Predictive Modelling Project Report

Preetam Sarmah

PGPDSBA.O.JAN23.A
January’ 23
Date: 04/06/2023

Table of Contents

List of Figures

List of Tables

References PreetamSarmah_04-June-2023.ipynb

Problem 1

Executive Summary

The comp-activ databases is a collection of a computer systems activity measures .
The data was collected from a Sun Sparcstation 20/712 with 128 Mbytes of memory running in a multi-user university department. Users would typically be doing a large variety of tasks ranging from accessing the internet, editing files or running very cpu-bound programs.

Introduction

As you are a budding data scientist you thought to find out a linear equation to build a model to predict 'usr'(Portion of time (%) that cpus run in user mode) and to find out how each attribute affects the system to be in 'usr' mode using a list of system attributes.

1.1 Read the data and do exploratory data analysis. Describe the data briefly. (Check the Data types, shape, EDA, 5 point summary). Perform Univariate, Bivariate Analysis, Multivariate Analysis.

Data types

```
RangeIndex: 8192 entries, 0 to 8191
Data columns (total 22 columns):
 #   Column    Non-Null Count  Dtype  
---  ------    --------------  -----  
 0   lread     8192 non-null   int64  
 1   lwrite    8192 non-null   int64  
 2   scall     8192 non-null   int64  
 3   sread     8192 non-null   int64  
 4   swrite    8192 non-null   int64  
 5   fork      8192 non-null   float64
 6   exec      8192 non-null   float64
 7   rchar     8088 non-null   float64
 8   wchar     8177 non-null   float64
 9   pgout     8192 non-null   float64
 10  ppgout    8192 non-null   float64
 11  pgfree    8192 non-null   float64
 12  pgscan    8192 non-null   float64
 13  atch      8192 non-null   float64
 14  pgin      8192 non-null   float64
 15  ppgin     8192 non-null   float64
 16  pflt      8192 non-null   float64
 17  vflt      8192 non-null   float64
 18  runqsz    8192 non-null   object 
 19  freemem   8192 non-null   int64  
 20  freeswap  8192 non-null   int64  
 21  usr       8192 non-null   int64  
dtypes: float64(13), int64(8), object(1)
```

EDA and 5 point Summary

- There are total 8192 rows and 22 columns in the dataset. Out of 22, 1 column is of object type and rest
  21 are of either integer or float data type.
- There are null values for rchar (104 nulls) and wchar (15 nulls)
- There are no duplicated data

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
lread	8192.0	NaN	NaN	NaN	19.559692	53.353799	0.0	2.0	7.0	20.0	1845.0
lwrite	8192.0	NaN	NaN	NaN	13.106201	29.891726	0.0	0.0	1.0	10.0	575.0
scall	8192.0	NaN	NaN	NaN	2306.318237	1633.617322	109.0	1012.0	2051.5	3317.25	12493.0
sread	8192.0	NaN	NaN	NaN	210.47998	198.980146	6.0	86.0	166.0	279.0	5318.0
swrite	8192.0	NaN	NaN	NaN	150.058228	160.47898	7.0	63.0	117.0	185.0	5456.0
fork	8192.0	NaN	NaN	NaN	1.884554	2.479493	0.0	0.4	0.8	2.2	20.12
exec	8192.0	NaN	NaN	NaN	2.791998	5.212456	0.0	0.2	1.2	2.8	59.56
rchar	8088.0	NaN	NaN	NaN	197385.728363	239837.493526	278.0	34091.5	125473.5	267828.75	2526649.0
wchar	8177.0	NaN	NaN	NaN	95902.992785	140841.707911	1498.0	22916.0	46619.0	106101.0	1801623.0
pgout	8192.0	NaN	NaN	NaN	2.285317	5.307038	0.0	0.0	0.0	2.4	81.44
ppgout	8192.0	NaN	NaN	NaN	5.977229	15.21459	0.0	0.0	0.0	4.2	184.2
pgfree	8192.0	NaN	NaN	NaN	11.919712	32.36352	0.0	0.0	0.0	5.0	523.0
pgscan	8192.0	NaN	NaN	NaN	21.526849	71.14134	0.0	0.0	0.0	0.0	1237.0
atch	8192.0	NaN	NaN	NaN	1.127505	5.708347	0.0	0.0	0.0	0.6	211.58
pgin	8192.0	NaN	NaN	NaN	8.27796	13.874978	0.0	0.6	2.8	9.765	141.2
ppgin	8192.0	NaN	NaN	NaN	12.388586	22.281318	0.0	0.6	3.8	13.8	292.61
pflt	8192.0	NaN	NaN	NaN	109.793799	114.419221	0.0	25.0	63.8	159.6	899.8
vflt	8192.0	NaN	NaN	NaN	185.315796	191.000603	0.2	45.4	120.4	251.8	1365.0
runqsz	8192	2	Not_CPU_Bound	4331	NaN	NaN	NaN	NaN	NaN	NaN	NaN
freemem	8192.0	NaN	NaN	NaN	1763.456299	2482.104511	55.0	231.0	579.0	2002.25	12027.0
freeswap	8192.0	NaN	NaN	NaN	1328125.959839	422019.426957	2.0	1042623.5	1289289.5	1730379.5	2243187.0
usr	8192.0	NaN	NaN	NaN	83.968872	18.401905	0.0	81.0	89.0	94.0	99.0

Table 1 - Problem 1 Data Description

Based on the mean and median , the data shows that there is a skew.

The runsqz has 2 value types, No_CPU_Bound which has a frequency of 4331 and CPU_Bound.

Usr (portion of time (%) that cpus run in user mode) is the dependent variable , which takes a value from 0 to 99%.

There are attributes, which have min value 0 , which are genuine values based on the dataset.

Univariate Analysis

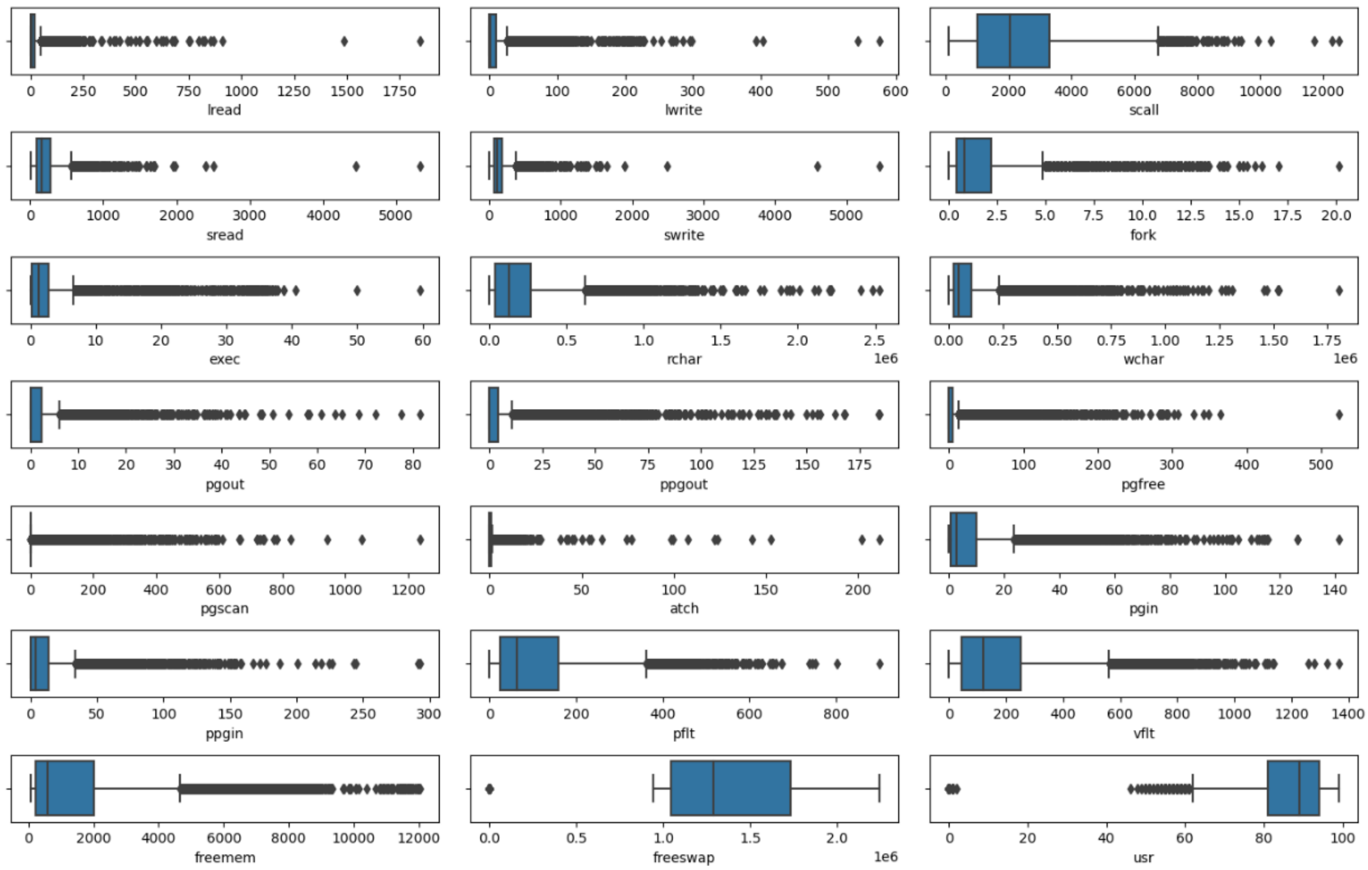

Fig 1 Problem 1 Univariate Analysis

From the initial univariate analysis its clear that the various attributes are not normally distributed and also that there are a lot of outliers present.

