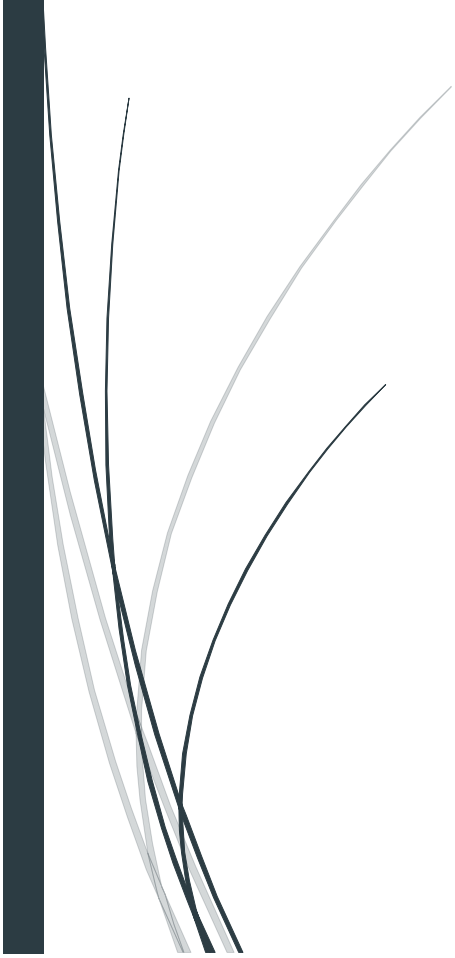
Business Report

Classification Problem Statement



Ayush Sharma

Table of Contents

A. 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.

Ans: Data Summary Data Info

	Wife_age	Wife_ education	Husband_education	No_of_children_born
count	1402.000000	1473	1473	1452.000000
unique	NaN	4	4	NaN
top	NaN	Tertiary	Tertiary	NaN
freq	NaN	577	899	NaN
mean	32.606277	NaN	NaN	3.254132
std	8.274927	NaN	NaN	2.365212
min	16.000000	NaN	NaN	0.000000
