROC AUC Score for Test Data = 63.97

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.

Solution: -

Let's plot Confusion matrix and Classification report for train and test data of all 3 model i.e., Logistic Regression, Linear Discriminant Analysis(LDA) and KART Model (Decision Tree).

1. Logistic Regression Classification Report and Confusion Matrix

Train Data

Logistic Regression Classification Report(with GridSearchCV) on Train Data

	precision	recall	f1-score	support
0	0.66	0.52	0.58	430
1	0.68	0.79	0.73	545
accuracy			0.67	975
macro avg	0.67	0.65	0.65	975
weighted avg	0.67	0.67	0.66	975

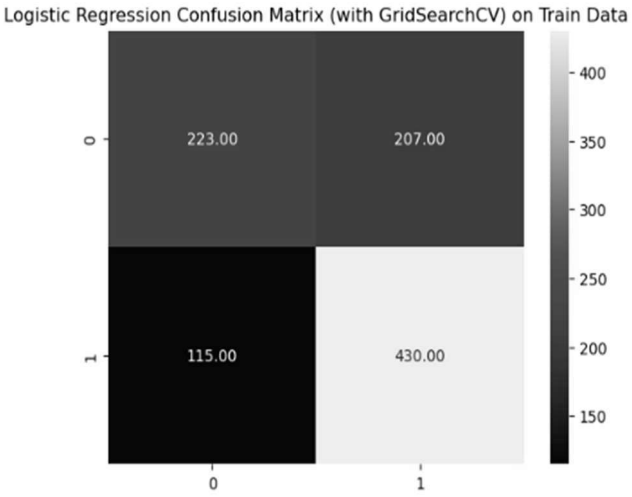


Fig. 2.3A

Logistic Regression Classification Report and Confusion matrix (with GridSearchCV) on Train Data

Test Data

Logistic Regression Classification Report(with GridSearchCV) on Test Data

	precision	recall	f1-score	support
0	0.64	0.46	0.53	184
1	0.65	0.79	0.72	234
accuracy			0.65	418
macro avg	0.64	0.63	0.62	418
weighted avg	0.64	0.65	0.63	418

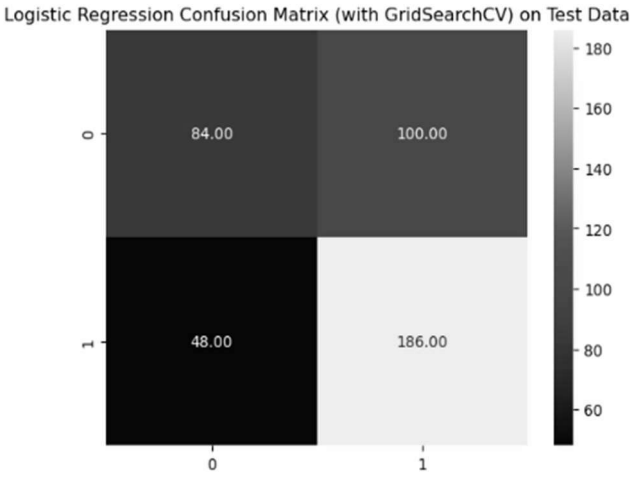


Fig. 2.3B

Logistic Regression Classification Report and Confusion matrix (with GridSearchCV) on Test Data

Inferences from Fig. 2.3A & Fig. 2.3B: -

- For Train Data out of 545 number of 1's(Yes), 430 were actually predicted as 1's (Yes)
- For Test Data out of 234 number of 1's(Yes), 186 were actually predicted as 1's (Yes)
- Recall for 1 (Yes) for train data as well as test data = 0.79
- Precision for 1(Yes) for train data = 0.68 | Precision for 1(Yes) test data are = 0.65
- F1 score for 1 (Yes) for train data = 0.73 | F1 score for 1 (Yes) for test data = 0.72
- Accuracy of train data = 0.67 | Accuracy of test data = 0.65

