Business Report

Index-

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.

Let’s check out the head of the Data-

	Wife_age	Wife_ education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_exposure	Contraceptive_method_used
0	24.0	Primary	Secondary	3.0	Scientology	No	2	High	Exposed	No
1	45.0	Uneducated	Secondary	10.0	Scientology	No	3	Very High	Exposed	No
2	43.0	Primary	Secondary	7.0	Scientology	No	3	Very High	Exposed	No
3	42.0	Secondary	Primary	9.0	Scientology	No	3	High	Exposed	No
4	36.0	Secondary	Secondary	8.0	Scientology	No	3	Low	Exposed	No

Columns are-

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=very-low, 2, 3, 4=high
9. Media exposure (binary) Good, Not good
10. Contraceptive method used (class attribute) No, Yes

Let’s check the shape of the Data-

We have 1473 rows and 10 columns in the Data set

There should be 2 numerical and 8 categorical columns as per column description mentioned above.

Husband occupation should be categorical so we will change it when we proceed further.

```
 #   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)
```

Let’s check for 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
```

We have imputed the missing values with mean of data, as both of them are numerical columns-

After imputation –

```
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
```

Let’s check for duplicate values-

There are 85 duplicate values found which are removed from the dataset-

Let’s check the 5 point summary of Data and check out the categorical columns too-

	Wife_age	No_of_children_born	Husband_Occupation
count	1388.000000	1388.000000	1388.000000
mean	32.533862	3.287464	2.177954
std	8.102151	2.385715	0.853782
min	16.000000	0.000000	1.000000
25%	26.000000	1.000000	1.000000
50%	32.000000	3.000000	2.000000
75%	38.000000	5.000000	3.000000
max	49.000000	16.000000	4.000000

Let’s check for categorical data –

```
Wife_ education
Tertiary      510
Secondary     398
Primary       330
Uneducated    150
Name: Wife_ education, dtype: int64
```

```
Husband_education
Tertiary      822
Secondary     347
Primary       175
Uneducated     44
Name: Husband_education, dtype: int64
```

```
Wife_religion
Scientology        1182
Non-Scientology     206
Name: Wife_religion, dtype: int64
```

```
Wife_Working
No      1040
Yes      348
Name: Wife_Working, dtype: int64
```

```
Husband_Occupation
3    570
2    414
1    377
4     27
Name: Husband_Occupation, dtype: int64
```

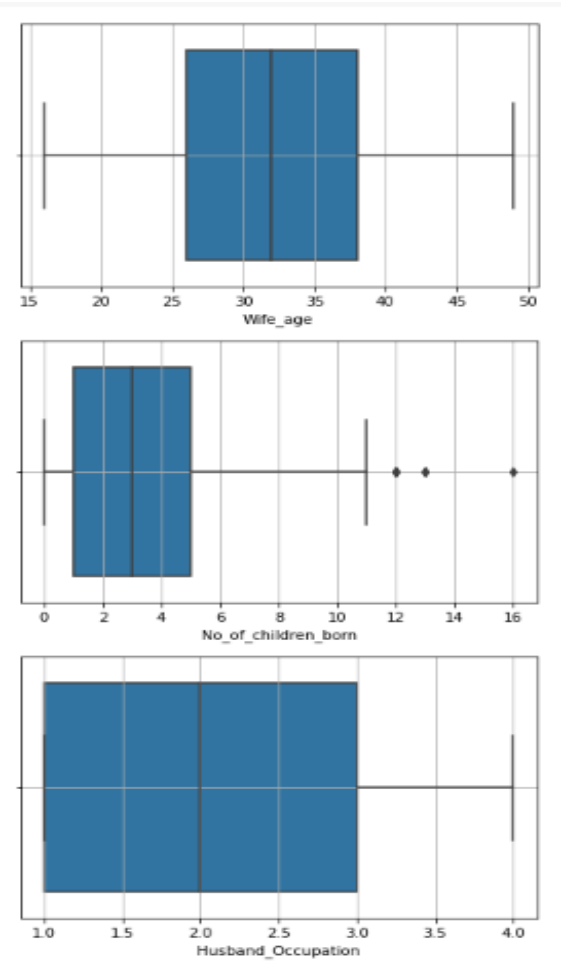
```
Standard_of_living_index
Very High    613
High         419
Low          227
Very Low     129
Name: Standard_of_living_index, dtype: int64
```