25%	26.000000	NaN	NaN	1.000000
50%	32.000000	NaN	NaN	3.000000
75%	39.000000	NaN	NaN	4.000000
max	49.000000	NaN	NaN	16.000000

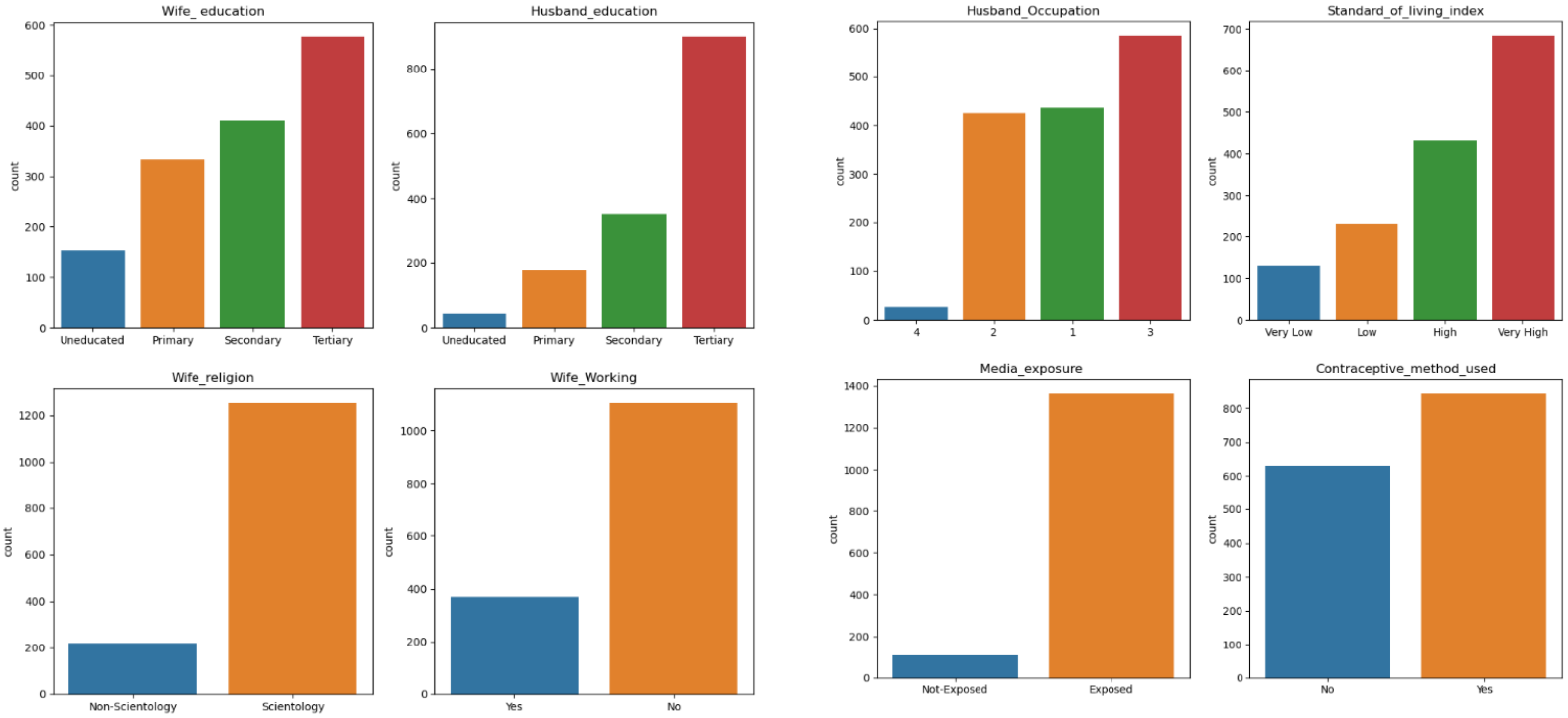
	Wife_religion	Wife_Working	Husband_Occupation	\
count	1473	1473	1473.000000	
unique	2	2	NaN	
top	Scientology	No	NaN	
freq	1253	1104	NaN	
mean	NaN	NaN	2.137814	
std	NaN	NaN	0.864857	
min	NaN	NaN	1.000000	
25%	NaN	NaN	1.000000	
50%	NaN	NaN	2.000000	
75%	NaN	NaN	3.000000	
max	NaN	NaN	4.000000	

