Predictive Modeling Project

by Anisha Sharma

PGPDSBA.O.Mar23.A

Great Learning

CONTENT:

Problem 1: Linear Regression

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.

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.

Problem 2: Logistic Regression, LDA and Cart

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.

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

Problem1:

1.1 Read the data and do exploratory analysis. Describe data briefly (checking datatype, EDA, 5 point summary). Perform Univariate, Bivariate and Multivariate analysis.

Solution:

- Below are the first five rows of dataset.

lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgscan	atch	pgin	ppgin	pflt	vflt	runqsz	freemem	freeswap	usr
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	95
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	97
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	87
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	98
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	90

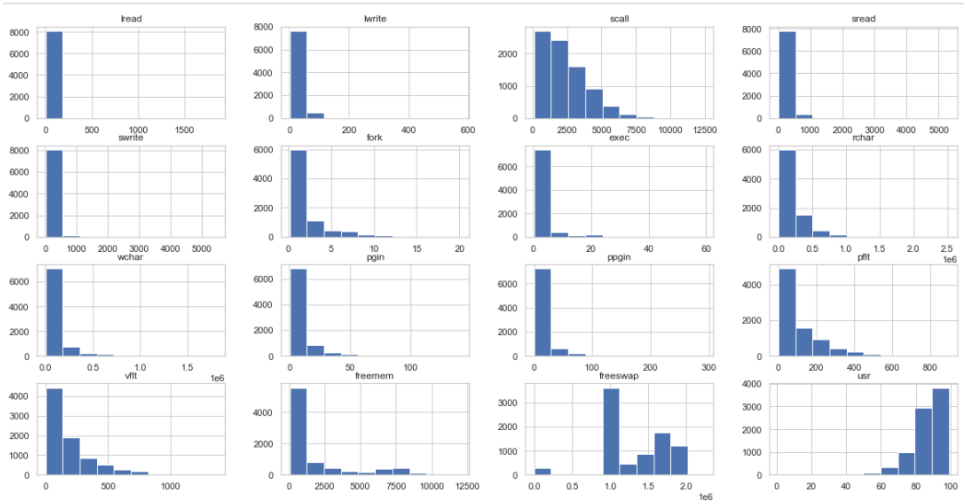
- The last five rows of the dataset.

lread	lwrite	scall	sread	swrite	fork	exec	rchar	wchar	pgout	...	pgscan	atch	pgin	ppgin	pflt	vflt	runqsz	freemem	freeswap	usr
16	12	3009	360	244	1.6	5.81	405250.0	85282.0	8.02	...	55.11	0.6	35.87	47.90	139.28	270.74	CPU_Bound	387	986647	80
4	0	1596	170	146	2.4	1.80	89489.0	41764.0	3.80	...	0.20	0.8	3.80	4.40	122.40	212.60	Not_CPU_Bound	263	1055742	90
16	5	3116	289	190	0.6	0.60	325948.0	52640.0	0.40	...	0.00	0.4	28.40	45.20	60.20	219.80	Not_CPU_Bound	400	969106	87
32	45	5180	254	179	1.2	1.20	62571.0	29505.0	1.40	...	18.04	0.4	23.05	24.25	93.19	202.81	CPU_Bound	141	1022458	83
2	0	985	55	46	1.6	4.80	111111.0	22256.0	0.00	...	0.00	0.2	3.40	6.20	91.80	110.00	CPU_Bound	659	1756514	94