 # Logistic Regression model (with GridSearchCV) ROC AUC Score and ROC Curve

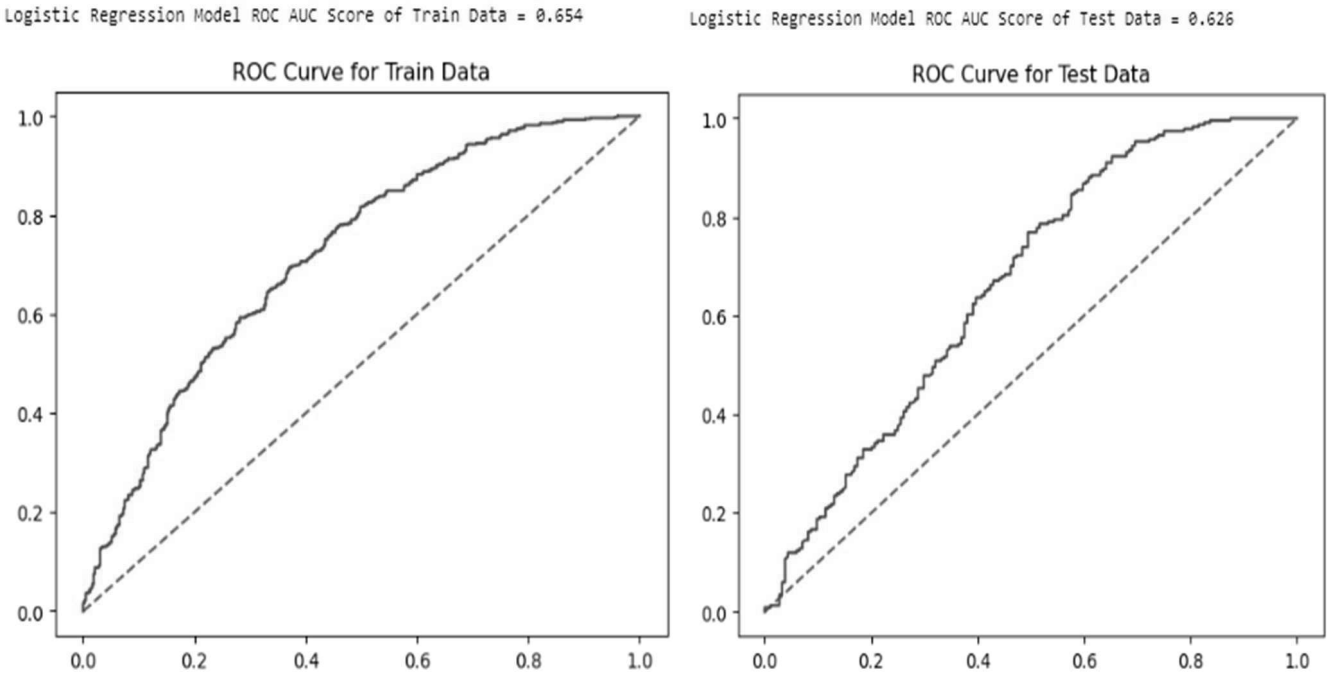


Fig. 2.3C
Linear Discriminant Analysis ROC AUC Score and ROC Curve for Train and Test Data

2. Linear Discriminant Analysis Classification Report / Confusion Matrix / ROC AUC Score/ROC Curve

Train Data

LDA Classification Report on train data

	precision	recall	f1-score	support
0	0.67	0.51	0.58	430
1	0.68	0.81	0.73	545
accuracy			0.67	975
macro avg	0.67	0.66	0.66	975
weighted avg	0.67	0.67	0.67	975

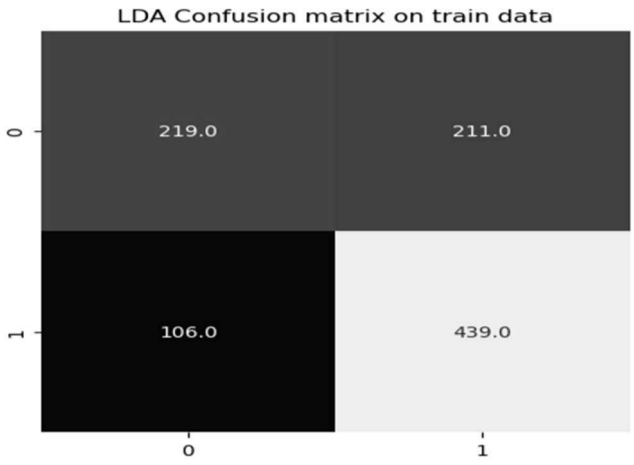


Fig. 2.3D
LDA Classification Report and Confusion matrix (Train Data)

Test Data

LDA Classification Report on test data

	precision	recall	f1-score	support
0	0.64	0.44	0.52	184
1	0.65	0.80	0.72	234
accuracy			0.64	418
macro avg	0.64	0.62	0.62	418
weighted avg	0.64	0.64	0.63	418