	Standard_of_living_index	Media_exposure	Contraceptive_method_used
count	1473	1473	1473
unique	4	2	2
top	Very High	Exposed	Yes
freq	684	1364	844
mean	NaN	NaN	NaN
std	NaN	NaN	NaN
min	NaN	NaN	NaN
25%	NaN	NaN	NaN
50%	NaN	NaN	NaN
75%	NaN	NaN	NaN
max	NaN	NaN	NaN

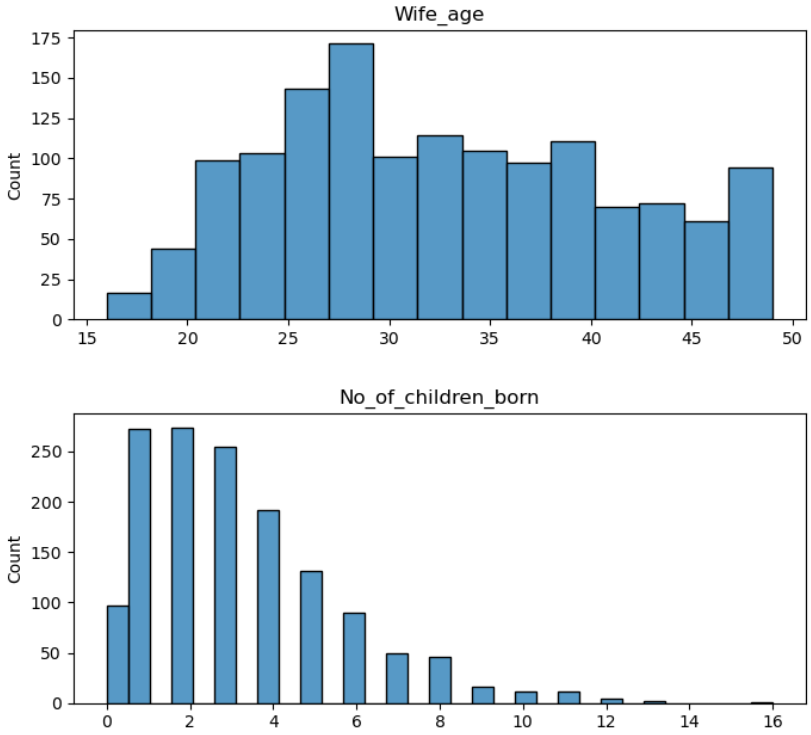
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1473 entries, 0 to 1472
Data columns (total 10 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Wife_age                   1402 non-null   float64
 1   Wife_ education            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   object
 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), object(8)
memory usage: 115.2+ KB
```

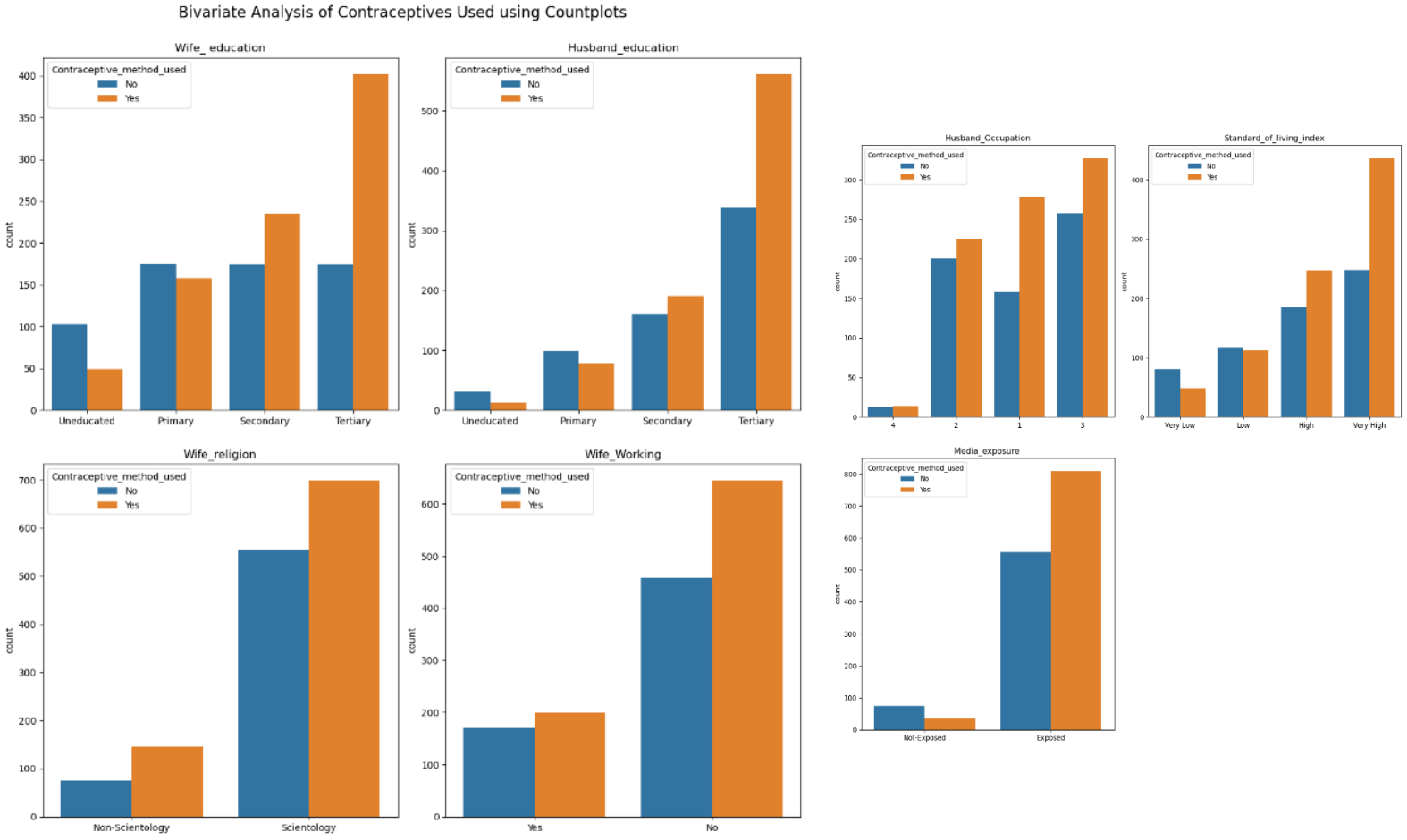
- The data consists of **1473 rows** and **10 columns**
- There is a total of **2 numeric columns** and **8 categoric columns**
- It can be observed from the data info that null values exist in the **Wife_age and No_of_children_born columns** of the dataset

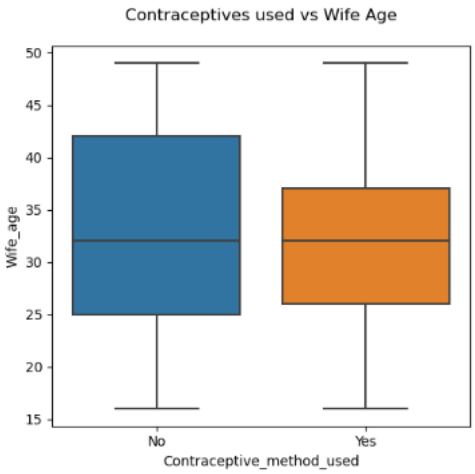
Univariate Analysis using Countplots

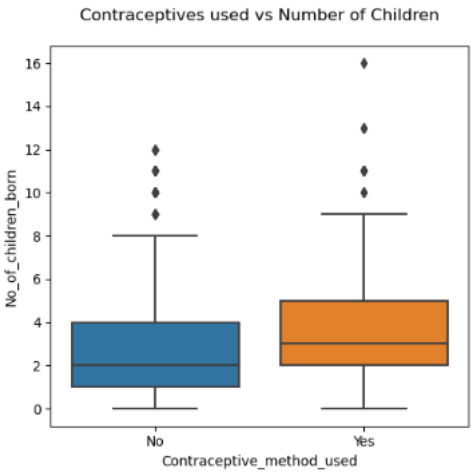