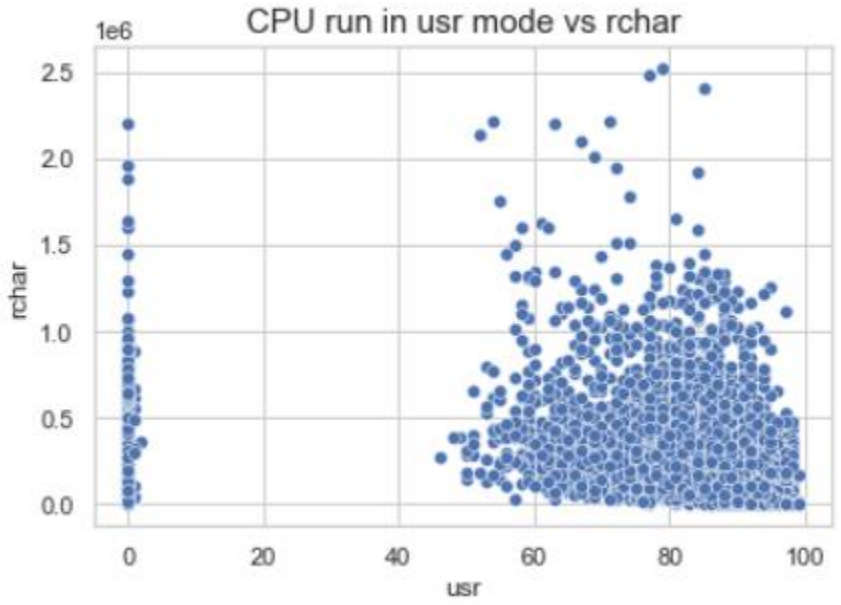
- Data summary.

```
<class 'pandas.core.frame.DataFrame'>
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)
memory usage: 1.4+ MB
```

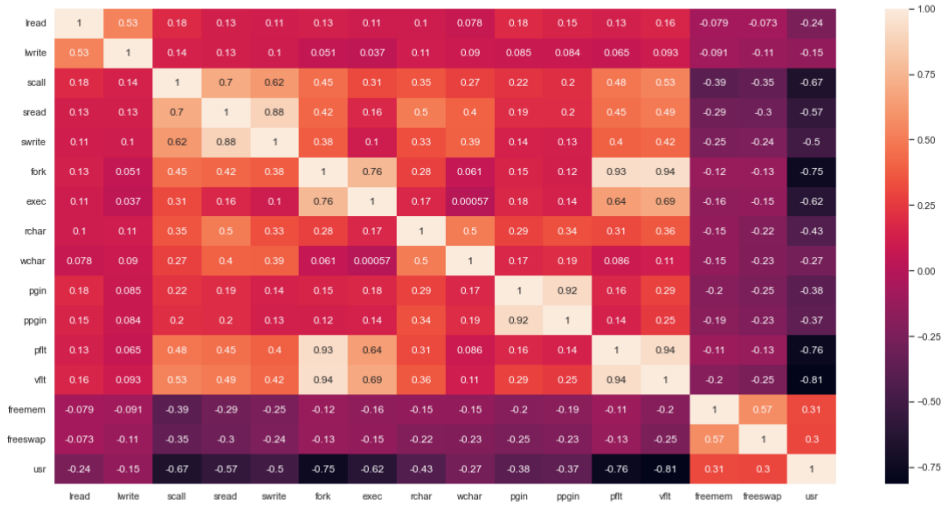
- Univariate Analysis



- Bivariate analysis



- Multivariate analysis



Insights:

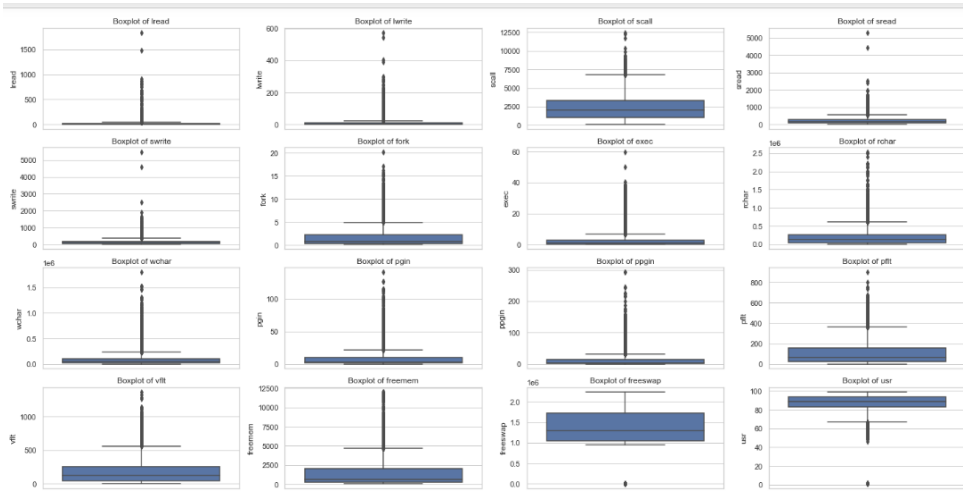
1. There are 22 column of variables in the dataset and 8192 rows.
2. There are no duplicate values in all the columns.
3. There is only one categorical variable runqsz in the dataset.
4. Outliers are present in the all variable.
5. There are null values present in the column rchar and wchar.
6. Dependent variable is usr.