```
Media_exposure
Exposed        1279
Not-Exposed     109
Name: Media_exposure , dtype: int64
```

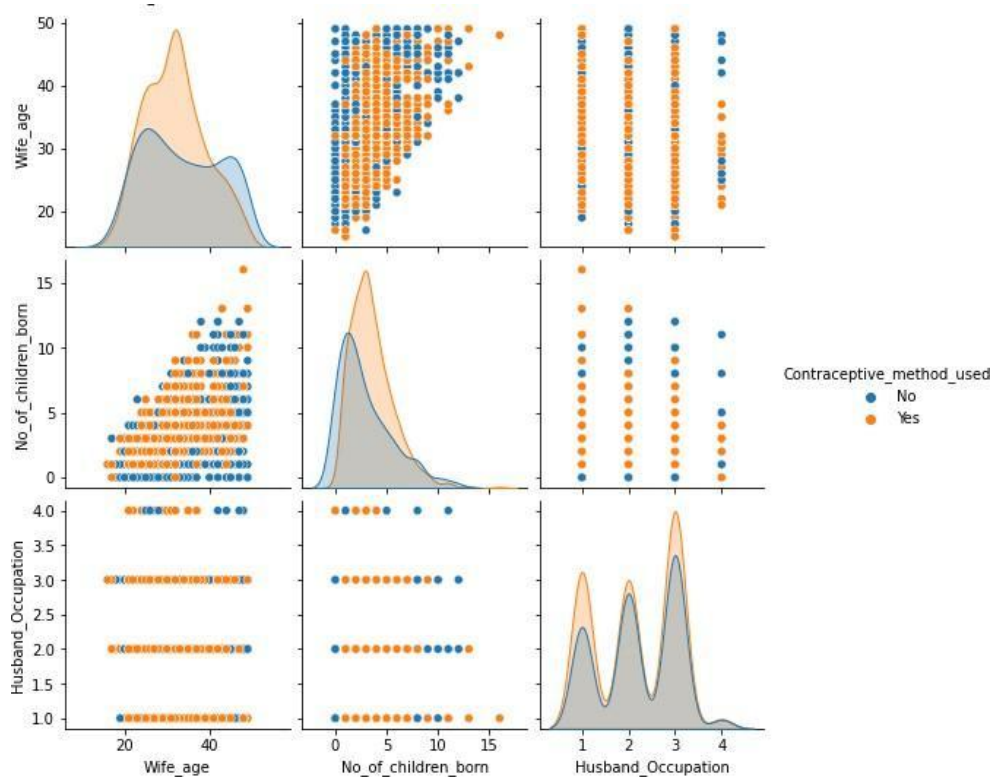
```
Contraceptive_method_used
Yes    774
No     614
Name: Contraceptive_method_used, dtype: int64
```

For target feature ‘Contraceptive-method-used’ data set is quite balanced.

Lets check forunivariate , bivariate and multivariate analysis-

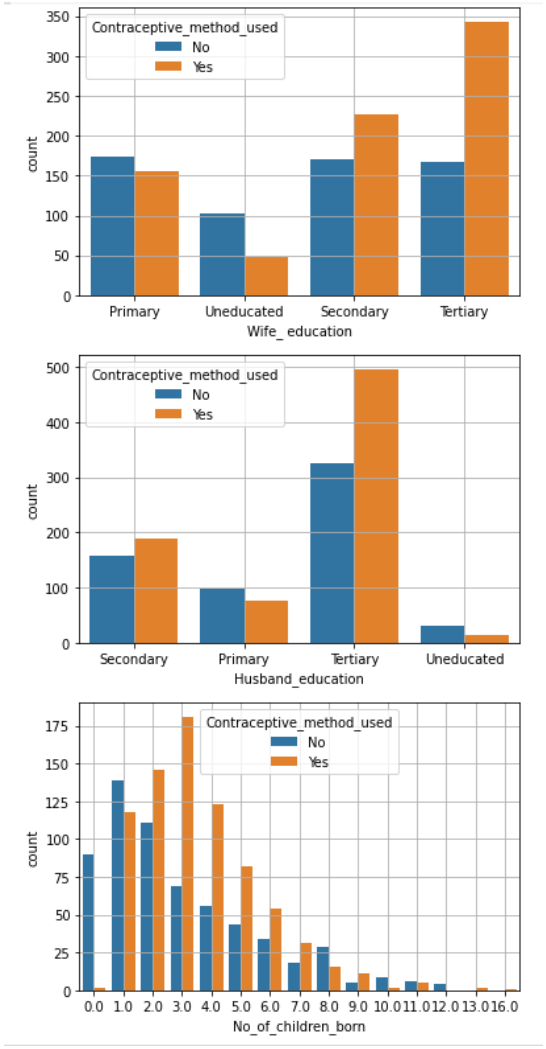


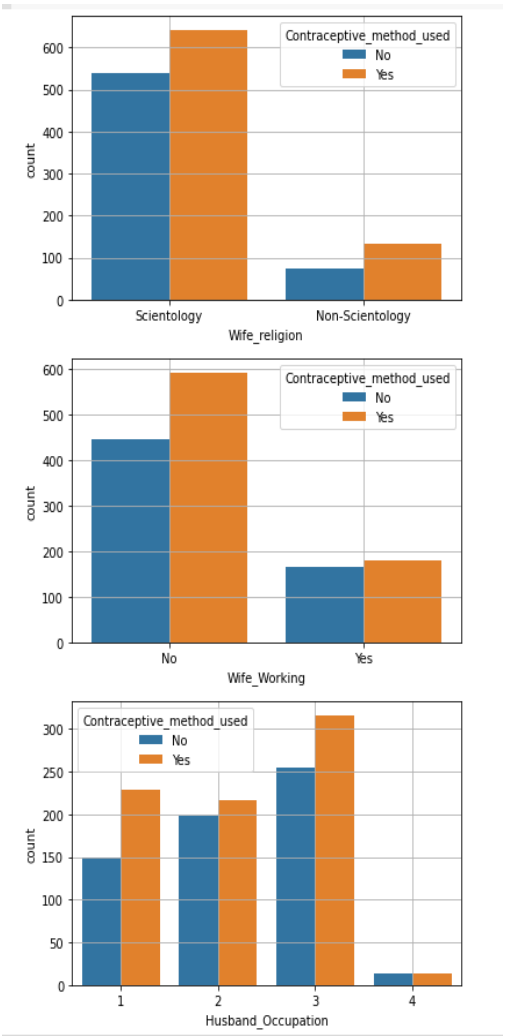
There are few outliers in column number of children born. There are very rare chances of having more than 10 or 11 children. If we even assume some of the data from very old ladies.

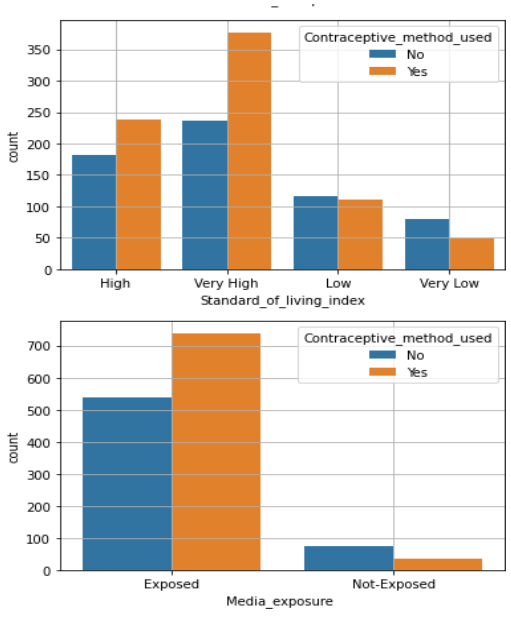


Number of children born and wife age could be a good differentiator between contraceptive method used or not.

Let’s check categorical columns:







Use of Contraceptive methods increases as wife's and husbands education increases, It also increases as standard of living index, the number of children born and media exposure increases.

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.

Let's encode the data

So according to the description of the data, we are encoding some of columns to ordinal data type and other to one-hot encoding.