Bivariate Analysis

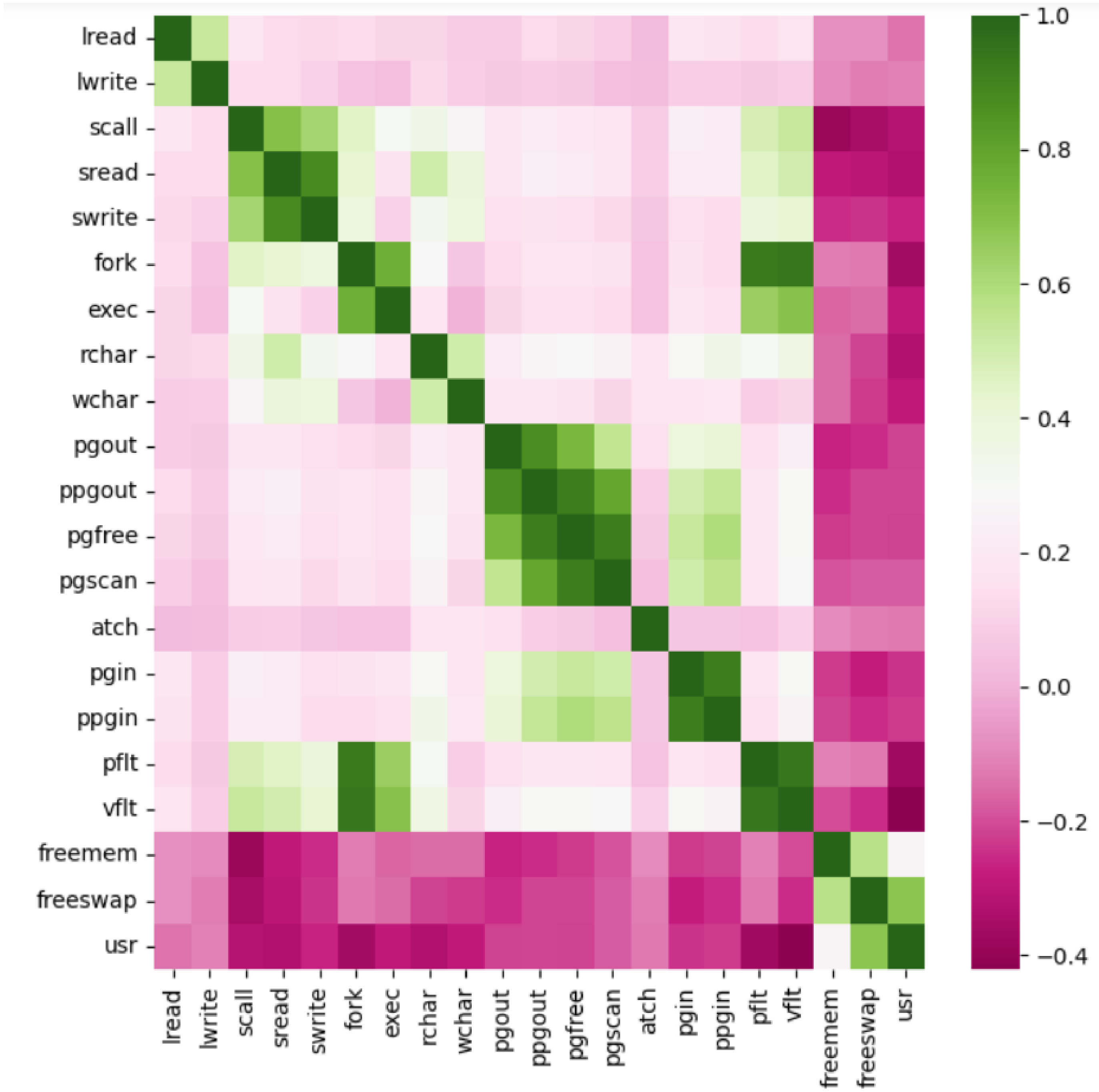

Fig 2 Problem 1 Correlation Plot

From the correlation plot, we can see that various attributes of the dataset are highly correlated to each other. Correlation values near to 1 are highly positively correlated. Correlation values near to 0 are not correlated to each other.

The attribute freeswap is highly correlated to the usr.

The attributes vflt and fork have highest negative correlation and freemem seems to have little to no correlation to the dependent variable

Multi-Variate Analysis

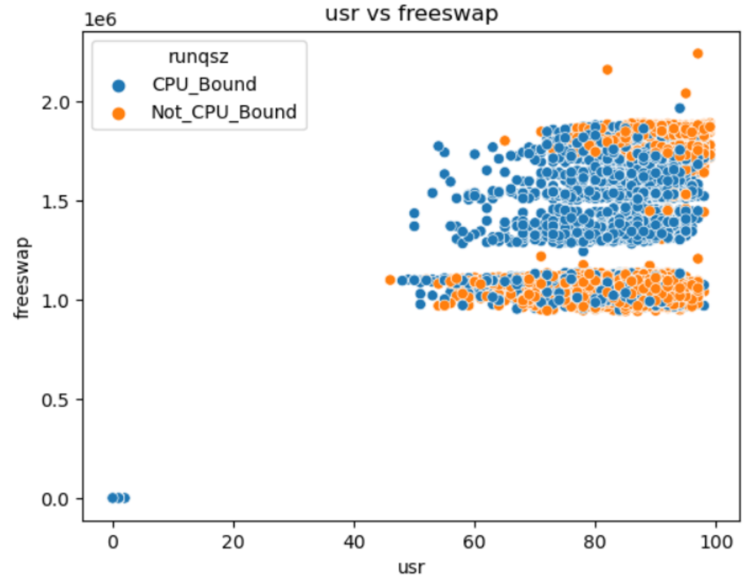

Fig 3 usr vs freeswap

Usr vs freeswap shows a net positive correlation , with process run queue time (runsqz) favouring CPU bound for high freeswap and usr , where as it favours not CPU bound for moderate freeswap and high usr.

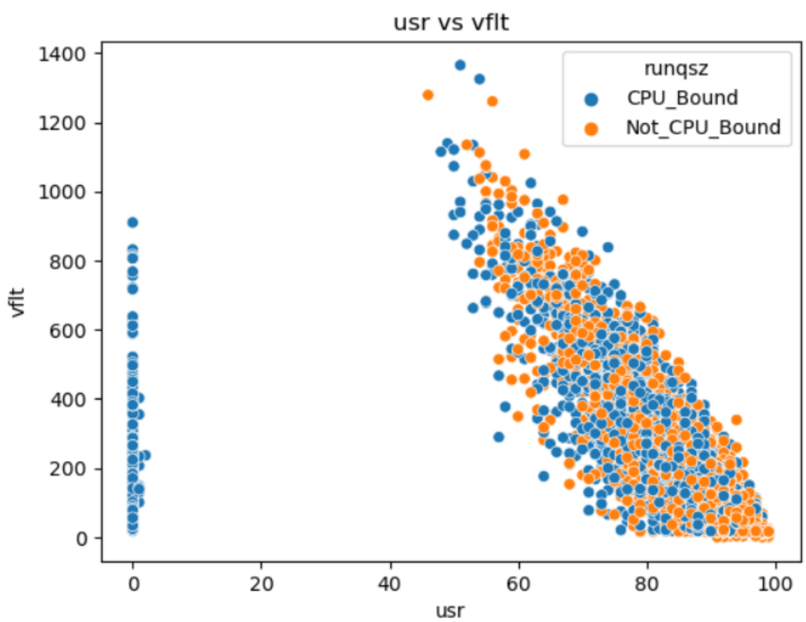

Fig 4 usr vs vflt

Usr vs vflt shows a net negative correlation , with process run queue time (runsqz) being equally distributed between CPU bound and not CPU bound. Also it can be noted that for usr equal to 0,Number of page faults caused by address translation (vflt) is high

1.2 Impute null values if present, also check for the values which are equal to zero. Do they have any meaning or do we need to change them or drop them? Check for the possibility of creating new features if required. Also check for outliers and duplicates if there.

Impute null values

The null values can be imputed with the median of the respective attributes rchar and wchar , as the attributes are not normally distributed.