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.

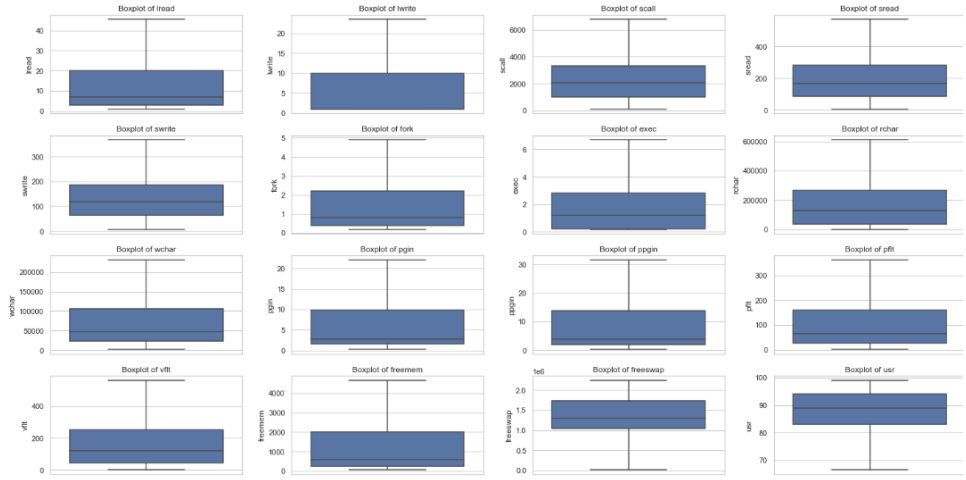
Solution: There were null values present in the dataset which need to be treated and also have zero values in the column

which is just unnecessary so it was required to treat them. There were no duplicate values in the dataset. There were outliers present in the every column which is needed to be treated.

This is graph before treating the outliers.



After treating outliers:



Below are insights of observation.

Insights

1. The column varible pgout, ppgout, pgfree, pgscan, atch have more than 50%, 0 values so we drop the column. The

remaining column having 0 values, we replace it with median value of the column.

2. Treat the variable columns having minimum value 0 with replace value with the median values.

3. We treated the null values with the median values.

4. There are outliers present in the every column so it is neccessary to treat them all.

5. Have checked the collinearity among all the variable using pairplot, done univariate analysis to analyse each and every column and done multivariate analysis usig heatmap.

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.

Solution:

Coefficients of each of the independent attribute.

```
The coefficient for lread is -0.011900566545768654
The coefficient for lwrite is -0.0023821393568151206
The coefficient for scall is -0.0014051923741018773
The coefficient for sread is 0.0006411282096692113
The coefficient for swrite is -0.003931752249884447
The coefficient for fork is 0.2043696725745215
The coefficient for exec is -0.3241911778899859
The coefficient for rchar is -1.196513414988441e-06
The coefficient for wchar is -5.135969681179832e-06
The coefficient for pgin is -0.013921553275935333
The coefficient for ppgin is -0.05295469120861869
The coefficient for pflt is -0.017642650641656817
The coefficient for vflt is -0.015474555520647732
The coefficient for freemem is 0.0002126116368606702
The coefficient for freeswap is -3.6692520047972195e-07
The coefficient for runqsz_Not_CPU_Bound is -0.12111793824531517
```

	usr	predictive	residual
3894	95	95.252590	-0.252590
4276	95	94.830537	0.169463
3414	89	90.495571	-1.495571
4165	80	79.370512	0.629488
7385	79	81.698794	-2.698794