Conversion on columns –

Ordinal Labeling – Wife education , husband education, standard of living, and contraceptive methods used.

One-Hot encoding – Wife religion, wife working, Husband occupation and media exposure.

After encoding following is the data –

	0	1	2	3	4
Wife_age	24.0	45.0	43.0	42.0	36.0
Wife_ education	1.0	0.0	1.0	2.0	2.0
Husband_education	2.0	2.0	2.0	1.0	2.0
No_of_children_born	3.0	10.0	7.0	9.0	8.0
Standard_of_living_index	2.0	2.0	2.0	1.0	2.0
Contraceptive_method_used	0.0	0.0	0.0	0.0	0.0
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
Husband_Occupation_2	1.0	0.0	0.0	0.0	0.0
Husband_Occupation_3	0.0	1.0	1.0	1.0	1.0
Husband_Occupation_4	0.0	0.0	0.0	0.0	0.0
Media_exposure _Not-Exposed	0.0	0.0	0.0	0.0	0.0

Then we split the Data (70:30) and converted the dataset into X_train and y_train with stratified sampling. Then we apply all the three algorithms mentioned (Logistic Regression and LDA (linear discriminant analysis) and CART.)

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.

Let's Check Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model.

Logistic Regression-

Train accuracy scores- 0.67

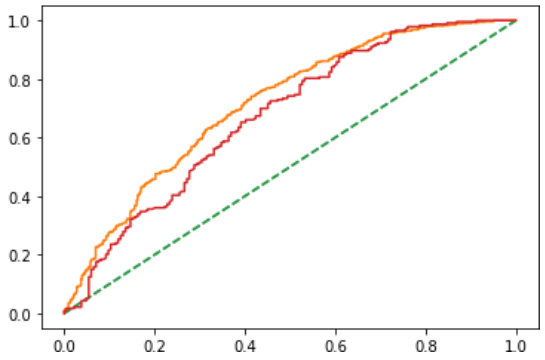