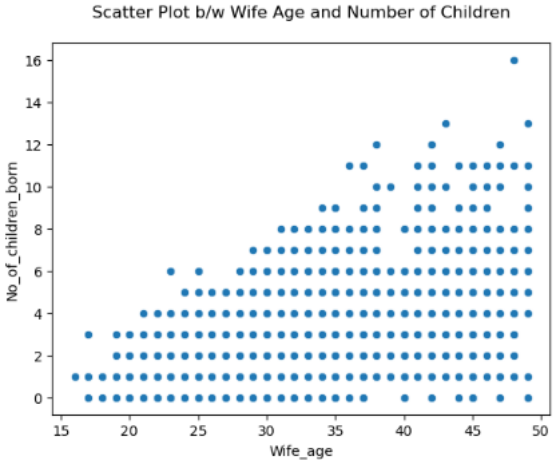
Univariate Analysis using Histograms




Bivariate Analysis of Contraceptives Used using Countplots
Wife_ education
Husband_education
Husband_Occupation
Standard_of_living_index
Wife_religion
Wife_Working
Media_exposure
Contraceptive_method_used
No
Yes
count
Uneducated
Primary
Secondary
Tertiary
Non-Scientology
Scientology
Very Low
Low
High
Very High
Not-Exposed
Exposed

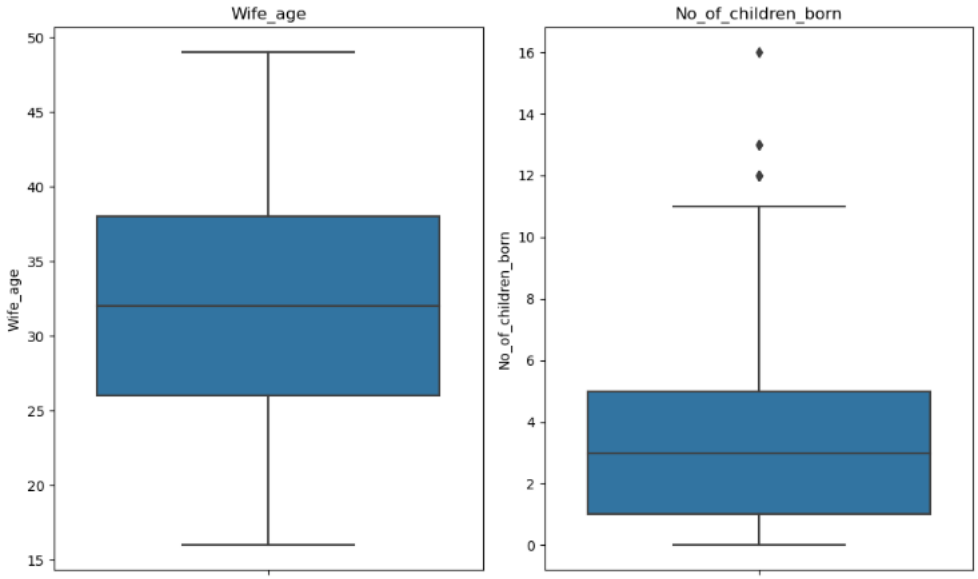

Contraceptives used vs Wife Age
Wife_age
No
Yes
Contraceptive_method_used


Contraceptives used vs Number of Children
No_of_children_born
No
Yes
Contraceptive_method_used


Scatter Plot b/w Wife Age and Number of Children
No_of_children_born
Wife_age

Inferences:

- The histogram for the number of children is right-skewed with the minimum count of children at 0 and maximum at 16.
- The histogram for the women's' age is slightly right-skewed with the minimum age value at 16 and maximum at 49.
- The median values for the women's' age and number of children born are 32 and 3 respectively.
- The dependent variable 'Contraceptive_method_used' is a binary variable with approximately 57% positive (Yes) values and 42% negative (No) values.