Linear regression model

Dep. Variable:	usr	**R-squared:**	0.799
Model:	OLS	**Adj. R-squared:**	0.799
Method:	Least Squares	**F-statistic:**	1421.
Date:	Sat, 29 Jul 2023	**Prob (F-statistic):**	0.00
Time:	15:44:45	**Log-Likelihood:**	-16397.
No. Observations:	5734	**AIC:**	3.283e+04
Df Residuals:	5717	**BIC:**	3.294e+04
Df Model:	16		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	98.0406	0.288	340.806	0.000	97.477	98.605
lread	-0.0119	0.001	-9.628	0.000	-0.014	-0.009
lwrite	-0.0024	0.002	-1.009	0.313	-0.007	0.002
scall	-0.0014	5.26e-05	-26.695	0.000	-0.002	-0.001
sread	0.0006	0.001	0.901	0.367	-0.001	0.002
swrite	-0.0039	0.001	-5.035	0.000	-0.005	-0.002
fork	0.2044	0.094	2.185	0.029	0.021	0.388
exec	-0.3242	0.019	-17.451	0.000	-0.361	-0.288
rchar	-1.197e-06	3.23e-07	-3.709	0.000	-1.83e-06	-5.64e-07
wchar	-5.136e-06	4.91e-07	-10.453	0.000	-6.1e-06	-4.17e-06
pgin	-0.0139	0.011	-1.231	0.218	-0.036	0.008
ppgin	-0.0530	0.007	-7.685	0.000	-0.066	-0.039
pflt	-0.0176	0.002	-10.622	0.000	-0.021	-0.014
vflt	-0.0155	0.001	-12.458	0.000	-0.018	-0.013
freemem	0.0002	2.91e-05	7.319	0.000	0.000	0.000
freeswap	-3.669e-07	1.76e-07	-2.090	0.037	-7.11e-07	-2.27e-08
runqsz_Not_CPU_Bound	-0.1211	0.118	-1.025	0.306	-0.353	0.111

Omnibus:	9057.361	Durbin-Watson:	1.997
Prob(Omnibus):	0.000	Jarque-Bera (JB):	9167845.549
Skew:	-9.883	Prob(JB):	0.00
Kurtosis:	197.889	Cond. No.	7.39e+06

The intercept of our model is 98.04.

Insights:

1. The R-squared value tells us that our model can explain 79.9% of the variance in the training set.
2. Coefficients tells us how one unit charge in X can affect y. The sign of the coefficients indicates if the relationship is positive or negative. So we can observe from the analysis that there is strong multicollinearity between variables.
3. The RMSE training data having value of 4.223 and RMSE test data having value of 4.202.

4. Linear Regression equation is (98.0406) * const (-0.0119) * lread (-0.0024) * lwrite (-0.0014) * scall (0.0006) * sread (-0.0039) * swrite (0.2044) * fork (-0.3242) * exec (-0.0) * rchar (-0.0) * wchar (-0.0139) * pgin (-0.053) * ppgin (-0.0176) * pflt (-0.0155) * vflt (0.0002) * freemem (-0.0) * freeswap (-0.1211) * runqsz_Not_CPU_Bound

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

Solution:

Insights:

Based on the above predictions, we have observed that the variables have high significance to determine the linearity between variables and how this affect the independent variable.

Linear Regression equation is (98.0406) * const (-0.0119) * lread (-0.0024) * lwrite (-0.0014) * scall (0.0006) * sread (-0.0039) * swrite (0.2044) * fork (-0.3242) * exec (-0.0) * rchar (-0.0) * wchar (-0.0139) * pgin (-0.053) * ppgin (-0.0176) * pflt (-0.0155) * vflt (0.0002) * freemem (-0.0) * freeswap (-0.1211) * runqsz_Not_CPU_Bound

The high coefficient variables are comparative to others:

Sread having coefficient value 0.006

Freeem having coefficient value 0.0002

Fork having coefficient value 0.2043

And all other variables have negative coefficient which will affect this way that if the value of X is increased by one unit, the value of y is decreased by the given unit.