	precision	recall	f1-score	support
0	0.66	0.51	0.58	430
1	0.67	0.79	0.73	541
accuracy			0.67	971
macro avg	0.67	0.65	0.65	971
weighted avg	0.67	0.67	0.66	971

Test accuracy scores-0.65

	precision	recall	f1-score	support
0	0.65	0.47	0.54	184
1	0.66	0.80	0.72	233
accuracy			0.65	417
macro avg	0.65	0.63	0.63	417
weighted avg	0.65	0.65	0.64	417

Let's take a look at ROC curve and ROC_AUC score-

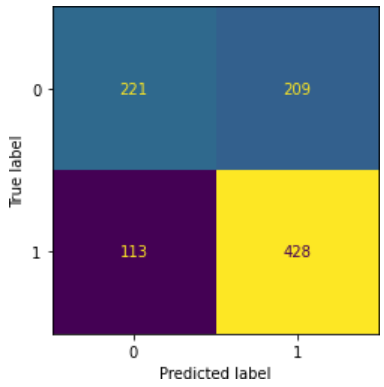
```
Train_AUC: 0.714
Test_AUC: 0.674
```

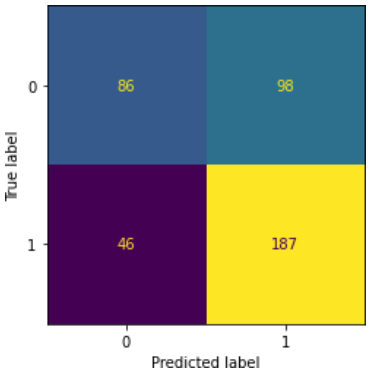


Let's take a look at confusion matrix for both train and test set-

Train set

Test set





LDA (linear discriminant analysis)

Classification Report of the training data:

	precision	recall	f1-score	support
0	0.66	0.50	0.57	430
1	0.67	0.80	0.73	541
accuracy			0.67	971
macro avg	0.67	0.65	0.65	971
weighted avg	0.67	0.67	0.66	971

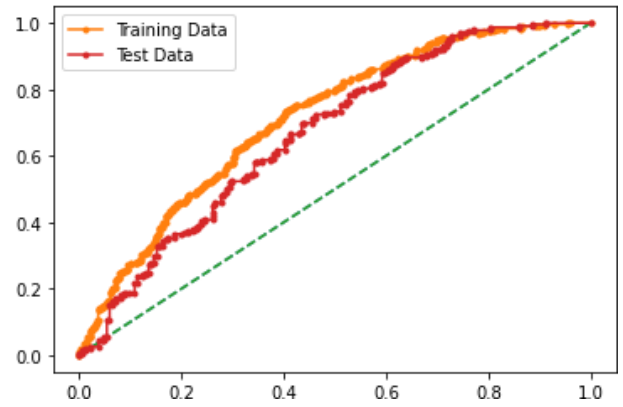
Classification Report of the test data:

	precision	recall	f1-score	support
0	0.63	0.43	0.52	184
1	0.64	0.80	0.71	233
accuracy			0.64	417
macro avg	0.64	0.62	0.61	417
weighted avg	0.64	0.64	0.63	417

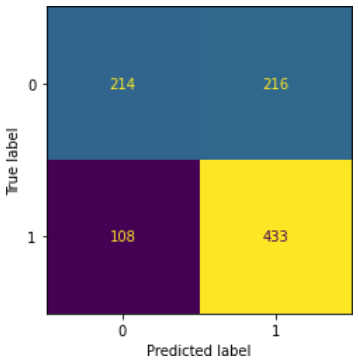
Train accuracy – 0.67

Test accuracy – 0.64