- Tertiary education dominates in both the wife and husband's education levels columns.
- Approximately 85% of the women follow scientology as their religion while the other 15% follow non-scientology.
- Approximately 25% of the women are working women while the other 75% are non-working.
- 46% of the women have very high standards of living and 29% of the women have high standards of living.
- A majority of the population (92%) has media exposure.



As per the boxplot, it can be observed that there are not many outliers in the dataset hence there is no need for outlier treatment for the dataset.

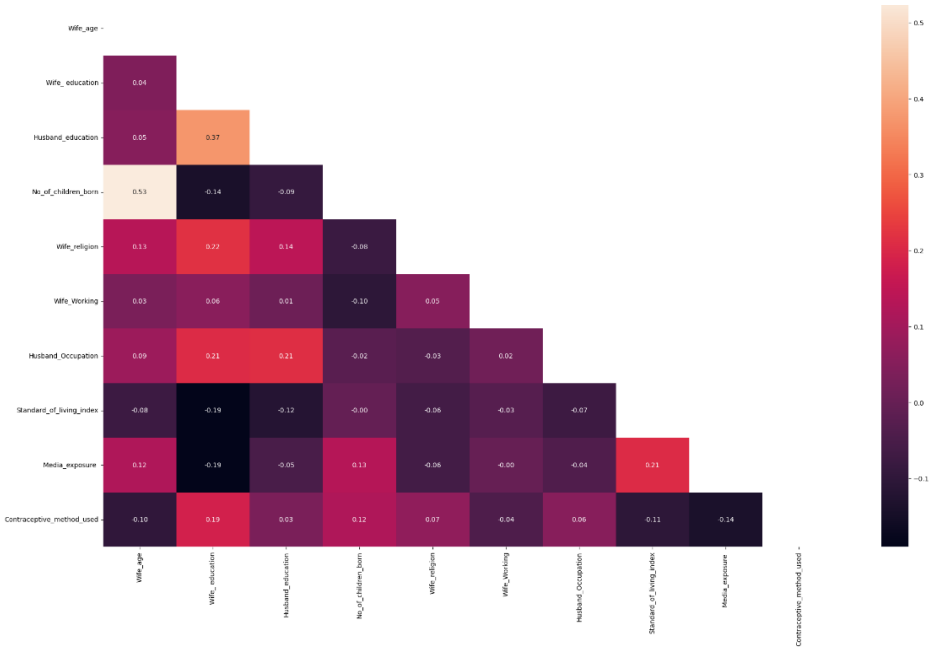
B. Do not scale the data. Encode the data (having string values) for Modelling. Data Split: Split the data into train and test (70:30). Apply Logistic Regression and LDA (linear discriminant analysis) and CART.

Ans: The data comprises of many categorical variables of different levels. A custom function has been created to assign numeric categories to these levels. The dependent variable has also been changed to numeric categories.