Duplicates

The dataset has no duplicates in the dataset

Outliers

There are several outliers present , since the linear regression model is sensitive to outliers we choose to treat them based on the Interquartile range (IQR)

lower_range= Q1 - (1.5 * IQR)

upper_range= Q3 + (1.5 * IQR)

Where Q1 = 1st Quartile , Q3 = 3rd Quartile and IQR = Q3-Q1

Possibility of creating new features

From the previous correlation plot it was observed that as few attributes had high correlation with one another ie multicollinearity, and hence in the model , only one of the two multicollinear features is likely to be considered in the optimized model.

1.3 Encode the data (having string values) for Modelling. Split the data into train and test (70:30). Apply Linear regression using scikit learn. Perform checks for significant variables using appropriate method from statsmodel. Create multiple models and check the performance of Predictions on Train and Test sets using Rsquare, RMSE & Adj Rsquare. Compare these models and select the best one with appropriate reasoning.

Encoding the data

The only object column in runqsz and since its a nominal in nature we choose to perform dummy variable encoding.

Data Splitting

The Data splitting is done in such that we have 70% of the data as train and 30% of the data as test, using model selection train_test_split

Linear regression

We create the linear regression model by fitting and transforming on the train data of dependent variable and the independent variables

Significant Variables

The coefficient for const or intercept of the model is 84.12174079532215

The coefficient for lread is	-0.06348150618192322
The coefficient for lwrite is	0.048161287091430305
The coefficient for scall is	-0.0006638280111671737
The coefficient for sread is	0.00030825210314083286
The coefficient for swrite is	-0.005421822297640978
The coefficient for fork is	0.029312727248894666
The coefficient for exec is	-0.3211664838986006
The coefficient for rchar is	-5.166841759434584e-06
The coefficient for wchar is	-5.402875235423325e-06
The coefficient for pgout is	-0.3688190638729695
The coefficient for ppgout is	-0.0765976821274773
The coefficient for pgfree is	0.08448414470555507
The coefficient for pgscan is	4.0018303196836174e-14
The coefficient for atch is	0.6275741574807907
The coefficient for pgin is	0.01998790767868225
The coefficient for ppgin is	-0.06733383975703425

The coefficient for pflt is -0.0336028293775232
The coefficient for vflt is -0.005463668798514263
The coefficient for freemem is -0.00045846718795069537
The coefficient for freeswap is 8.831840263021474e-06
The coefficient for runqsz_Not_CPU_Bound is 1.6152978488248837

The most significant predictor variable is runqsz_Not_CPU_Bound , followed by atch.

Rsquare, RMSE & Adj Rsquare

The Models prepared by Sklearn and statsmodels both have approximately similar values for RSquare, Root Mean Square Error and Adjusted RSquare

- The Root Mean Square Error of train data is 4.42 , where as Root Mean Square error for the test data is 4.65
- The R Square of train data is 0.7961 , whereas the R Square of the test data is 0.7677 i.e, 79.6 % of the variation in the usr is explained by the predictors in the model for train set and 76.8 % of the variation in the usr is explained by the predictors in the model for test set.
- The Adj R Sqaure of the train data is 0.795 ,

Variance Inflation Factor (VIF)

The Variance Inflation Factor tells what percentage of te variance is inflated for each coefficient,
The following are the ViF values for each coefficient
VIF values:

const	29.229332
lread	5.350560
lwrite	4.328397
scall	2.960609
sread	6.420172
swrite	5.597135
fork	13.035359
exec	3.241417
rchar	2.133616
wchar	1.584381
pgout	11.360363
ppgout	29.404223
pgfree	16.496748
pgscan	NaN

atch	1.875901
pgin	13.809339
ppgin	13.951855
pflt	12.001460
vflt	15.971049
freemem	1.961304
freeswap	1.841239
runqsz_Not_CPU_Bound	1.156815

Ideally we try not to have VIF values more that 5 and above 10 is not acceptable.

- The VIF values indicate that the features ppgout, pgfree, vflt,ppgin,pgin,fork,pflt, pgout are correlated with one or more independent features.
- To treat multicollinearity, we will have to drop one or more of the correlated features.
- We will drop the variable that has the least impact on the adjusted R-squared of the model.

Best Model

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                    usr   R-squared:                       0.795
Model:                            OLS   Adj. R-squared:                  0.794
Method:                 Least Squares   F-statistic:                     1705.
Date:                Sun, 04 Jun 2023   Prob (F-statistic):               0.00
Time:                        15:05:02   Log-Likelihood:                -16675.
No. Observations:                5734   AIC:                         3.338e+04
Df Residuals:                    5720   BIC:                         3.347e+04
Df Model:                          13
Covariance Type:            nonrobust
=========================================================================================
                          coef    std err          t      P>|t|      [0.025      0.975]
-----------------------------------------------------------------------------------------
const                  84.1149      0.311    270.640      0.000      83.506      84.724
lread                  -0.0364      0.004     -8.196      0.000      -0.045      -0.028
scall                  -0.0007   5.96e-05    -11.311      0.000      -0.001      -0.001
swrite                 -0.0058      0.001     -5.533      0.000      -0.008      -0.004
exec                   -0.3707      0.048     -7.664      0.000      -0.466      -0.276
rchar               -5.329e-06   4.36e-07    -12.233      0.000   -6.18e-06   -4.47e-06
wchar               -4.581e-06   1.02e-06     -4.505      0.000   -6.57e-06   -2.59e-06
pgout                  -0.3452      0.038     -9.018      0.000      -0.420      -0.270
pgscan               1.899e-13   7.25e-16    261.945      0.000    1.88e-13    1.91e-13
atch                    0.6046      0.143      4.240      0.000       0.325       0.884
ppgin                  -0.0645      0.006    -10.009      0.000      -0.077      -0.052
pflt                   -0.0406      0.001    -38.939      0.000      -0.043      -0.039
freemem                -0.0005   5.06e-05     -9.226      0.000      -0.001      -0.000
freeswap             8.937e-06   1.86e-07     48.020      0.000    8.57e-06     9.3e-06
runqsz_Not_CPU_Bound    1.6380      0.126     13.012      0.000       1.391       1.885
==============================================================================
Omnibus:                     1048.939   Durbin-Watson:                   2.012
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2212.451
Skew:                          -1.075   Prob(JB):                         0.00
Kurtosis:                       5.153   Cond. No.                     8.68e+21
==============================================================================
```

79.5 % of the variation in the usr is explained by the predictors in the model and runqsz_Not_CPU_Bound , atch have the highest coefficients which indicates these are the strong predictor variables

Variance Inflation Factor of predictors are well below 5 as follows

VIF values:

lread	1.299453
scall	2.649728
swrite	3.011905
exec	2.830182
rchar	1.695027

wchar	1.528573
pgout	2.044776
pgscan	NaN
atch	1.860195
ppgin	1.484607
pflt	3.303635
freemem	1.944783
freeswap	1.757192
runqsz_Not_CPU_Bound	1.148886

Equation of Linear Regression

usr = 84.11486488706325 + -0.03635273052516947 * (lread) + -0.0006739702280041565 * (scall) + -0.005840354651889987 * (swrite) + -0.3707104216153422 * (exec) + -5.328664415674294e-06 * (rchar) + -4.581266015721766e-06 * (wchar) + -0.3451883883905269 * (pgout) + 1.898722143755558e-13 * (pgscan) + 0.6046496339972265 * (atch) + -0.06453366898124792 * (ppgin) + -0.04058413274785204 * (pflt) + -0.0004671884185384003 * (freemem) + 8.937361937692906e-06 * (freeswap) + 1.6380431703984373 * (runqsz_Not_CPU_Bound)

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

1. Our goal is to predict which attributes significantly affect Portion of time (%) that cpus run in user mode ie usr
2. The data is initially sanitized and outliers are treated as they can significantly affect the linear regression model
3. When the model is being run, we ensure that all the predictor variables and dependent variable in float or int data type , as the model cannot , take direct string values and hence need to be encoded
4. We split the data into training and testing data in 70:30 ratio,we train the model on train data and predict on the test data.
5. The linear equation from above states that for unit increase in the process run queue size or runqsz of type not CPU bound there is a 1.64 times increase the overall usr given all the other attribute values remain the same.
6. exec - Number of system exec calls per second has the most significant effect in reducing usr by 0.370 , given all the other attribute values remain the same.
7. pgout - Number of page out requests per second will reduce usr by 0.345 given all the other attribute values remain the same.

Problem 2

Executive Summary

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

Introduction

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

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.

Data types