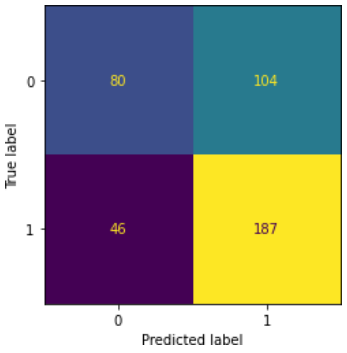
```
AUC for the Training Data: 0.714
AUC for the Test Data: 0.671
```



Train set



Test Set



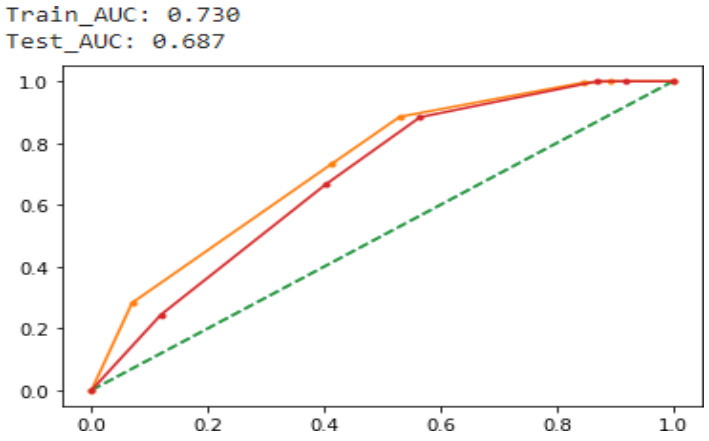
Let's check for CART-

	precision	recall	f1-score	support
0	0.77	0.47	0.58	430
1	0.68	0.89	0.77	541
accuracy			0.70	971
macro avg	0.72	0.68	0.67	971
weighted avg	0.72	0.70	0.69	971

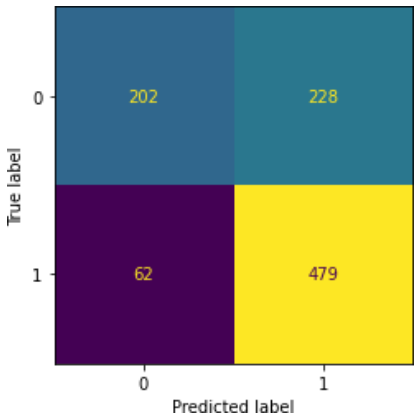
	precision	recall	f1-score	support
0	0.75	0.43	0.55	184
1	0.66	0.88	0.76	233
accuracy			0.69	417
macro avg	0.71	0.66	0.65	417
weighted avg	0.70	0.69	0.67	417

Train accuracy – 0.70

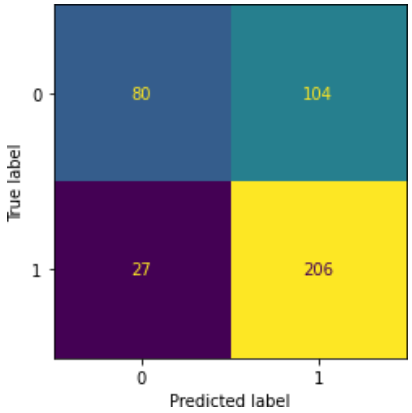
Test accuracy – 0.69



Train Set



Test Set



It could be clearly seen that CART has performed exceptionally well as compared to Logistic regression and LDA.

Model	Train Accuracy	Test Accuracy
Logistic regression	**0.67**	**0.65**
LDA	**0.67**	**0.64**
CART	**0.70**	**0.69**

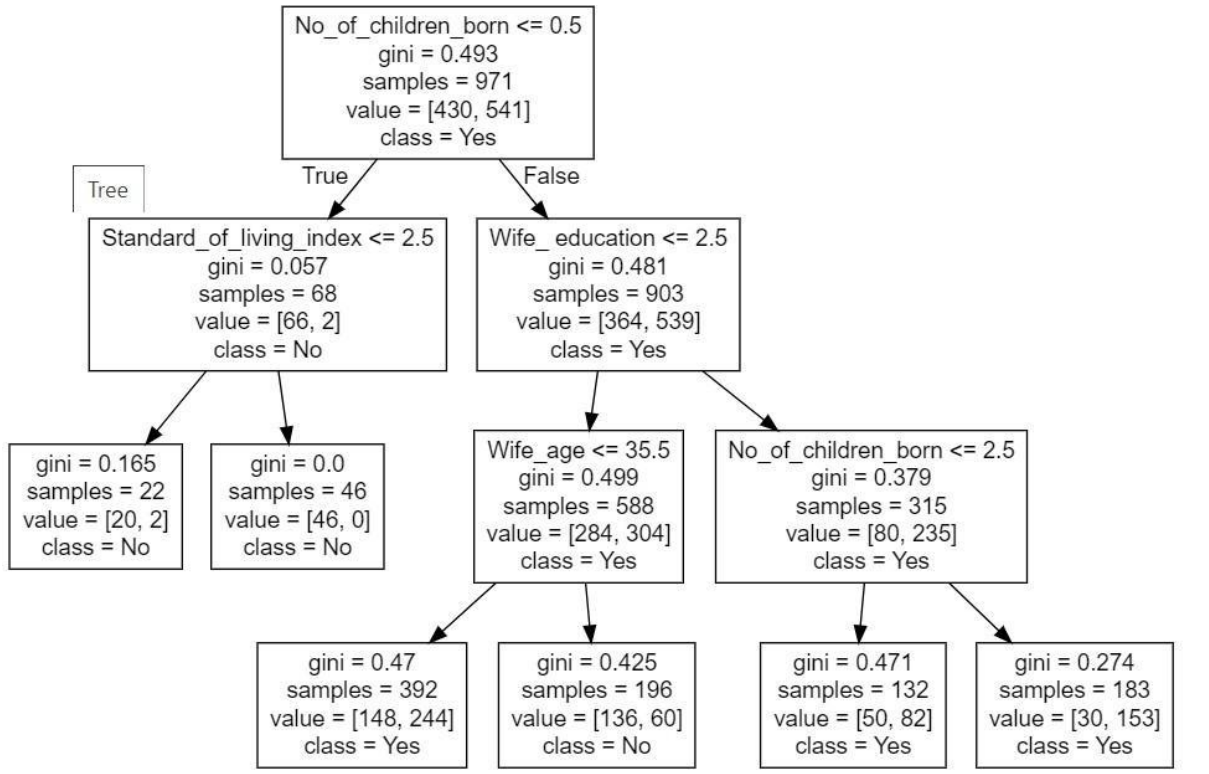
It could be clearly seen that there is no overfitting issue with CART and AUC score is also better as compared to LDA and LG.

Let's see for CART which features contributed the most-

	Imp
No_of_children_born	0.499463
Wife_age	0.273178
Wife_ education	0.224789
