PREDICTIVE MODELLING

REGRESSION METHODS(LINEAR,LOGISTIC,LDA & CART)

LIST OF CONTENTS

LIST OF FIGURES

1. PROBLEM 1

The comp-active database comprises activity measures of computer systems. Data was gathered from a Sun Sparcstation 20/712 with 128 Mbytes of memory, operating in a multi-user university department. Users engaged in diverse tasks, such as internet access, file editing, and CPU-intensive programs.

Being an aspiring data scientist, you aim to establish a linear equation for predicting 'usr' (the percentage of time CPUs operate in user mode). Your goal is to analyze various system attributes to understand their influence on the system's 'usr' mode.

1.1) Perform Exploratory Data Analysis

	lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgscan	atch	pgin	ppgin	pflt	vflt	runqsz	freemem	freeswap
0	1	0	2147	79	68	0.2	0.2	40671.0	53995.0	0.0	...	0.0	0.0	1.6	2.6	16.00	26.40	CPU_Bound	4670	1730946
1	0	0	170	18	21	0.2	0.2	448.0	8385.0	0.0	...	0.0	0.0	0.0	0.0	15.63	16.83	Not_CPU_Bound	7278	1869002
2	15	3	2162	159	119	2.0	2.4	NaN	31950.0	0.0	...	0.0	1.2	6.0	9.4	150.20	220.20	Not_CPU_Bound	702	1021237
3	0	0	160	12	16	0.2	0.2	NaN	8670.0	0.0	...	0.0	0.0	0.2	0.2	15.60	16.80	Not_CPU_Bound	7248	1863704
4	5	1	330	39	38	0.4	0.4	NaN	12185.0	0.0	...	0.0	0.0	1.0	1.2	37.80	47.60	Not_CPU_Bound	633	1760253

Fig 1- Sample data

- There are 8192 rows and 22 columns in the dataset
- There are 21 numerical and 1 object datatypes.
- runsqz is an categorical data and we are converting it into a numerical datatype..

Data Description:

lread - Reads (transfers per second) between system memory and user memory
lwrite - writes (transfers per second) between system memory and user memory
scall - Number of system calls of all types per second
sread - Number of system read calls per second .
swrite - Number of system write calls per second .
fork - Number of system fork calls per second.
exec - Number of system exec calls per second.
rchar - Number of characters transferred per second by system read calls
wchar - Number of characters transfreed per second by system write calls
pgout - Number of page out requests per second
ppgout - Number of pages, paged out per second

pgfree - Number of pages per second placed on the free list.
pgscan - Number of pages checked if they can be freed per second
atch - Number of page attaches (satisfying a page fault by reclaiming a page in memory) per s
econd
pgin - Number of page-in requests per second
ppgin - Number of pages paged in per second
pflt - Number of page faults caused by protection errors (copy-on-writes).
vflt - Number of page faults caused by address translation .
runqsz - Process run queue size (The number of kernel threads in memory that are waiting for
a CPU to run.
Typically, this value should be less than 2. Consistently higher values mean that the system m
ight be CPU-bound.)
freemem - Number of memory pages available to user processes
freeswap - Number of disk blocks available for page swapping.
---
usr - Portion of time (%) that cpus run in user mode

1.1.1) Univariate and Bivariate Analysis

- No linear relationship is visible with independent variables.
- variables fork, pflt and vflt show slight linearity.
- The CPU running on user mode is either 0 or above 40% of the time.
- No linear relationships can be seen in the pairplot also.

1.2) Data Preprocessing

- Imputed the median values of rchar and wchar to their corresponding null values.
- runqsz is a categorical variable having values CPU bound and CPU not bound , the values of CPU bound will be changed to 3 and for CPU not bound we will keep the values to be 1.
- created dummy variables for the column runqsz to avoid misconception.
- Here runqsz_3 is for CPU bound and runqsz is for NOT CPU bound .

- There are no missing values in the dataset.
- Null values of rchar and wchar have been imputed with its corresponding median values.
- runqsz has been converted into numerical datatype and true or false dummy variables have been created for the same.

1.3) Model building- Linear Regression

- The dataset has been splitted into independent variables X and dependent variables y.

- Further the data has been split into X_train, y_train, X_test, y_test using sklearn
- Using linear regression from sklearn model was built and the data was fit into the model.
- Using statsmodel the regression summary was taken.

OLS Regression Results

Dep. Variable:	usr	R-squared:	0.640
Model:	OLS	Adj. R-squared:	0.639
Method:	Least Squares	F-statistic:	692.0
Date:	Fri, 16 Feb 2024	Prob (F-statistic):	0.00
Time:	17:33:09	Log-Likelihood:	-31296.
No. Observations:	8192	AIC:	6.264e+04
Df Residuals:	8170	BIC:	6.279e+04
Df Model:	21		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
lread	-0.0198	0.003	-7.088	0.000	-0.025	-0.014
lwrite	0.0074	0.005	1.522	0.128	-0.002	0.017
scall	0.0010	0.000	8.759	0.000	0.001	0.001
sread	-9.749e-05	0.002	-0.061	0.952	-0.003	0.003
swrite	-0.0016	0.002	-0.889	0.374	-0.005	0.002
fork	-1.8871	0.209	-9.039	0.000	-2.296	-1.478
exec	-0.0416	0.041	-1.013	0.311	-0.122	0.039
rchar	-3.657e-06	7.19e-07	-5.084	0.000	-5.07e-06	-2.25e-06
wchar	-1.066e-05	1.1e-06	-9.731	0.000	-1.28e-05	-8.51e-06
wchar	-1.066e-05	1.1e-06	-9.731	0.000	-1.28e-05	-8.51e-06
pgout	-0.2048	0.054	-3.799	0.000	-0.310	-0.099
ppgout	0.1270	0.031	4.109	0.000	0.066	0.188
pgfree	-0.0880	0.016	-5.553	0.000	-0.119	-0.057
pgscan	0.0125	0.005	2.644	0.008	0.003	0.022
atch	-0.0396	0.022	-1.773	0.076	-0.083	0.004
pgin	0.0581	0.024	2.383	0.017	0.010	0.106
ppgin	-0.0392	0.016	-2.507	0.012	-0.070	-0.009
pflt	-0.0401	0.004	-11.182	0.000	-0.047	-0.033
vflt	0.0228	0.003	8.215	0.000	0.017	0.028
freemem	-0.0017	6.37e-05	-26.141	0.000	-0.002	-0.002
freeswap	3.325e-05	3.82e-07	87.082	0.000	3.25e-05	3.4e-05
runqsz_1	50.7881	0.587	86.472	0.000	49.637	51.939
runqsz_3	42.8406	0.623	68.749	0.000	41.619	44.062