```
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)
```

EDA and 5 point Summary

- There are total 1473 rows and 10 columns in the dataset. Out of 10, 7 columns are of object type and rest 3 are of either integer or float data type.
- There are null values for Wife_age(71 nulls) and No_of_children_born(21 nulls)
- There are 80 rows of duplicated data

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

Table 2 Problem 2 Data Description

Based on the mean and median , the numerical attributes shows that they are almost normally distributed

Most of the households seem to have media exposure at 1364 entries

Most of the wives are Non Working and adhere to Scientology religion

46% of the households have a Very High Standard of Living

The median age of wives is 32 years.

Imputing missing values

We Impute missing values for wife_age, No_of_children_born using median of the respective two attributes

Check for duplicates

After imputing the missing values we have 85 duplicates , we choose to drop these duplicates

Outlier Check

There are Ouliers as the maximum number of children seems to 16 , so we treat the outliers using the Interquartile Range or IQR lower_range= Q1 - (1.5 * IQR)

upper_range= Q3 + (1.5 * IQR)

Where Q1 = 1st Quartile , Q3 = 3rd Quartile and IQR = Q3-Q1

Univariate Analysis

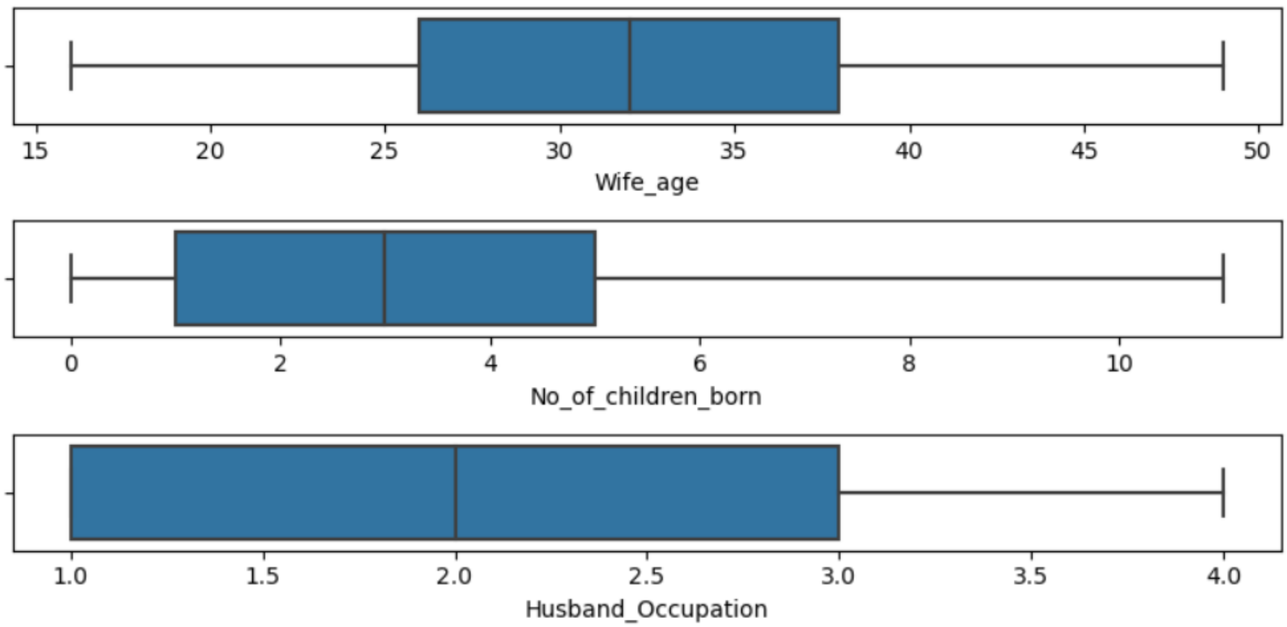

Fig 5 Problem 2 Univariate Analysis Numerical

From the initial univariate analysis its clear that the various attributes are almost normally distributed .

The Mean number of children is 3 and maximum is 11 , The mean wife age is 32 and maximum is 49.

The husband_occupation is actually a categorical data between 1 and 4

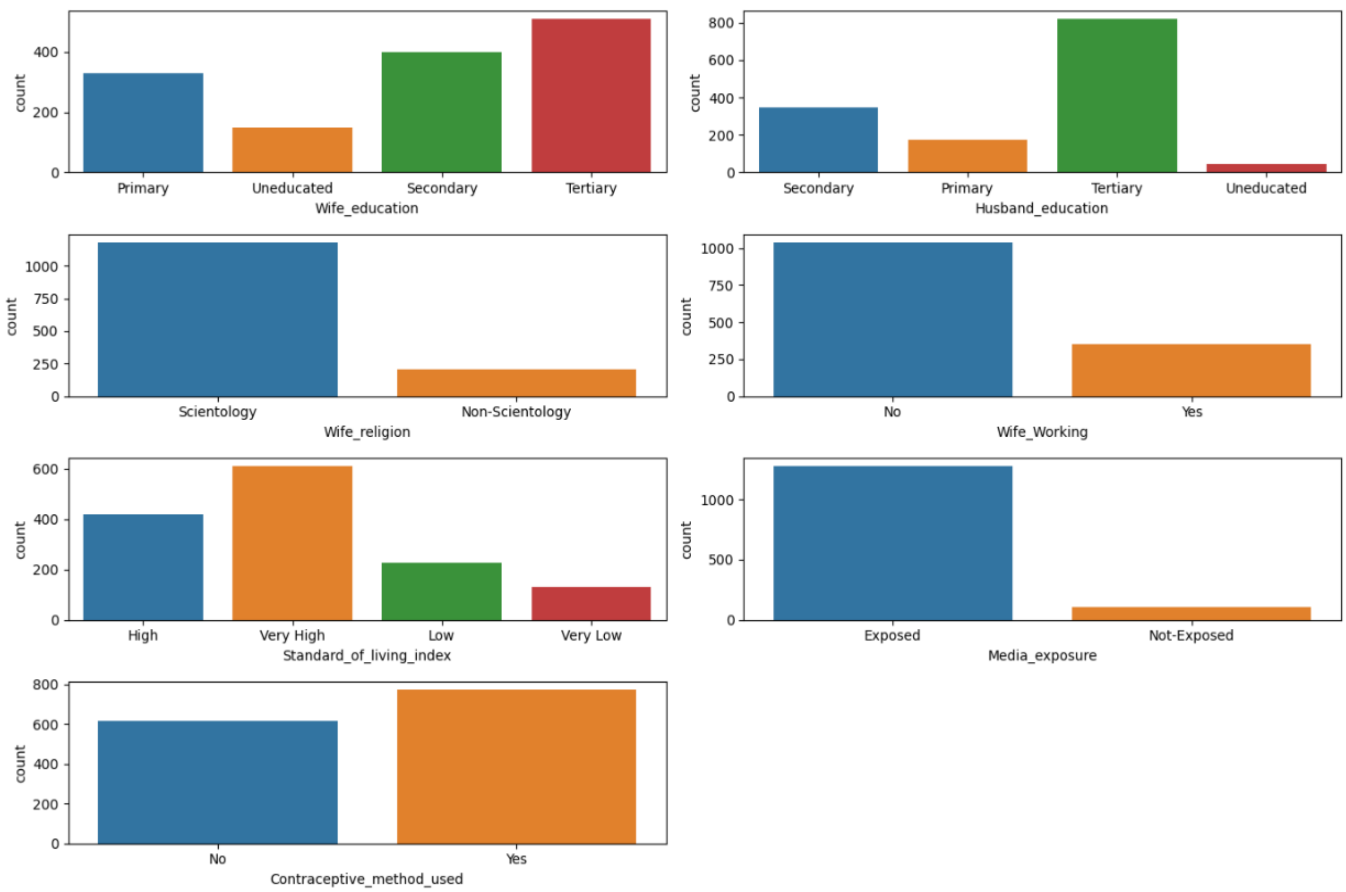