Standard_of_living_index	0.002570
Husband_education	0.000000
Wife_religion_Scientology	0.000000
Wife_Working_Yes	0.000000
Husband_Occupation_2	0.000000
Husband_Occupation_3	0.000000
Husband_Occupation_4	0.000000
Media_exposure _Not-Exposed	0.000000

No of children born, wife's age and wife's education are the most important 3 features among all.

Let's now check for the decision tree we used-



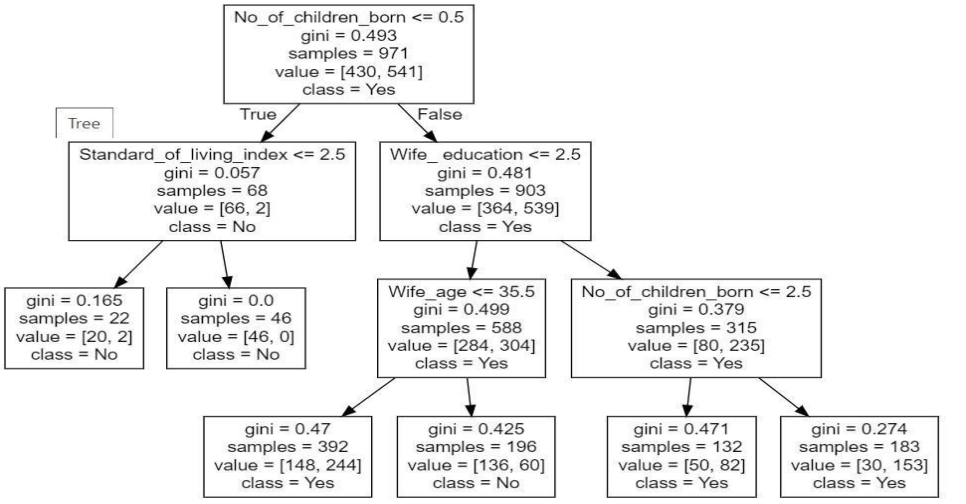
2.4 Inference: Basis on these predictions, what are the insights and recommendations.

Various steps followed in this model are-

We initially check the data and tried to understand the variables after that we checked for missing values and removed them after that we checked for duplicates and removed them too. Then we checked for outliers we found few outliers for No. of children born, then we check out the relation of all other categorical variables according to the target variable and check what features have most impact on target variable.

Then we encode the data according to the need either ordinal or with one-hot encoding.
Once encoding is done we split the data into train and test sets.
After that we applied all the given modules and tried to optimize them by using hyper-parameters and we found that CART is performing best out of all of the models.

Then we plot the decision tree and the confusion matrix to understand its business case and how it can be used.



Yes – is for women who has used contraceptive, No- is for women who has not used contraceptive-

The flow of DT –

If we have No. of child born less than 0.5, Then they have not used any contraceptive.

If No. of children born are greater than 0.5,

Then it will check with wife education,

if wife education is less than or equal to 2.5 it will check for wife's age , if age is less than 35.5 , then yes they have used a contraceptive method , if wife's age is greater than 35.5 then it they have not used a contraceptive method.

If wife’s education is greater than 2.5 then, we will check for no. of children born if it is greater than 0.5 then then yes they have used a contraceptive method.

Suggestions-

To restrict the increasing population we should really look forward to improve wife’s education to at least a Tertiary education. This is even more critical for women’s above age 35.5, they need to be educated and informed that contraceptive methods are a must after having a child.

END