Omnibus:	1928.707	Durbin-Watson:	2.026
Prob(Omnibus):	0.000	Jarque-Bera (JB):	5359.778
Skew:	-1.244	Prob(JB):	0.00
Kurtosis:	6.084	Cond. No.	9.66e+06

Fig 2- Regression summary

From the above results summary:

- pvalues of variables having less than 0.05 can be considered as significant variables.
- lread,lwrite,scall,fork,rchar,wchar,pgout,ppgout,pgfree,pgscan,pgin,ppgin,pflt,vflt,freemem,freeswap,runqsz_1,runqsz_3 are the significant variables.

Linear regression on train and test:

- Finding R-squared value, RMSE value and Adjusted R squared value on both train and test splits.

```
Training Set:
R-squared: 0.6428396267060906
RMSE: 10.813213974052196
Adjusted R-squared: 0.6414637681502395
```

- R-squared value is 0.64 which means 64% of the variance of the dependent variables is explained by the independent variables.
- The RMSE value is 10.81 which means there is an average difference of 10.81 from the actual and predicted variables.
- Adjusted R square value is 0.64 which is similar to the r squared value hence the model does not suffer from unnecessary overfitting.
- It means that the predictor variables are relevant in the model.

```
Test Set:
R-squared: 0.631217100611971
RMSE: 11.594013992326564
Adjusted R-squared: 0.6278851811924487
```

- R-squared value is 0.63 which means 63% of the variance of the dependent variables is explained by the independent variables.
- The RMSE value is 11.59 which means there is an average difference of 11.59 from the actual and predicted variables.
- Adjusted R square value is 0.62 which is similar to the r squared value hence the model does not suffer from unnecessary overfitting.
- It means that the predictor variables are relevant in the model.

Linear equation of the final model:

```
Linear Equation: y = 46.81 + -0.02 * lread + 0.01 * lwrite + 0.00 * sca
ll + -0.00 * sread + -0.00 * swrite + -1.89 * fork + -0.04 * exec + -0.
00 * rchar + -0.00 * wchar + -0.20 * pgout + 0.13 * ppgout + -0.09 * pg
free + 0.01 * pgscan + -0.04 * atch + 0.06 * pgin + -0.04 * ppgin + -0.
04 * pflt + 0.02 * vflt + -0.00 * freemem + 0.00 * freeswap + 3.97 * ru
nqsz_1 + -3.97 * runqsz_3
```

- From the above equation we can say that variables 'fork','pgout','ppgout','runqsz_1','runqsz_3' has most significance in the equation.

1.4) Business insights and Recommendations

- The models performance is consistent when compared to the train and test split data.
- Both the model captures a moderate amount of variance of the dependent variable about 64%.
- The level of error is also moderate in the sets.
- Hence the model generalises well to unseen data.
- Fork is the process which creates a new space for a new task, basically a process which is used while in multitasking hence multitasking and initiating new process can increase the fork hence usr mode.
- pgout and ppgout are number of page out requests and number of pages paged out request which enables the system to run in user mode.
- runqsz is also similar to fork where it is a queue that is waiting for a CPU to run.

Hence the process which are asked by the user to be done by the computer will make the CPU run more on User bound while other parameters like kernal base codes or being idle will not come under user bound time.¶

2. PROBLEM 2

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

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

Data Description:

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

2.1) Define the problem and perform Exploratory Data Analysis

- The Data set has 1473 rows and 10 columns.
- There are 3 numerical and 7 object datatypes.
- Before proceeding for the regression the object categorical variables need to be converted to numerical or should be encoded.
- There are null values in 2 columns.

	Wife_age	No_of_children_born	Husband_Occupation
count	1402.000000	1452.000000	1473.000000
mean	32.606277	3.254132	2.137814
std	8.274927	2.365212	0.864857
min	16.000000	0.000000	1.000000
25%	26.000000	1.000000	1.000000
50%	32.000000	3.000000	2.000000
75%	39.000000	4.000000	3.000000
max	49.000000	16.000000	4.000000

Fig 3- Statistical summary

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

Husband_Occupation	Standard_of_living_index	Media_exposure	Contraceptive_method_used
2	High	Exposed	No
3	Very High	Exposed	No
3	Very High	Exposed	No
3	High	Exposed	No
3	Low	Exposed	No