Fig 6 Problem 2 Univariate Analysis Object

Wife and Husbands education of maximum households belong to tertiary education.

Most wives are in favour of a contraceptive.

Bivariate Analysis

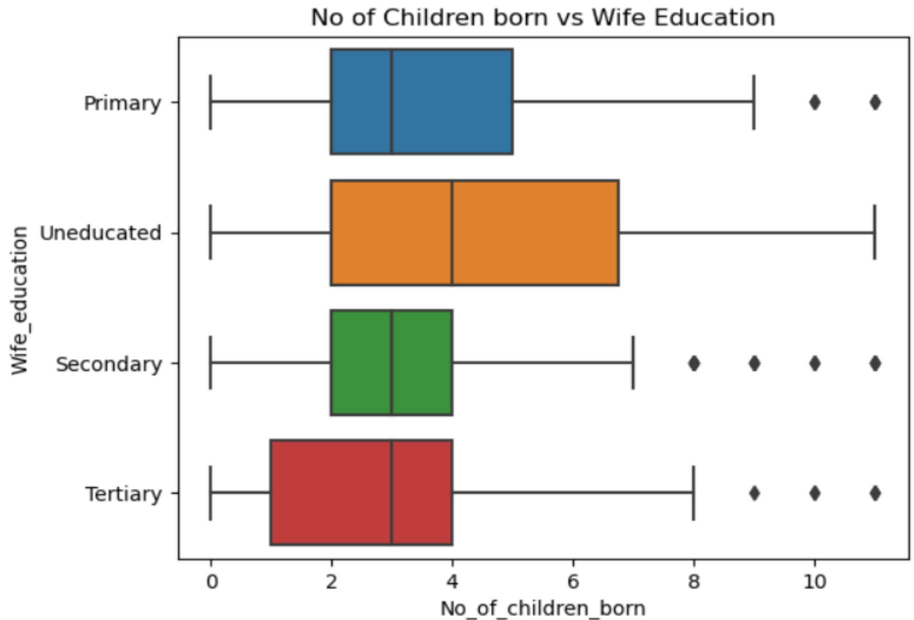

Fig 7 No of children born vs wife education

Uneducation women on averag tend to have more children , than those having some form of education, though there are a few exceptions to the same

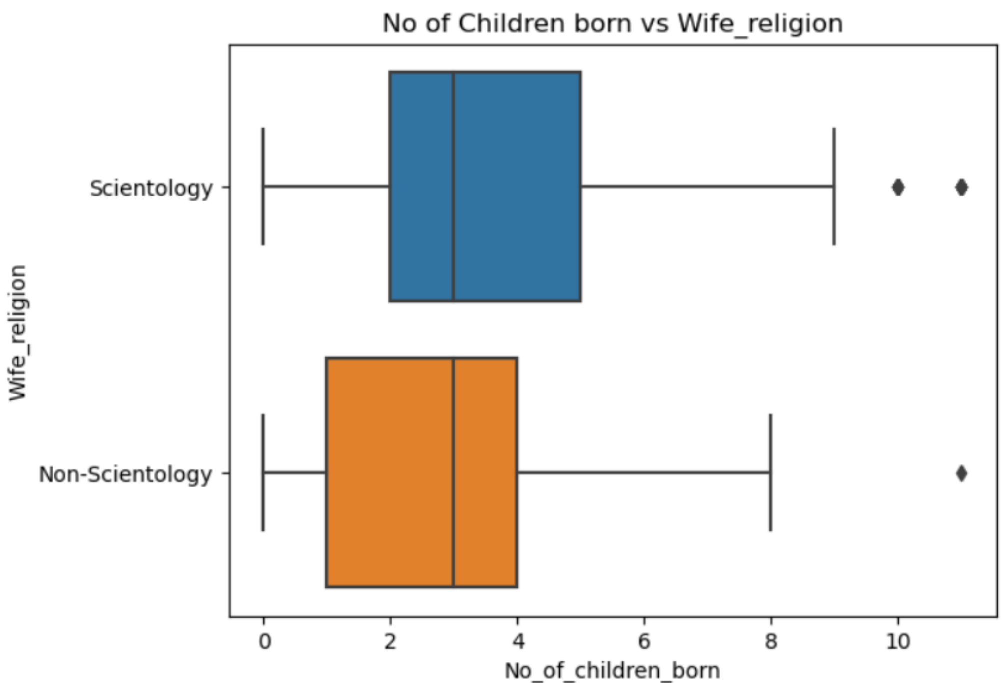

Fig 8 No of children born vs wife religion

Women adhering to Scientology , tend to have more children , despite the median being same as Non-Scientology adherents

Multi-Variate Analysis

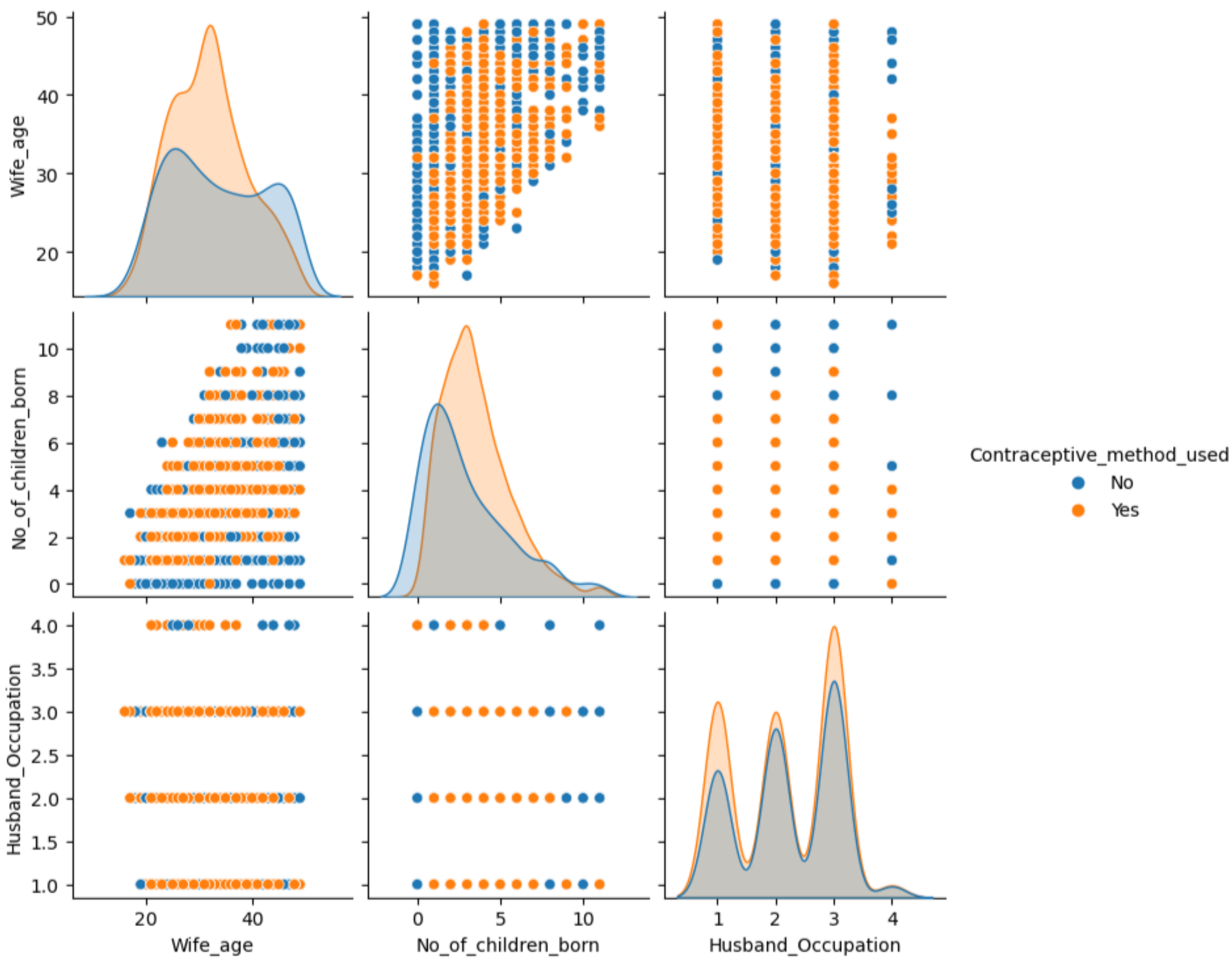

Fig 9 Pairplot

Wife_age and No_of_children_born show a weak positive correlation

Wifes with more children tend to prefer contraceptive methods

With increasing age wives tend to not use contraceptive methods , this might be due to fertility issue making them unable to bear kids.

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.

Encoding the data

Wife Education , Husband Education and Standard of Living Index are ordinal type and hence they are encoded as 1,2,3,4 in increasing order of weight and other object attributes are binary yes or no and are likewise encoded as 0,1

Data Split

The Data splitting is done in such that we have 70% of the data as train and 30% of the data as test, using model selection train_test_split and we use Stratify with Dependent Variable ie contraceptive_method_used , so as to preserve the ratio of Yes(1) and No(0) in both the test and train data.

Logistic Regression

Logistic Regression is a classification modelling where a logit function is used to classify a observation into a class