Recommendation: Go for the higher value of coefficiency of the variable that will be most important predictor.

Problem2:

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:

Below is the first five rows of the column.

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

Below is the last five rows of the column.

Wife_age	Wife_ education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_exposure	Co
33.0	Tertiary	Tertiary	NaN	Scientology	Yes	2	Very High	Exposed	
33.0	Tertiary	Tertiary	NaN	Scientology	No	1	Very High	Exposed	
39.0	Secondary	Secondary	NaN	Scientology	Yes	1	Very High	Exposed	
33.0	Secondary	Secondary	NaN	Scientology	Yes	2	Low	Exposed	
17.0	Secondary	Secondary	1.0	Scientology	No	2	Very High	Exposed	

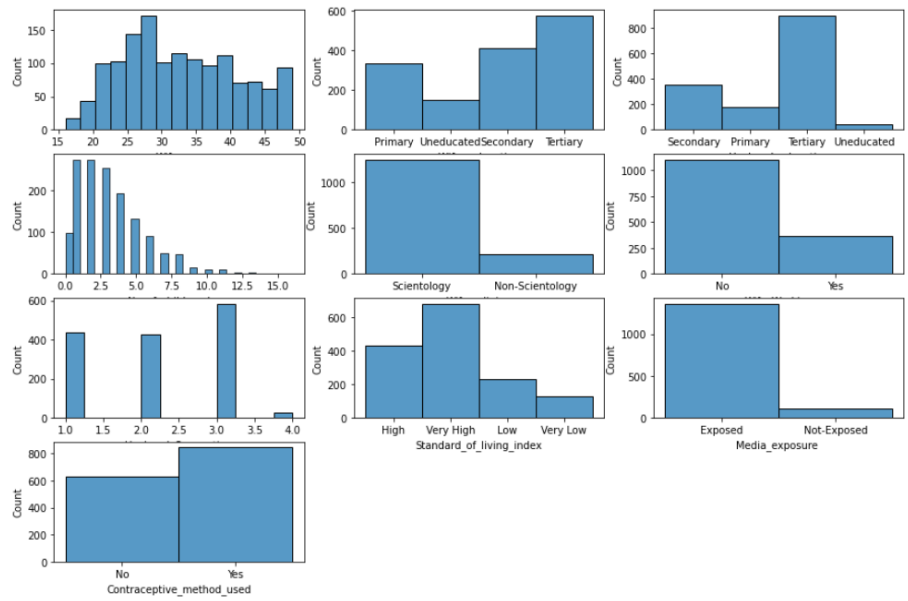
Data summary of the dataset.

```
<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
```