Fig 4- Sample data 2

2.1.1) Univariate and Bivariate Analysis

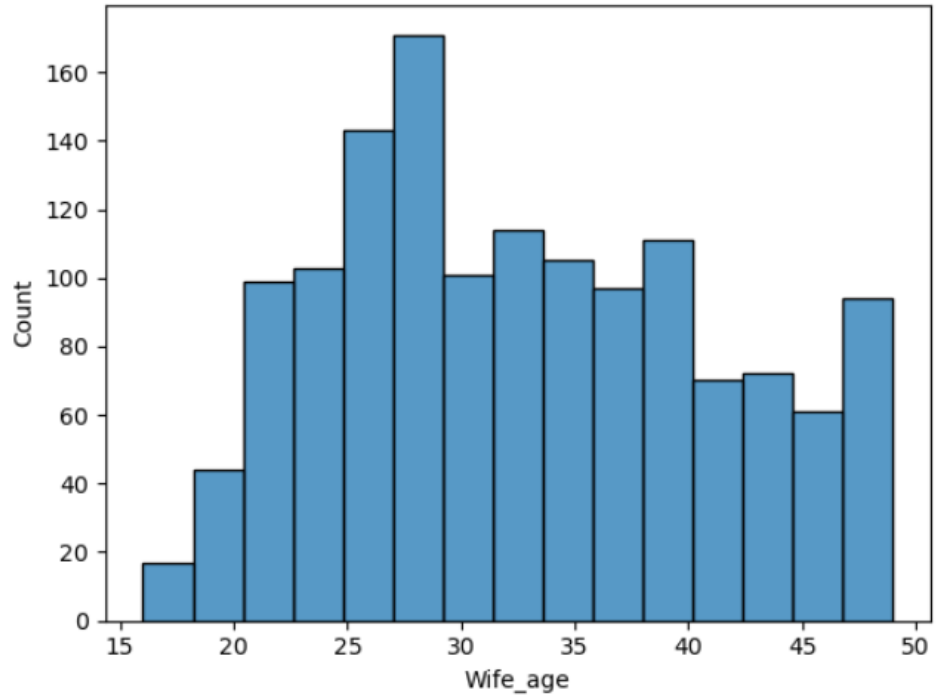

Fig 5- Wife age histogram

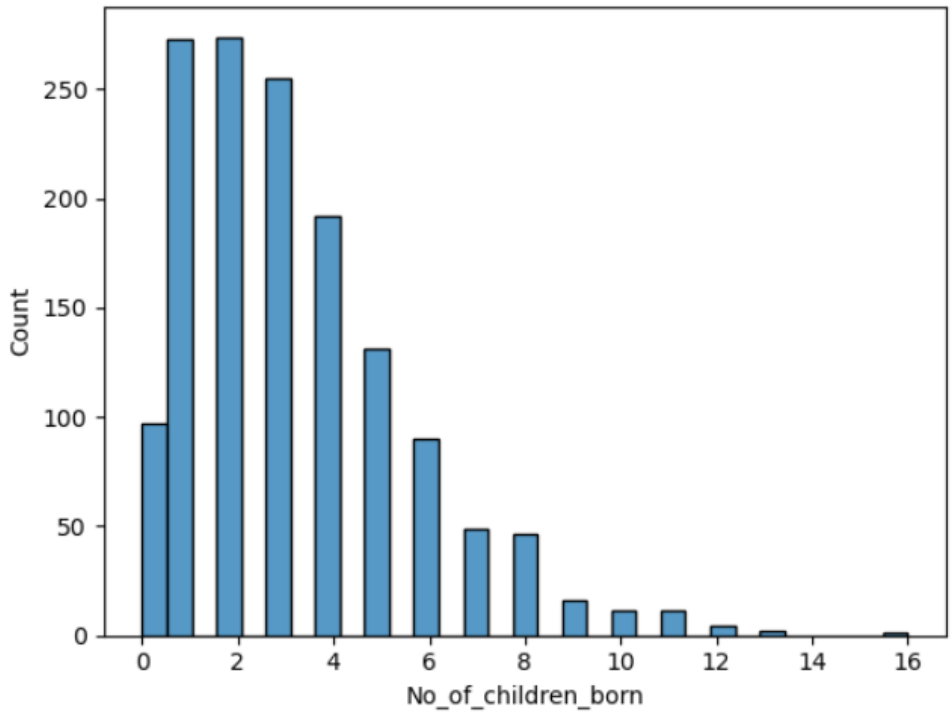

Fig 6- No of children born histogram

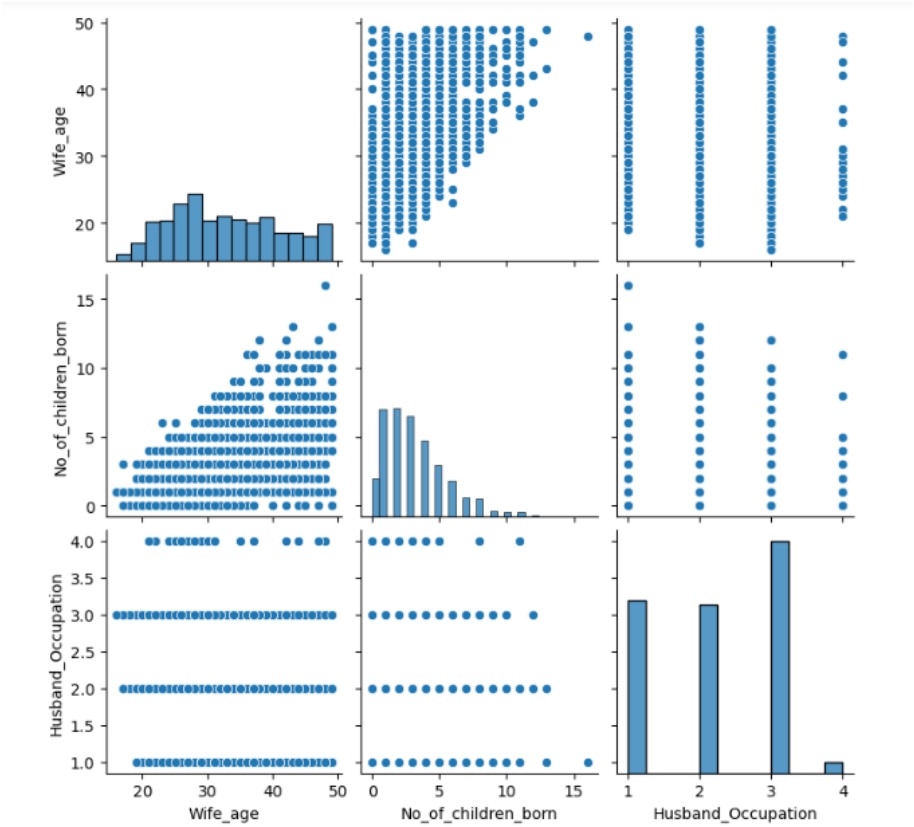

Fig 7- Pairplot between numerical variable

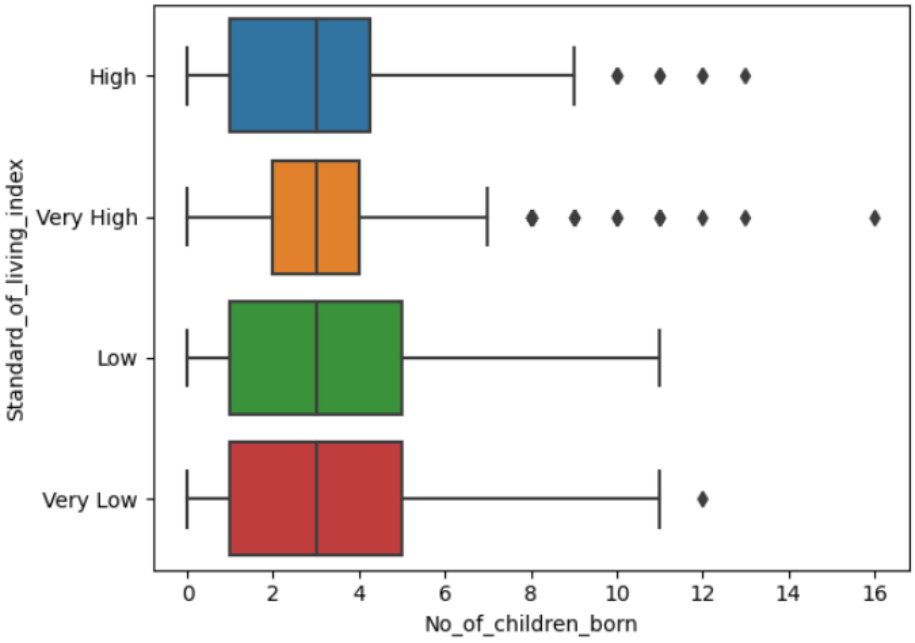

Fig 8- Boxplot of no of children born and standard of living

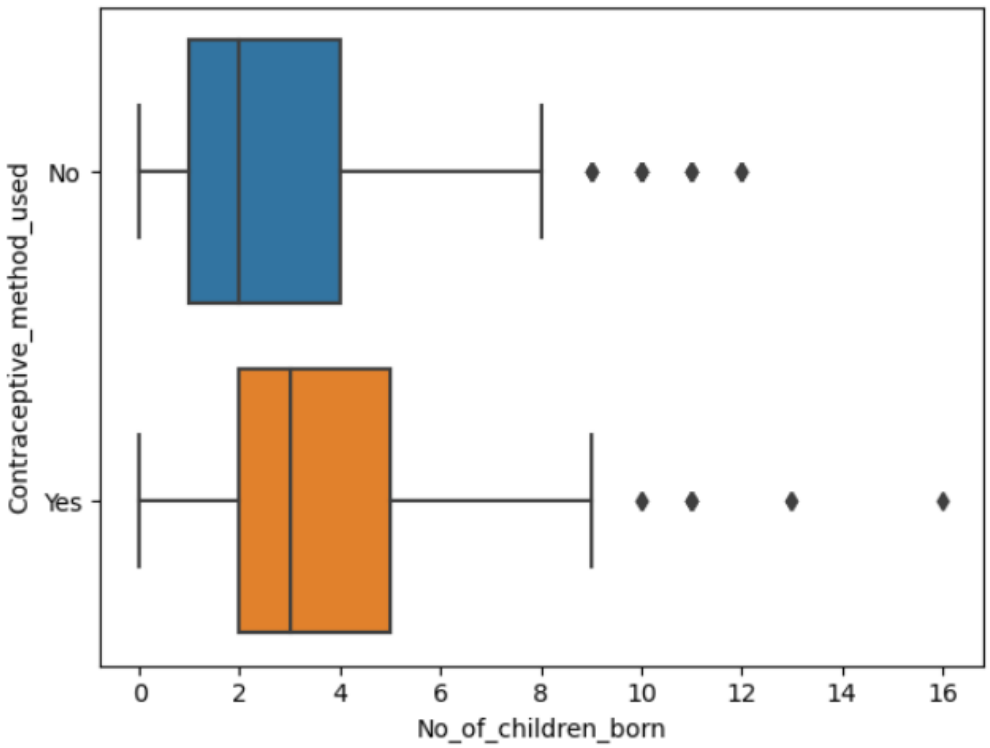

Fig 9- Boxplot of no of children born and contraceptive methods

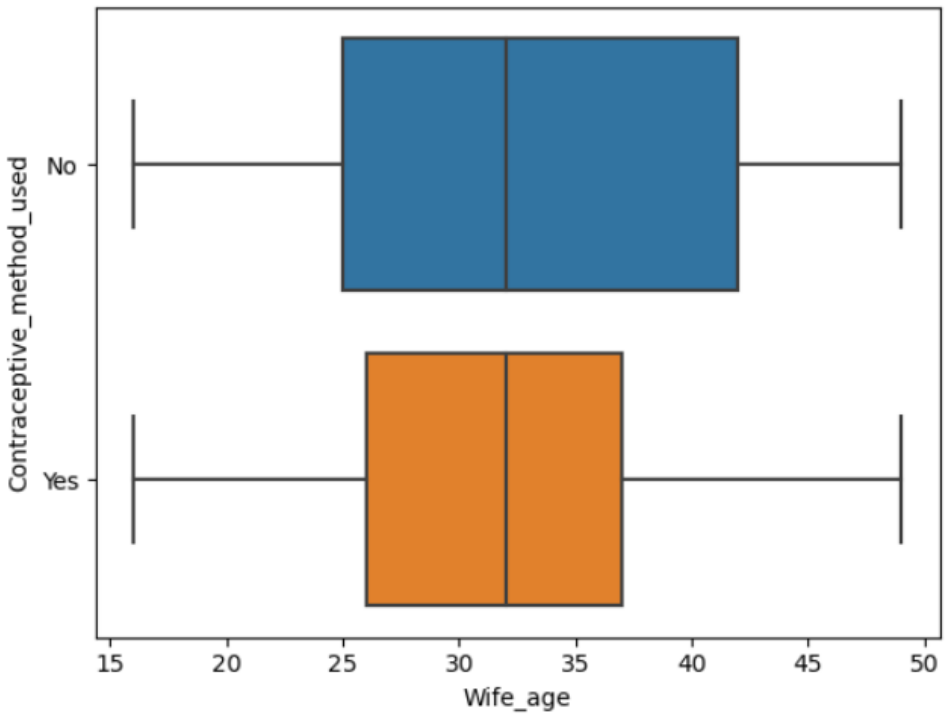

Fig 10- boxplot of wife age and contraceptive methods

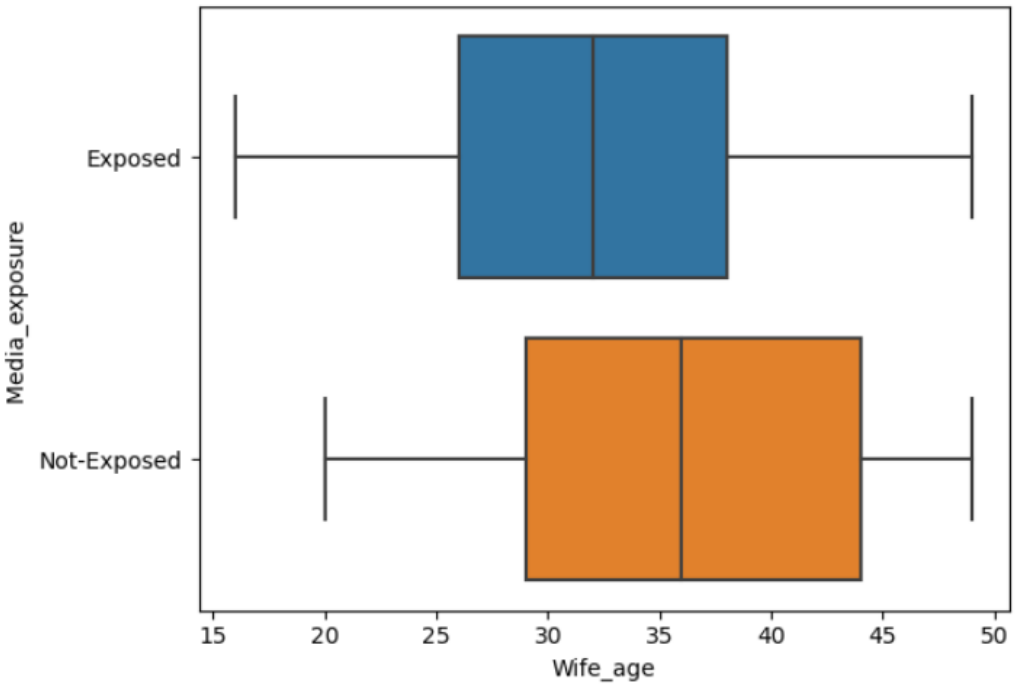

Fig 11- boxplot of wife age and media exposure

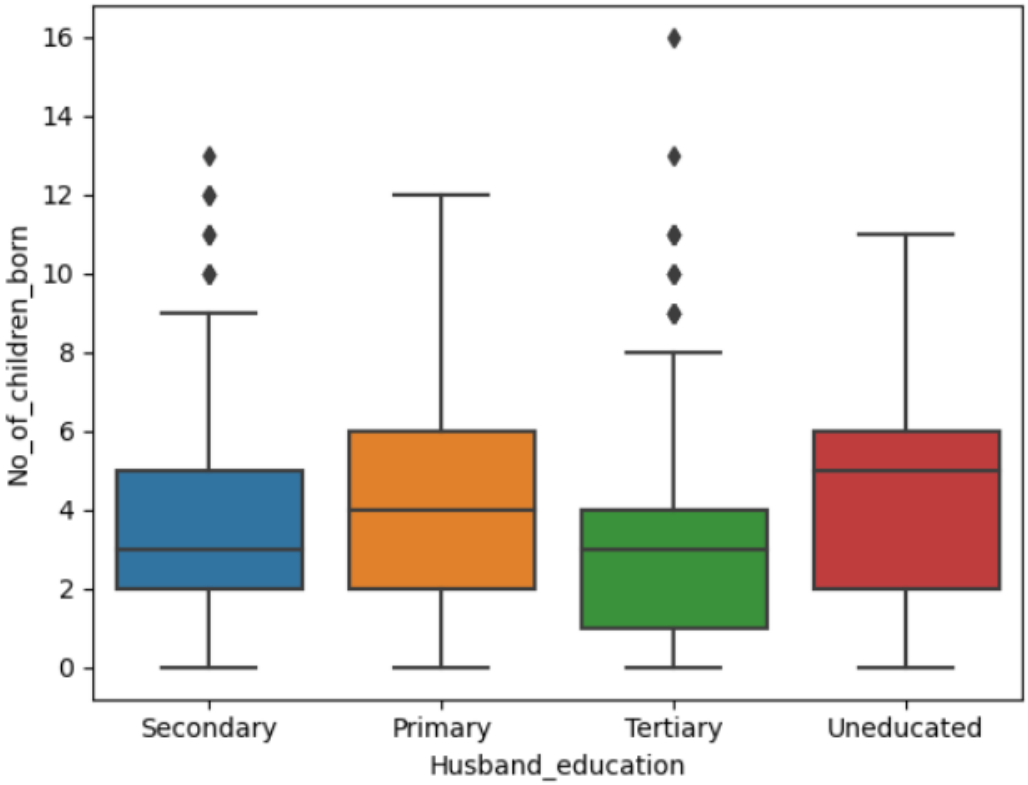

Fig 12- boxplot of husband education and no of children born

2.1.2) Observations and Insights

From the above plots and analysis,