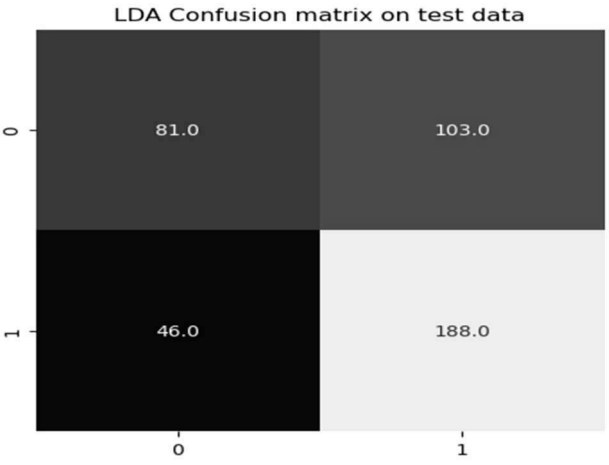


Fig. 2.3E
LDA Classification Report and Confusion matrix (Test Data)

Inferences from Fig. 2.3D & Fig. 2.3E: -

- For Train Data out of 545 number of 1's(Yes), 439 were actually predicted as 1's (Yes)
- For Test Data out of 234 number of 1's(Yes), 188 were actually predicted as 1's (Yes)
- Recall for 1 (Yes) for train data = 0.81 | Recall for 1 (Yes) for test data = 0.80
- Precision for 1(Yes) for train data = 0.68 | Precision for 1(Yes) test data are = 0.65
- F1 score for 1 (Yes) for train data = 0.73 | F1 score for 1 (Yes) for test data = 0.72
- Accuracy of train data = 0.67 | Accuracy of test data = 0.64

 # Linear Discriminant Analysis model ROC AUC Score and ROC Curve

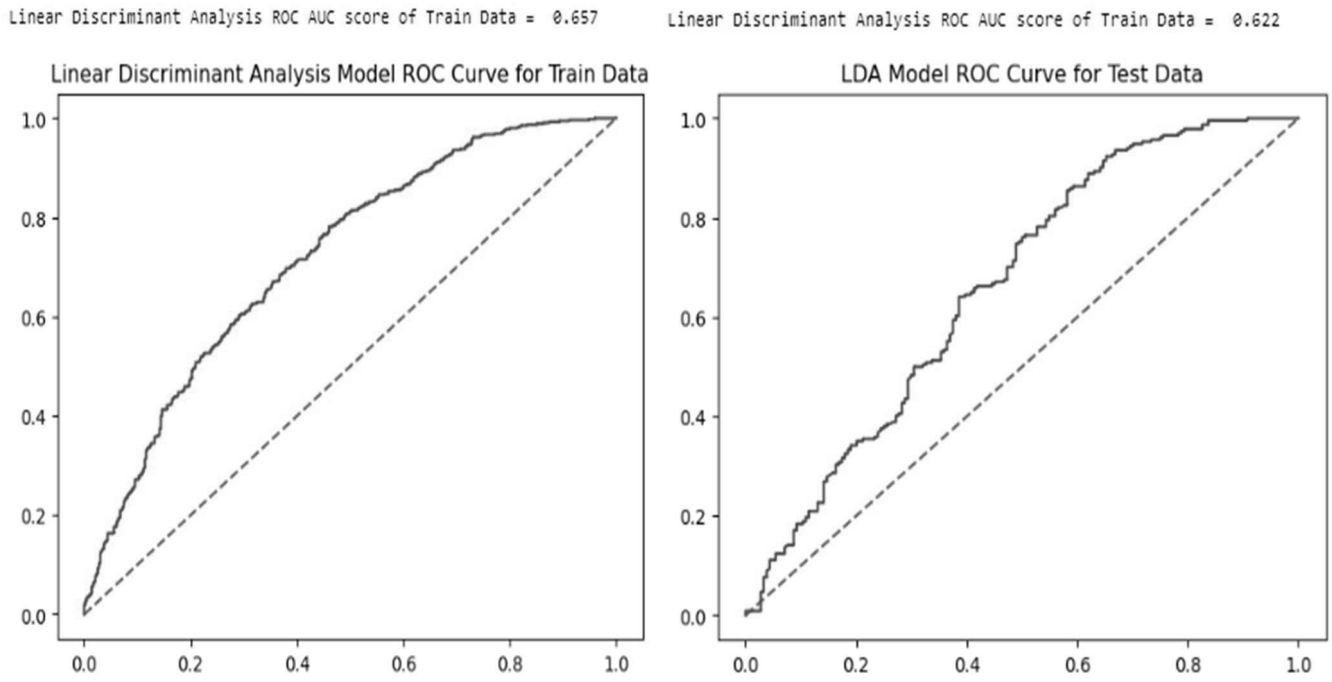


Fig. 2.3F
Linear Discriminant Analysis ROC AUC Score and ROC Curve for Train and Test Data