The following are the feature coefficients per Logistic Regression

Wife_age : -0.08

Wife_education : 0.51

Husband_education : 0.03

No_of_children_born : 0.33

Wife_religion : -0.49

Wife_Working : -0.2

Husband_Occupation : 0.18

Standard_of_living_index : 0.31

Media_exposure : 0.36

Wife Education, No of Children Born and Media Exposure are the important features as per LDA

Linear Discriminant Analysis

LDA finds a linear combination of predictor variables (a Linear Discriminant Function) that best separates the classes of the response variable.

The following are the feature coefficients per LDA

Wife_age : -0.08

Wife_education : 0.5

Husband_education : 0.02

No_of_children_born : 0.32

Wife_religion : -0.51

Wife_Working : -0.19

Husband_Occupation : 0.18

Standard_of_living_index : 0.32

Media_exposure : 0.39

Wife Education, No of Children Born and Media Exposure are the important features as per LDA

Classification and Regression Tree

CART is a decision tree used for classification as well as regression and is based on gini gain and gini index of nodes

A node with gini index 0.5 is a highly impure node and node with gini index 0 is a highly pure node

The following are the feature importances for CART

Wife_age	0.328187
No_of_children_born	0.246930
Wife_education	0.106439
Husband_Occupation	0.080228
Standard_of_living_index	0.072396
Husband_education	0.063915
Wife_Working	0.055484
Wife_religion	0.034044
Media_exposure	0.012376

Wife Age, No of Children Born and Wife Education are the important features as per 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.

Logistic Regression

The accuracy of the logistic regression model on the training data is 0.67 and the accuracy is 0.65 for the testing data