- Wifes age start from 16 upto 49
- No of children born starts from 0 and has very large values like 16.
- From the pairplot we can see that no pattern is observed hence there are no relavant relationship between husband occupation,wifes age and number of children.
- We can see that the number of children does not depend on the standard of living but for people with high standard of living number of children are less when compared to lower standard of living.
- Even if contraceptive measures were used the number of children are same in both the cases.
- Media exposure is equal irrespective of change in age.
- Education level of husbands doesnt effect on the number of children born.

2.2) Data Preprocessing

- Null values for column wifes age and number of children born were treated by imputing the median values of their respective columns.
- All the categorical variables were converted into numerical variables based on the data description provided.
    1. Wife's education :
        - 1=uneducated, 2= primary , 3= secondary , 4=tertiary
    2. Husband's education :
        - 1=uneducated, 2= primary , 3= secondary , 4=tertiary
    3. Wife's religion :
        - 0=Non-Scientology, 1=Scientology
    4. Husband's occupation :
        - 1,2,3,4= random encoding
    5. Standard-of-living index :
        - 1=very low, 2= low, 3= high, 4= very high
    6. Media exposure :
        - 0=Not good, 1= good
    7. Contraceptive method used:
        - 0= No, 1=Yes

Train-Test split:

y= dependent variable(contraceptive method used)

X= independent variables(all the other variables except y)

- Using sklearn the data was converted to Train and test in the ration 70:30
- Further the X_train, X_test, y_train, y_test was created .