3. CART Model Classification Report / Confusion Matrix / ROC AUC Score/ROC Curve

Train Data

CART Model Classification Report on train data

	precision	recall	f1-score	support
0	0.70	0.56	0.62	430
1	0.70	0.81	0.75	545
accuracy			0.70	975
macro avg	0.70	0.68	0.69	975
weighted avg	0.70	0.70	0.69	975

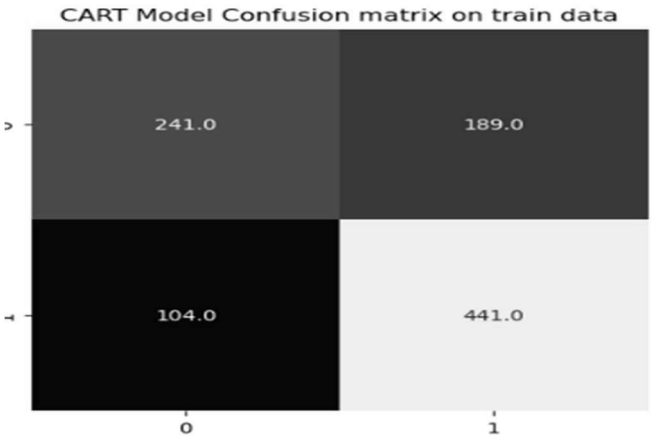


Fig 2.3G
CART Classification Report and Confusion matrix (Train Data)

Test Data

CART Model Classification Report on test data

	precision	recall	f1-score	support
0	0.66	0.47	0.55	184
1	0.66	0.81	0.73	234
accuracy			0.66	418
macro avg	0.66	0.64	0.64	418
weighted avg	0.66	0.66	0.65	418

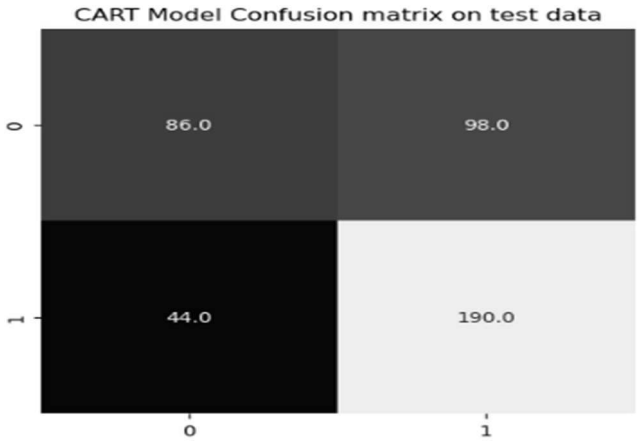


Fig 2.3H
CART Classification Report and Confusion matrix (Test Data)

Inferences from Fig. 2.3G & Fig. 2.3H: -

- For Train Data out of 545 number of 1's(Yes), 441 were actually predicted as 1's (Yes)
- For Test Data out of 234 number of 1's(Yes), 190 were actually predicted as 1's (Yes)
- Recall for 1 (Yes) for train data = 0.81 | Recall for 1 (Yes) for test data = 0.81
- Precision for 1(Yes) for train data = 0.70 | Precision for 1(Yes) test data are = 0.66
- F1 score for 1 (Yes) for train data = 0.75 | F1 score for 1 (Yes) for test data = 0.73
- Accuracy of train data = 0.70 | Accuracy of test data = 0.66