Confusion Matrix for Training Data and Testing Data.

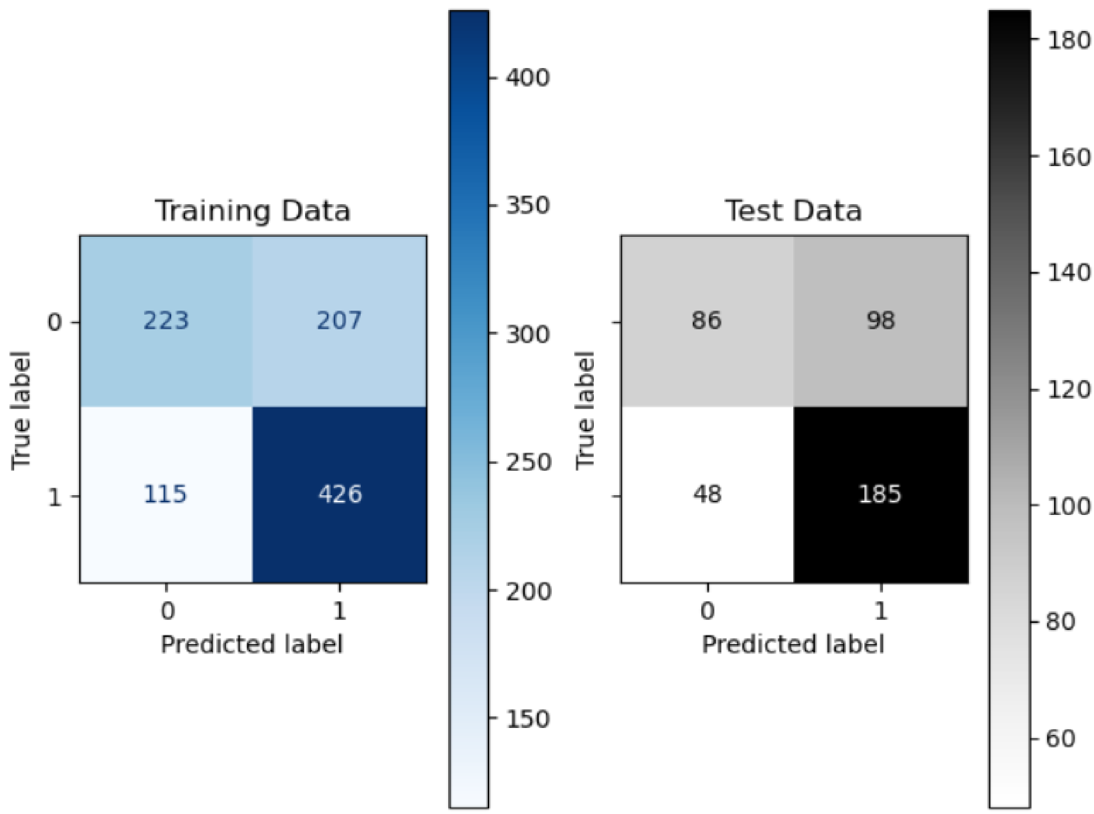

Fig 10 Confusion Matrix Logistic Regression

AUC Score for the Training Data: 0.719
AUC Score for the Test Data: 0.663

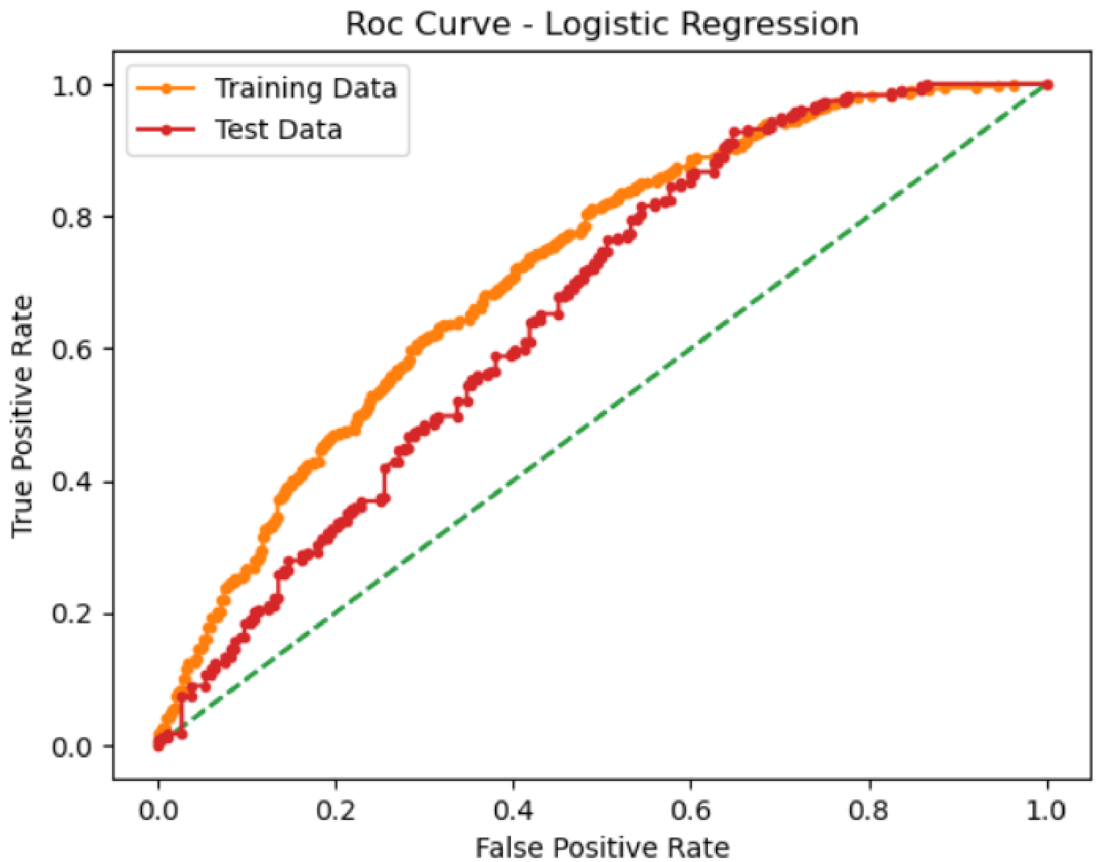

Fig 11 ROC Curve Logistic Regression

Linear Discriminant Analysis

The accuracy of the LDA model on the training data is 0.68 and the accuracy is 0.65 for the testing data

Confusion Matrix for Training Data and Testing Data.

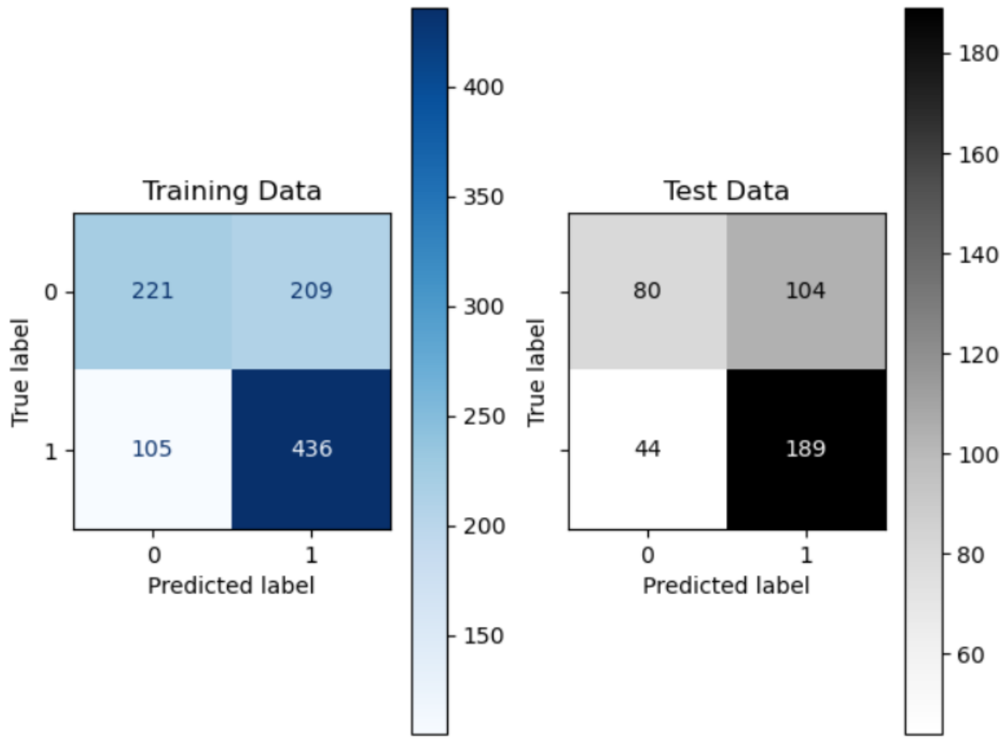

Fig 12 Confusion Matrix LDA

AUC Score for the Training Data: 0.719
AUC Score for the Test Data: 0.662

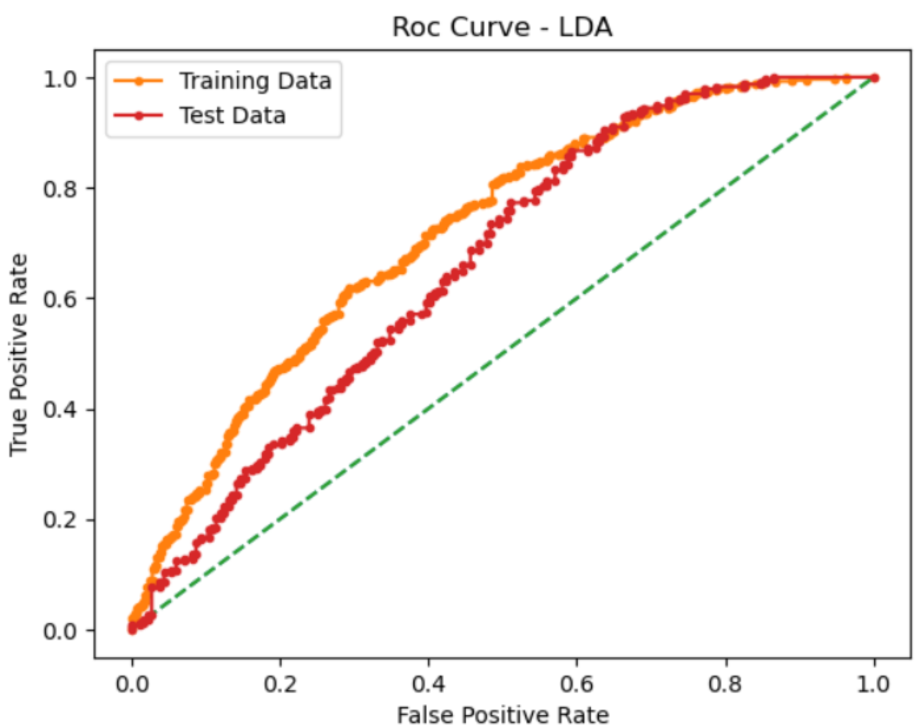

Fig 13 ROC Curve LDA

Classification and Regression Tree

The accuracy of the CART model on the training data is 0.99 and the accuracy is 0.59 for the testing data, this indicates that there is over fitting which the CART models are highly prone to.

Confusion Matrix for Training Data and Testing Data.

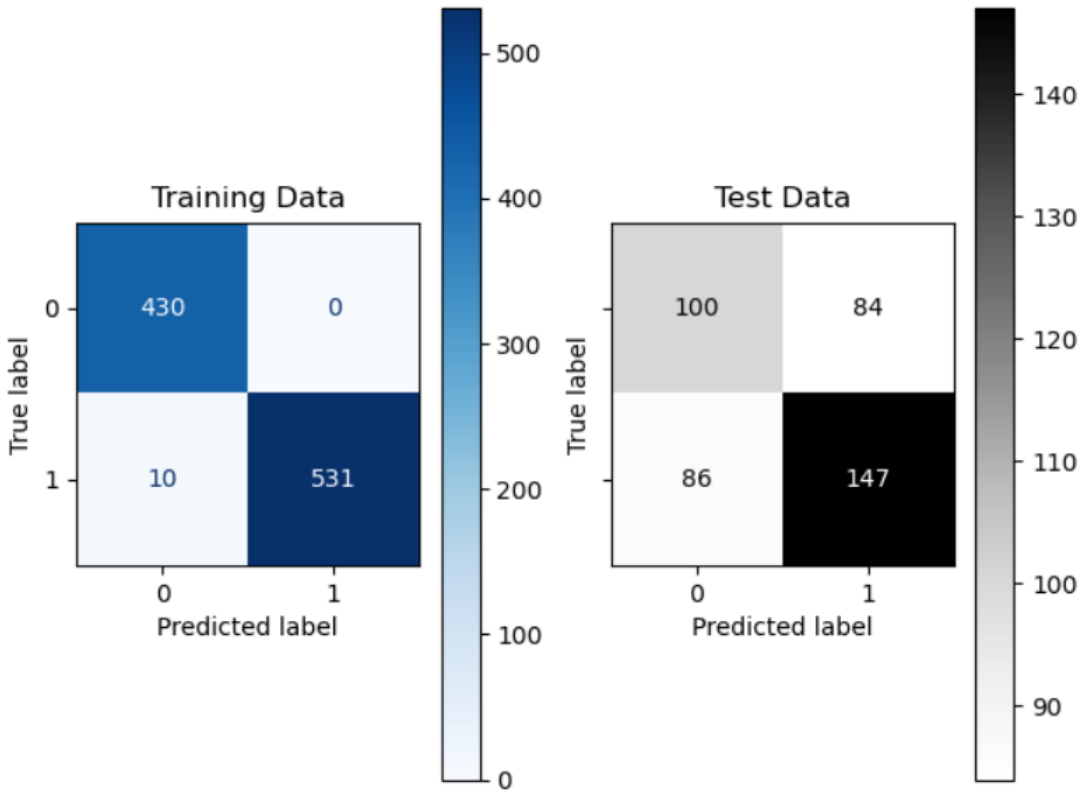

Fig 14 Confusion Matrix CART

AUC Score for the Training Data: 1.000
AUC Score for the Test Data: 0.590

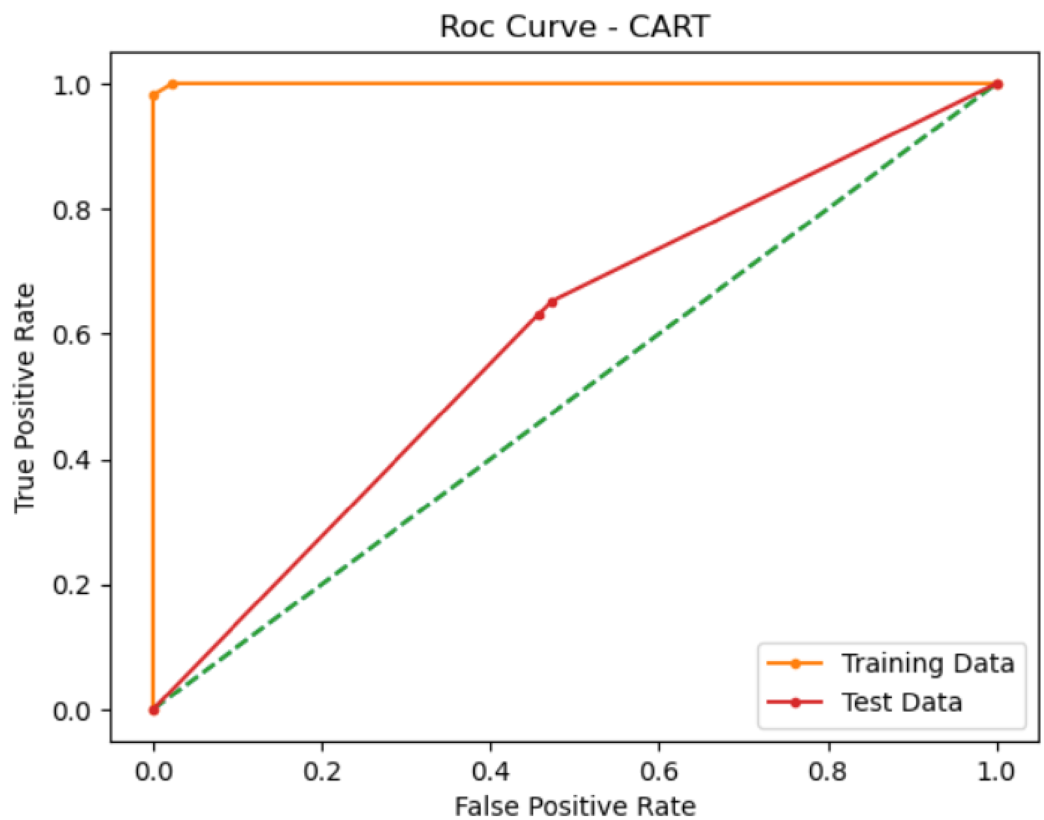

Fig 15 ROC Curve CART

Optimized Model

The accuracy of the optimized CART model is 0.70 for training data and 0.68 for test data.
The below is the confusion matrix

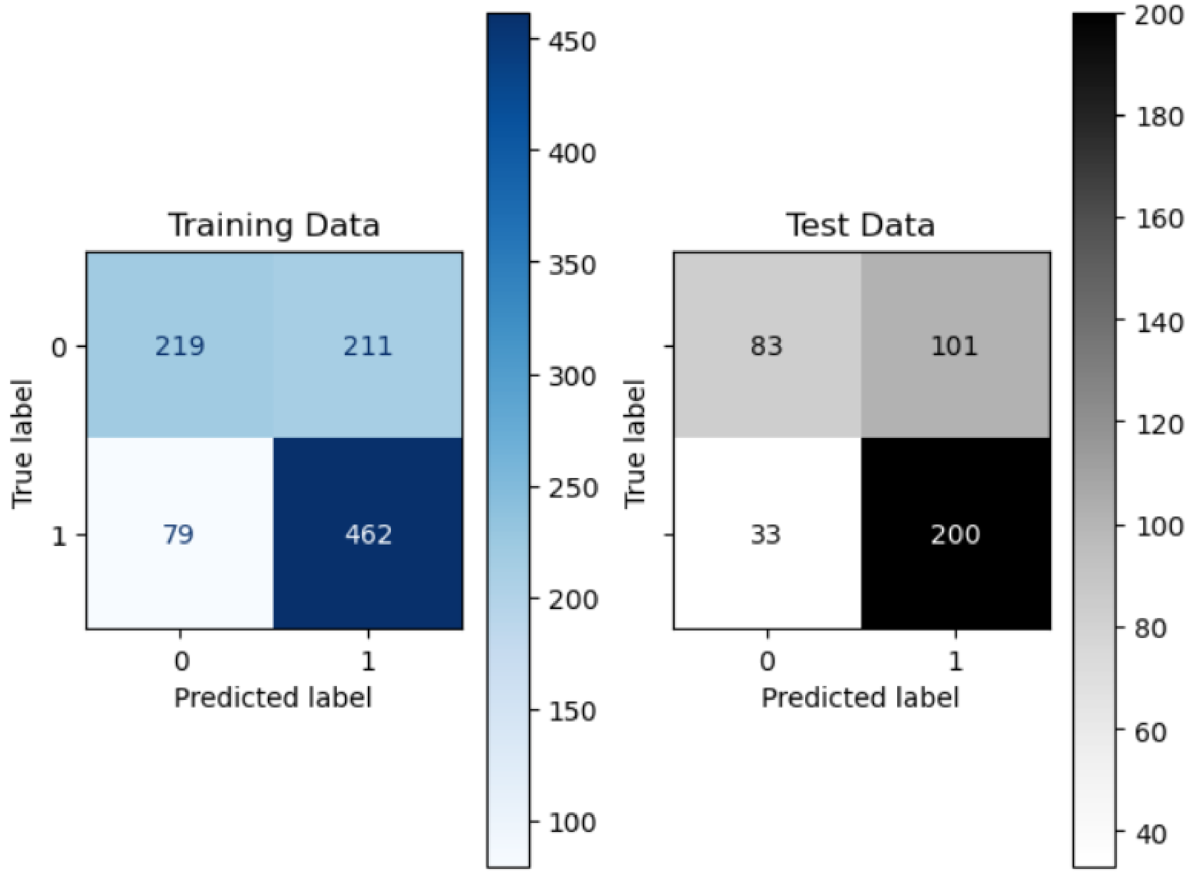

Fig 16 Confusion Matrix Optimized Model

AUC Score for the Training Data: 0.756

AUC Score for the Test Data: 0.685

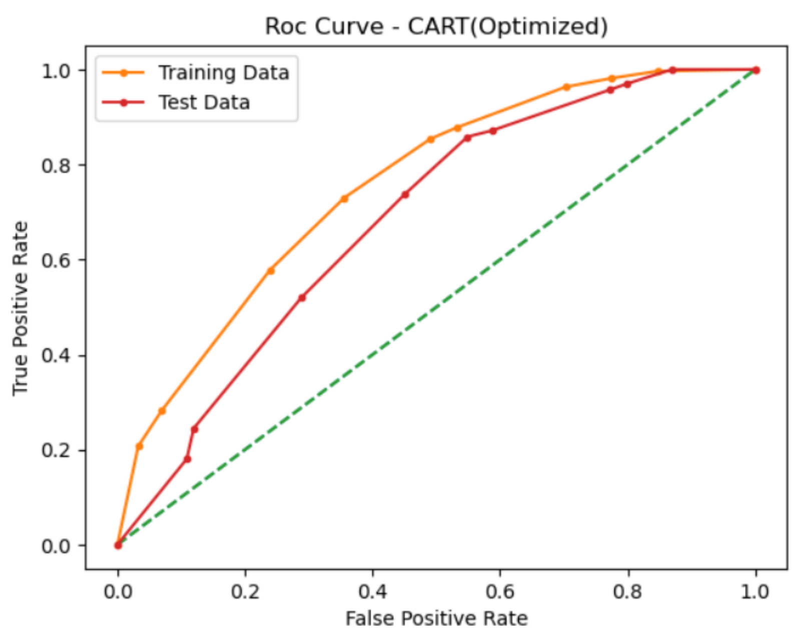

Fig 17 ROC Curve Optimized Model

The following are the feature importance of the optimized model

Wife_age	0.188133
Wife_education	0.208024
Husband_education	0.000000
No_of_children_born	0.477538
Wife_religion	0.000000
Wife_Working	0.000000