2.3) Model Building and Compare the Performance of the Models

2.3.1) Logistic Regression

- Using logistic regression from sklearn the model was created and the data was fit into the model.
- The model was carried out at the training data and it was tested on the testing data, accuracy , classification Report and confusion matrix was calculated.

```
Accuracy: 0.6832579185520362
Classification Report:
              precision    recall  f1-score   support

           0       0.70      0.48      0.57       193
           1       0.68      0.84      0.75       249

    accuracy                           0.68       442
   macro avg       0.69      0.66      0.66       442
weighted avg       0.69      0.68      0.67       442

Confusion Matrix:
[[ 93 100]
 [ 40 209]]
```

- AUC value = 0.714

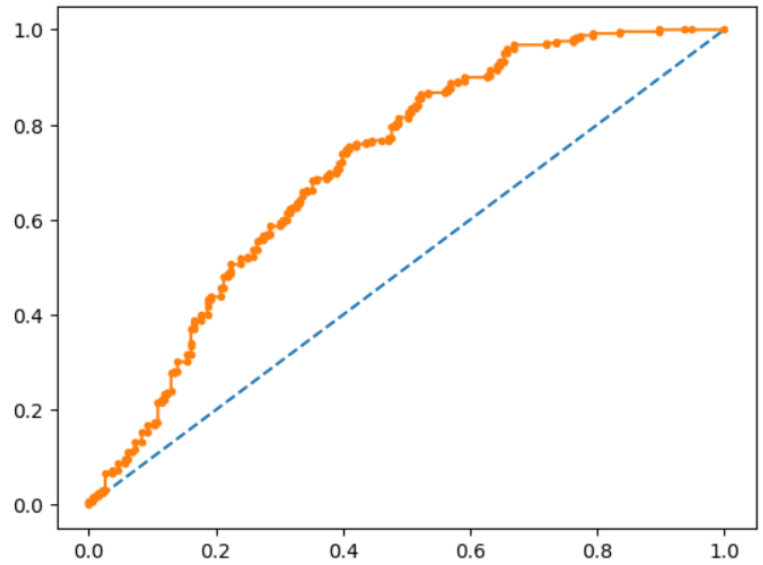

Fig 13- ROC curve (logistic Regression)

From the above logistic regression:

- The accuracy of the model is 0.68 that means about 68% of the predictions are correct for the test data.
- Precision for class 0 is 0.7 which means about 70% of the predictions of class 0 are correct and precision for class 1 is 0.68 which means that about 68% of the predictions of class 1 are correct.
- From the confusion matrix 302 values are correctly predicted while 140 are incorrectly predicted.
- The precision of predicting class 0 is higher than predicting class 1 but the recall for predicting class 1 is much higher than predicting class 0.
- The total number of class 0 is lesser when compared to class 1 which is class imbalance.
- The AUC score is 0.71 which means the distinguishing between the class is not so very good and can be improved.
- The ROC curve is also close to the diagonal line which means the classification is similar to a random guessing.

2.3.2) LDA

- Using LinearDiscriminantAnalysis from sklearn the model was created and the data was fit into the model.
- The model was carried out at the training data and it was tested on the testing data, accuracy , classification Report and confusion matrix was calculated.

```
Accuracy: 0.6855203619909502
Classification Report:
              precision    recall  f1-score   support

           0       0.71      0.48      0.57       193
           1       0.68      0.85      0.75       249

    accuracy                           0.69       442
   macro avg       0.69      0.66      0.66       442
weighted avg       0.69      0.69      0.67       442

Confusion Matrix:
[[ 92 101]
 [ 38 211]]
```

AUC value: 0.712

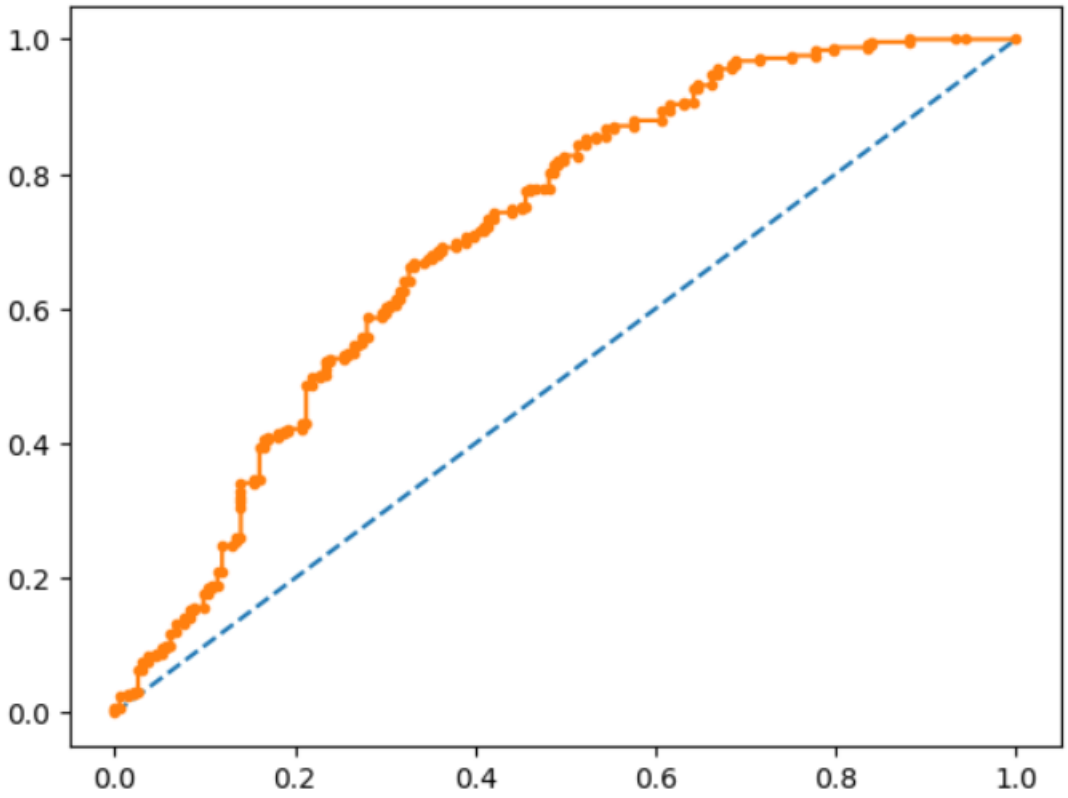

Fig 14- ROC curve(LDA)

From the above LDA:

- The precision, recall and accuracy of the model is exactly the same as that of the logidtic regression.
- The number of classification of 1 and 0 are also same as that of the Logistic regression.
- The AUC score is same as well as ROC curve is also similar.