```
Value counts for Wife_ education:
4    510
3    398
1    330
2    150
Name: Wife_ education, dtype: int64


Value counts for Husband_education:
3    822
1    347
2    175
4     44
Name: Husband_education, dtype: int64


Value counts for Wife_religion:
1    1182
2     206
Name: Wife_religion, dtype: int64


Value counts for Wife_Working:
1    1040
2     348
Name: Wife_Working, dtype: int64


Value counts for Husband_Occupation:
2    570
1    414
3    377
4     27
Name: Husband_Occupation, dtype: int64


Value counts for Standard_of_living_index:
2    613
1    419
3    227
4    129
Name: Standard_of_living_index, dtype: int64


Value counts for Media_exposure :
1    1279
2     109
Name: Media_exposure , dtype: int64
```

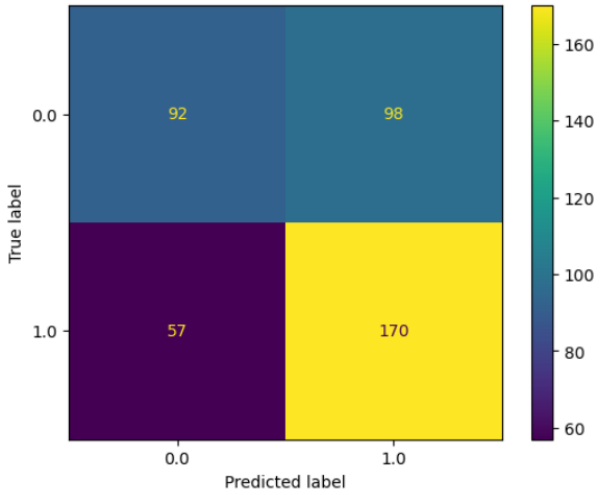
C. Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.

Ans:

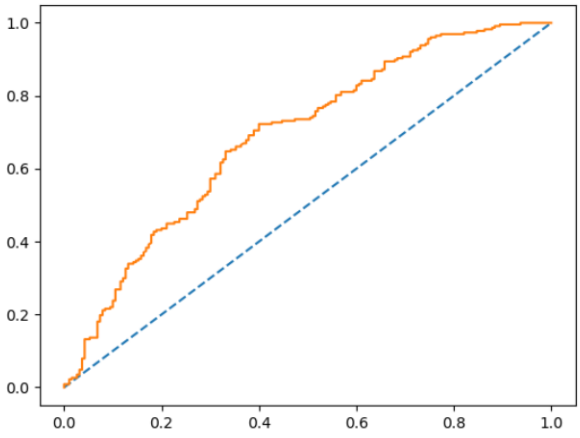


It can be observed from the correlation heatmap that there isn't high correlation amongst the variables. The highest correlation exists between the '*no_of_children_born*' and *'wife_age'* columns which need not be treated for now.

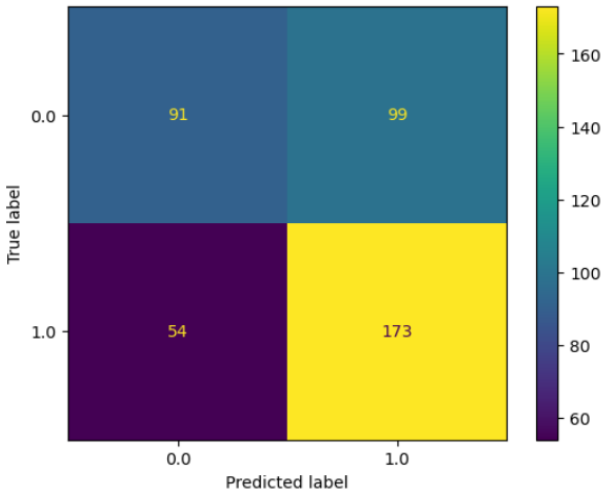
Confusion Matrix for the Logistic Regression testing datset:



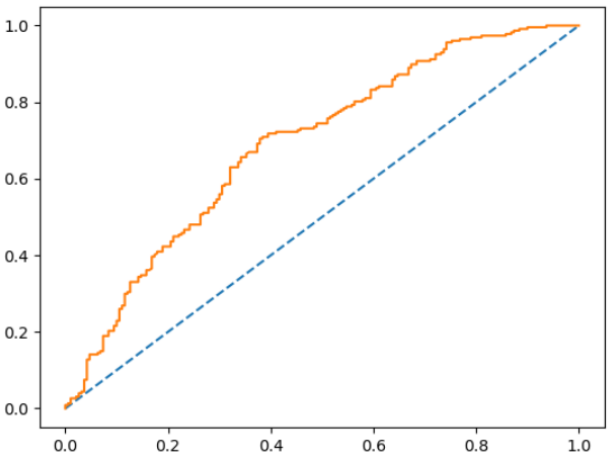
ROC curve for the Logistic Regression test datset:



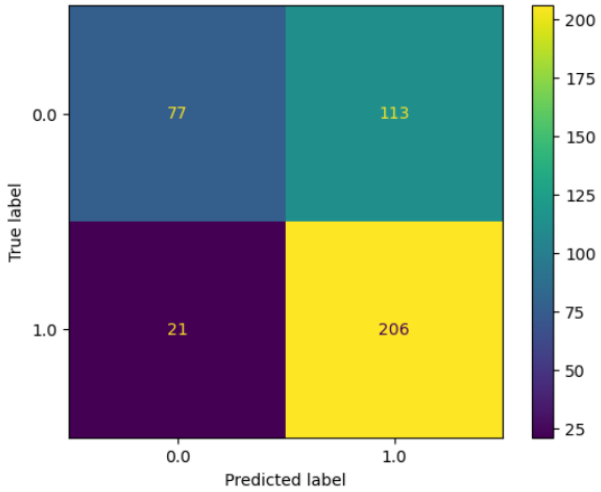
Confusion Matrix for the LDA test datset:



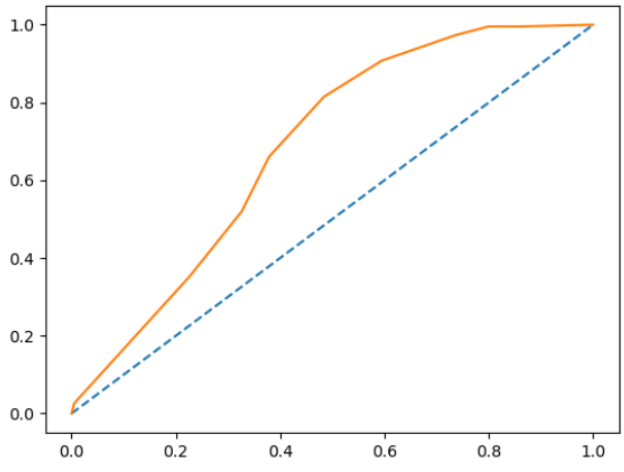
ROC curve for the LDA test datset:



Confusion Matrix for the CART test datset:



ROC curve for the CART test datset:



D. **Inference: Basis on these predictions, what are the insights and recommendations. Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.**

Ans:

Following inferences can be drawn from the **Logistic Regression model** for the testing data:

- The model predicted that 149 women did not use contraceptives while 268 women did use contraceptives while the actual values stand at 190 and 227 respectively.
- The model score and accuracy both stand at approximately 63%.