Husband_Occupation	0.000000
Standard_of_living_index	0.053954
Media_exposure	0.072350

As per the optimized model , No of children is the most important feature , followed by wife education and wife age.

The previous models had AUC score of 0.71 and 0.66 for train and test data respectively whereas the optimized CART model has AUC score of 0.75 and 0.68 for the train and test data respectively. And the higher AUC ROC score the better the model is in classifying, hence we choose this.

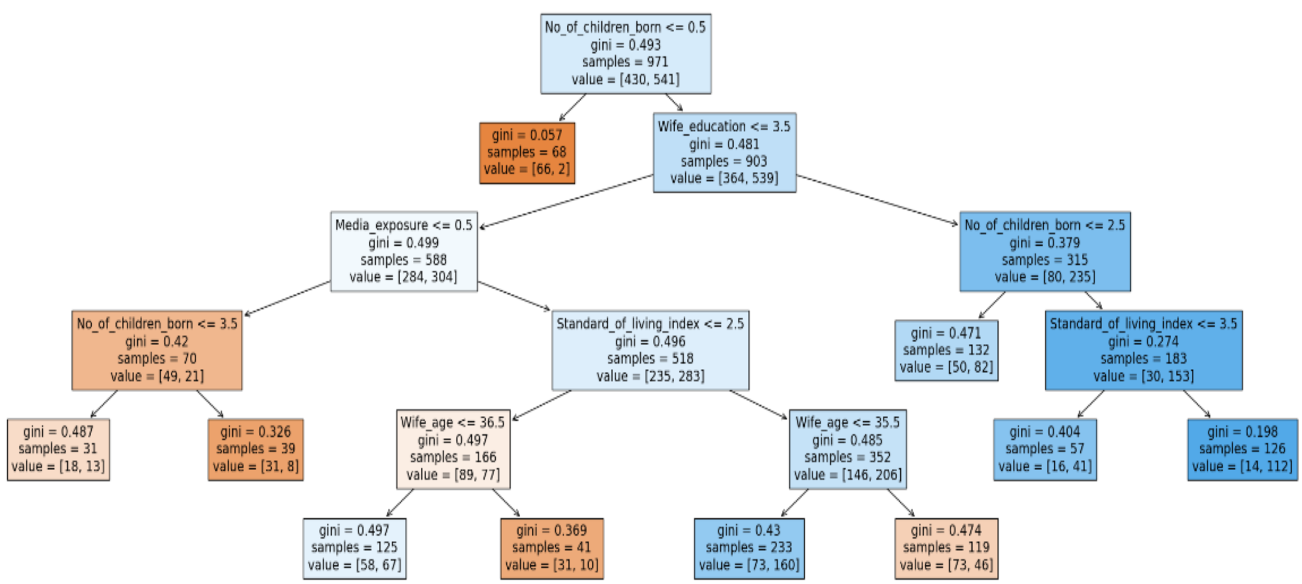

Fig 18 Tree Classifier

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

1. Our goal is to classify which demographic of people are likely to use contraceptive methods and which do not.
2. The data is initially sanitized and outliers are treated as they can significantly affect the logistic regression model, but CART model is immune to the presence of outliers.
3. When the model is being run, we ensure that all the predictor variables and dependent variable in float or int data type , as the model cannot , take direct string values and hence needs to be encoded
4. We split the data into training and testing data in 70:30 ratio, and pass Stratify as the dependent variable so the ratio of number of classes of the dependent variable is maintained , we train the model on train data and predict on the test data.
5. The Optimized Model states that Number of Children born is the most strong indicator of whether contraceptive methods are used or not
6. As we had seen the the Bivariate Analysis , the Wife education level tends to dictate the number of children born and wife age when its in 40s due , to fertility issues , menopause etc the number of households that use contraceptive decrease.