2.3.3) CART Model

- Using DecisionTreeClassifier with criterion as gini from sklearn the model was created and the data was fit into the model.
- Decision tree was created.

Regularising the Tree:

- After pruning decision tree was again created.
- Below shown are the important features of the Tree.

```
Wife_age                   0.347049
Wife_ education            0.084039
Husband_education          0.044396
No_of_children_born        0.244199
Wife_religion              0.041993
Wife_Working               0.040691
Husband_Occupation         0.093230
Standard_of_living_index   0.086374
Media_exposure             0.018029
```

Before pruning:

- Classification report, accuracy and confusion matrix along with AUC value and ROC plot was calculated.

```
Accuracy: 0.667420814479638
Classification Report:
              precision    recall  f1-score   support

           0       0.62      0.60      0.61       193
           1       0.70      0.72      0.71       249

    accuracy                           0.67       442
   macro avg       0.66      0.66      0.66       442
weighted avg       0.67      0.67      0.67       442

Confusion Matrix:
[[115  78]
 [ 69 180]]
```

AUC Value: 0.666

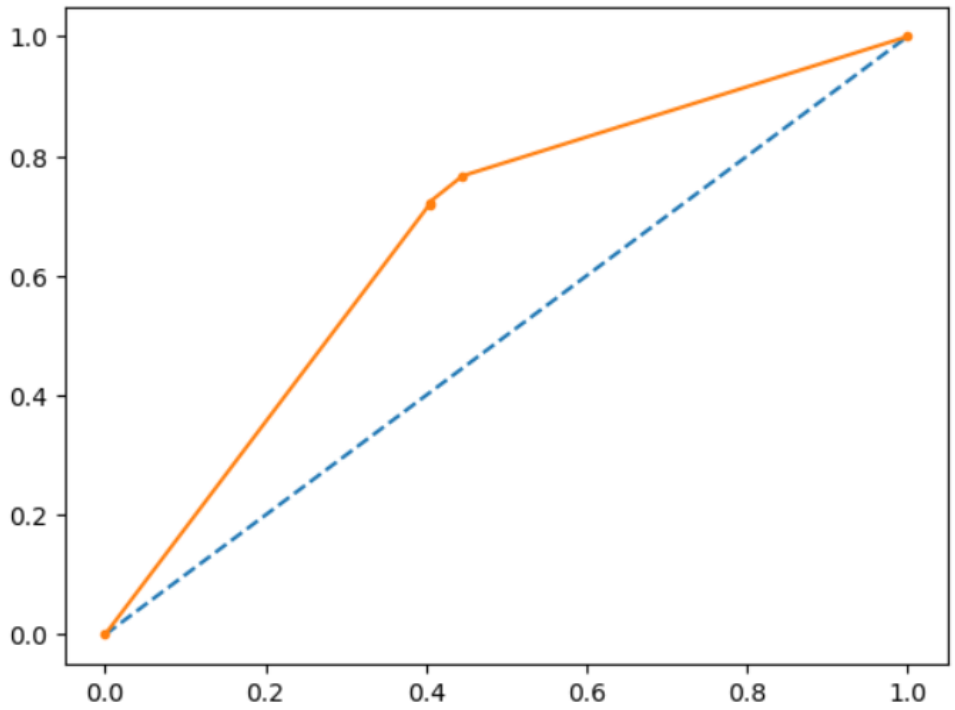

Fig 15- ROC Curve(CART before pruning)

- The overall accuracy of the model is 0.66.
- The precision and the recall values of 0 are 0.62 and 0.60.
- The precision and the recall values of 1 are 0.70 and 0.72.
- The classification of 0 and 1 are 193 and 249.
- From the confusion matrix the True positives are 115 , True negatives are 180, False positives are 78 and False negatives are 69.
- The AUC value is 0.666 , and the ROC curve is very close to the imaginary line.

After Pruning:

Model performance on test data:

```
Accuracy: 0.7036199095022625
Classification Report:
              precision    recall  f1-score   support

           0       0.71      0.54      0.61       193
           1       0.70      0.83      0.76       249

    accuracy                           0.70       442
   macro avg       0.71      0.69      0.69       442
weighted avg       0.71      0.70      0.70       442

Confusion Matrix:
[[104  89]
 [ 42 207]]
```

AUC Value: 0.748

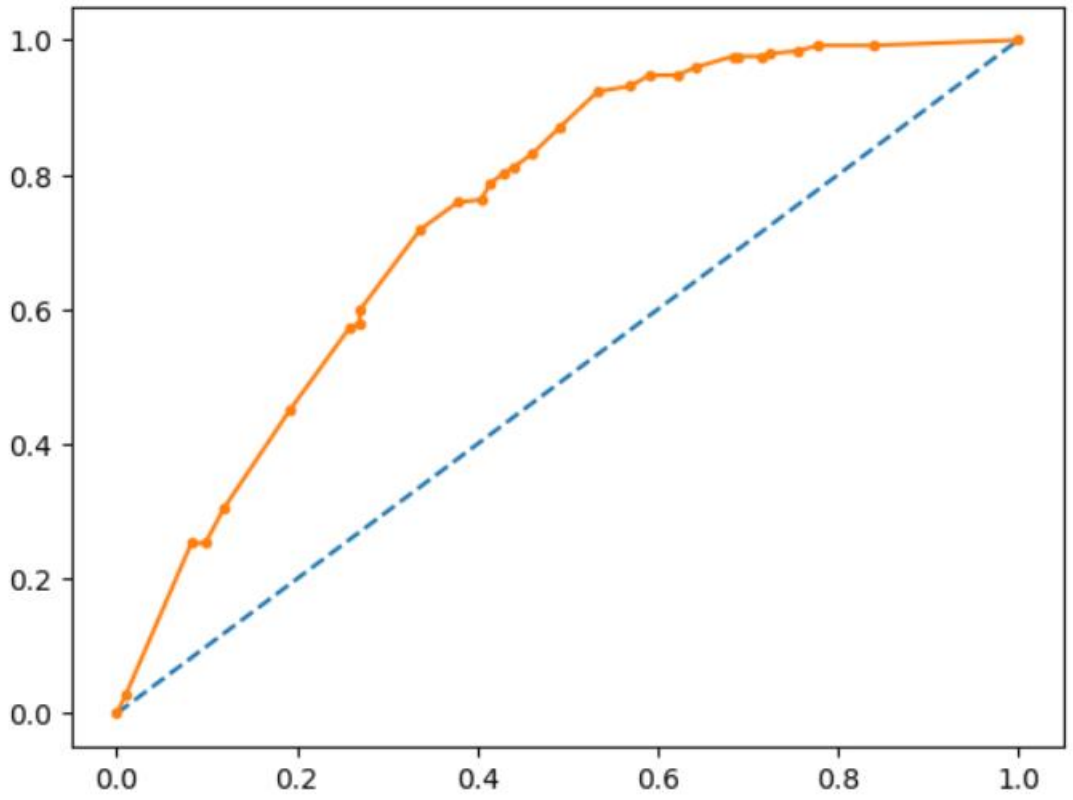

Fig 16- ROC Curve of test(CART after pruning)