 # CART Model ROC AUC Score and ROC Curve

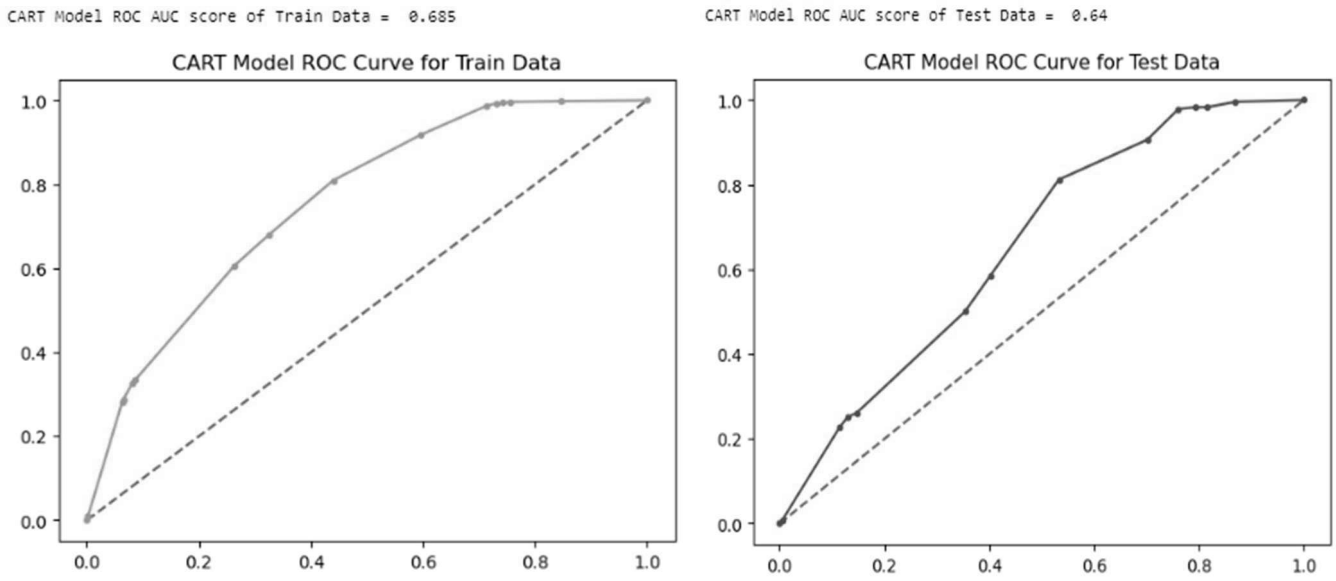


Fig. 2.3I
CART Model ROC AUC Score and ROC Curve for Train and Test Data

Lets Create a Comparision Chart between all 3 different models i.e., Linear Reg. LDA, CART

Comparison Between Different Classification Model						
Performance Parameters	**Logistic Reg. Train Data**	**Logistic Reg. Test Data**	**LDA Train Data**	**LDA Test Data**	**CART Train Data**	**CART Test Data**
Accuracy	**0.67**	**0.65**	**0.67**	**0.64**	**0.7**	**0.66**
Recall	**0.79**	**0.79**	**0.81**	**0.8**	**0.81**	**0.81**
Precision	**0.68**	**0.65**	**0.68**	**0.65**	**0.7**	**0.66**
F1 Score	**0.73**	**0.72**	**0.73**	**0.72**	**0.75**	**0.73**
ROC AUC Score	**0.654**	**0.626**	**0.657**	**0.622**	**0.685**	**0.64**

Table 2.3A
Comparison Between Different Classification Model

From above chart we can see that the model performace parameters of CART model for both train and test data are better than logistic regression model and LDA Model.

Hence, Best and Optimized model for the given dataset is CART Model

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

Solution: -

- In this dataset we have 10 variables and 1473 rows. Out of these 10 variable, 3 variables are numeric variable whereas 7 are categorical variable.
- While performing Feature Engineering where we found that the dataset has 80 duplicated rows (which were dropped), very few outliers were present in only 1 variable i.e., Number of Children Born (we had treated the outlier using IQR method), also there were null entries in 2 columns i.e., Wife Age and Number of children born (which were imputed with mean values of their columns)
- Then we perform several visualization techniques on the variables/columns/features
- While performing Univariate Analysis we find: - maximum number of wife's age are in between 25 to 35, maximum number of children born are either 1 or 2, most husbands are having occupation of Type 3, most of the wife's and husbands are having education level of Tertiary, most of the family are having very high standard of living, most the family are exposed to media, most wife's have used contraceptive methods
- While performing Bivariate Analysis between Independent Variables: -
    - Maximum percentage of woman who have used contraceptive method are of Age 33 and who doesn't used contraceptive method are of age 47
    - Most number of wife and husband who used contraceptive method are having education level of tertiary whereas who doesn't used contraceptive method are having education level of primary and uneducated resp.
    - Most number of children born even after using contraceptive method = 3 whereas who doesn't used contraceptive method = 1
    - Non-working wife's have mostly used contraceptive method
    - Wife's religion doesn't affect much about the contraceptive method used
- As we completed EDA then we have separated our dependent and Independent Variables.
- After separation of independent and dependent variable we have split our dataset into train and test data in 70:30 ratio

- **Steps performed in Logistic Regression**
    - We created a model using Simple Logistic Reg. and Logistic Reg. using GridSearchCV and found that both model output is slightly similar.
    - After creating a model, we calculated prediction probabilities for occurrence of 0 and 1 on both training and test data
    - Later we calculated model performance using confusion matrix and classification report for both Train and test data
    - ROC AUC Score and ROC Curve of train and test data was also calculated for checking model performance.