- True Positive -> 170 women who did use contraceptives (1) were predicted correctly (1) by the model.
- True Negative -> 92 women who didn't use any contraceptives (0) were predicted correctly (0) by the model.
- False Positive -> 98 women who didn't use any contraceptives (0) were predicted incorrectly (1) by the model.
- False Negative -> 57 women who did use contraceptives (1) were predicted incorrectly (0) by the model.
- The AUC Score for the model is approximately 70%.

Following inferences can be drawn from the LDA model for the testing data:

- The model predicted that 145 women did not use contraceptives while 272 women did use contraceptives while the actual values stand at 190 and 227 respectively.

- The model score and accuracy both stand at approximately 63%.

- True Positive -> 173 women who did use contraceptives (1) were predicted correctly (1) by the model.

- True Negative -> 91 women who didn't use any contraceptives (0) were predicted correctly (0) by the model.

- False Positive -> 99 women who didn't use any contraceptives (0) were predicted incorrectly (1) by the model.

- False Negative -> 54 women who did use contraceptives (1) were predicted incorrectly (0) by the model.

- The AUC Score for the model is approximately 69%.

Following inferences can be drawn from the CART model for the test data:

- The model predicted that 98 women did not use contraceptives while 319 women did use contraceptives while the actual values stand at 190 and 227 respectively.
- The model score and accuracy both stand at 68%
- True Positive -> 206 women who did use contraceptives (1) were predicted correctly (1) by the model.
- True Negative -> 77 women who didn't use any contraceptives (0) were predicted correctly (0) by the model.
- False Positive -> 113 women who didn't use any contraceptives (0) were predicted incorrectly (1) by the model.
- False Negative -> 21 women who did use contraceptives (1) were predicted incorrectly (0) by the model.
- The AUC Score for the model is approximately 69%.

It can thus be said that the CART model is the most efficient model out of the three models.