Model performance on Train data:

```
Accuracy: 0.7478176527643065
Classification Report:
              precision    recall  f1-score   support

           0       0.75      0.61      0.67       436
           1       0.75      0.85      0.80       595

    accuracy                           0.75      1031
   macro avg       0.75      0.73      0.73      1031
weighted avg       0.75      0.75      0.74      1031

Confusion Matrix:
[[266 170]
 [ 90 505]]
```

AUC Value: 0.807

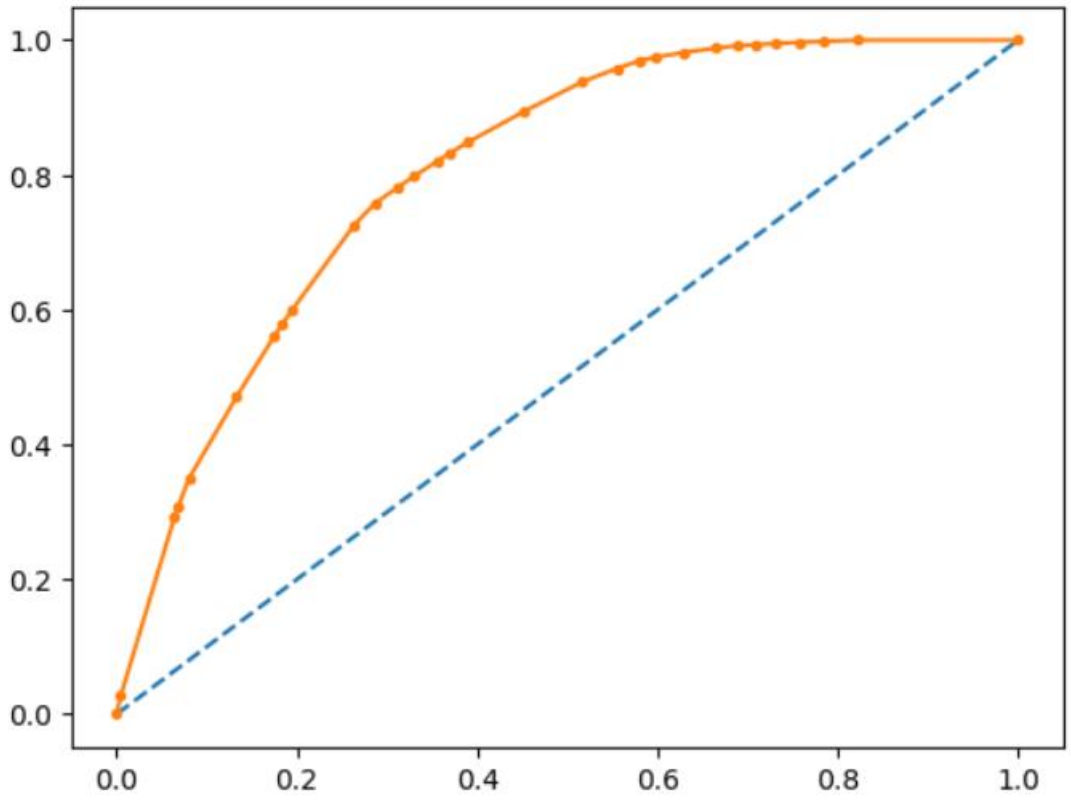

Fig 17- ROC Curve of training (CART after pruning)

From the above cart model:

- From the desicion tree without pruning the data was overfitted, it was complex, and the terminal nodes had sample size of 1 and 2.
- After pruning the tree , the sample size of the terminal nodes have increased , became less complex and became easy to intrepret.

For test data:

- The overall accuracy of the model is 0.70.
- The precision and the recall values of 0 are 0.71 and 0.54.
- The precision and the recall values of 1 are 0.7 and 0.83.
- The classification of 0 and 1 are 193 and 249.
- From the confusion matrix the True positives are 104 , True negatives are 207, False positives are 89 and False negatives are 42.
- The AUC value is 0.74 , and the ROC curve is slightly away from the imaginary line.

For train data:

- The overall accuracy of the model is 0.74.
- The precision and the recall values of 0 are 0.75 and 0.61.
- The precision and the recall values of 1 are 0.75 and 0.85.
- The classification of 0 and 1 are 436 and 595.
- From the confusion matrix the True positives are 266 , True negatives are 505, False positives are 170 and False negatives are 90.
- The AUC value is 0.80 , and the ROC curve is far away from the imaginary line.

2.4) Insights and Recommendations

- Wife age and the number of childrens born are most important parameters.
- Standard of living, husbands occupation as well as wifes education also have relevance in using the contraceptive methods.
- From all the different types of regression techniques used here The regularised cart model provides more relevant classification of the 0 and 1 .
- Before pruning the tree the sample size of the terminal nodes were very less and after pruning it increased drastically hence pruning was effective in improving the model's generalization ability and interpretability.
- The overall accuracy of the model on the test data is 0.70, while on the train data it is 0.74. This indicates that the model performs reasonably well in predicting the target variable. But we cannot let alone take the accuracy value to measure the models performance.
- The precision and recall values for both classes (0 and 1) provide insights into the model's performance for each class.
- The confusion matrix also helps to identify the places of error and give insights on how to improve.
- The AUC value here is 0.80 which is closer to 1 than most other regression model hence the seperation of the postive and negative instances are much better.
- The ROC curve is also far away to the left most corner of the graph hence the classification of the 0 and 1 are much better.