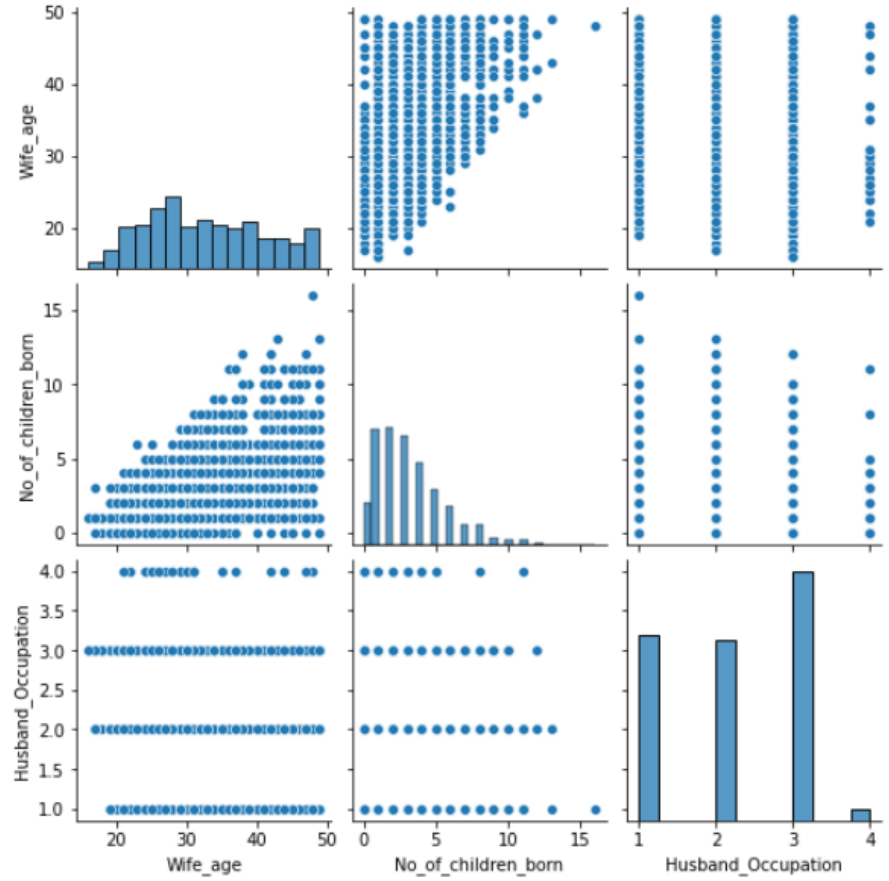
Description of the dataset.

	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

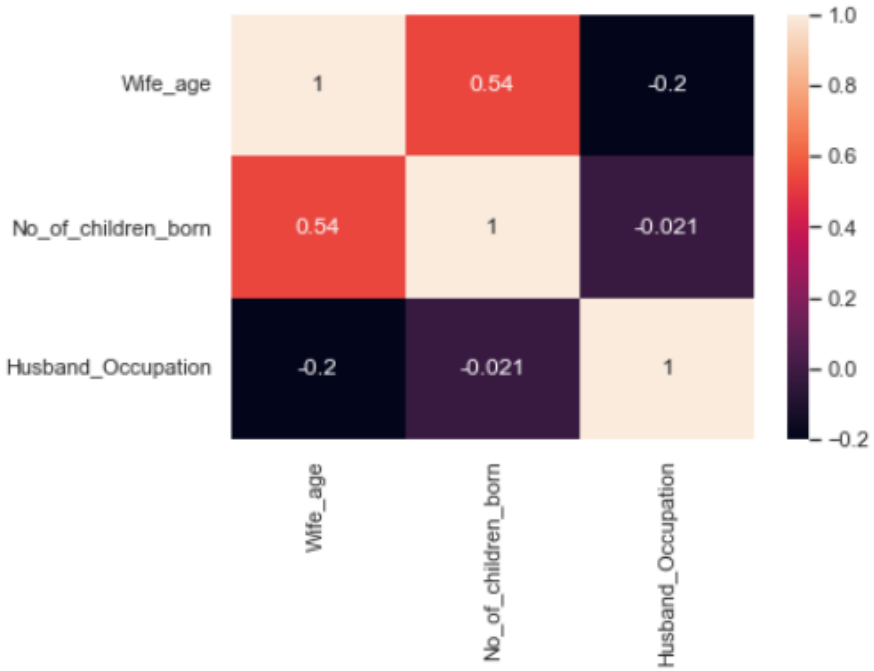
Univariate Analysis.



Bivariate Analysis.



Multivariate Analysis.



Insights:

- There are 1473 rows and 10 columns in the dataset.
- There is 1 int64, 2 float64 and 7 objects in the dataset.
- There were 80 duplicate values present in the dataset which is treated by dropping duplicate values.
- Null values were present in the column wife_age and no_of_children_born which are replaced by the median value of each column.
- There were outliers present in the column no_of_children_born.
- Have done univariate, bivariate and multivariate analysis on the dataset from which we have observed each variable to in order to analyse the variable.

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:

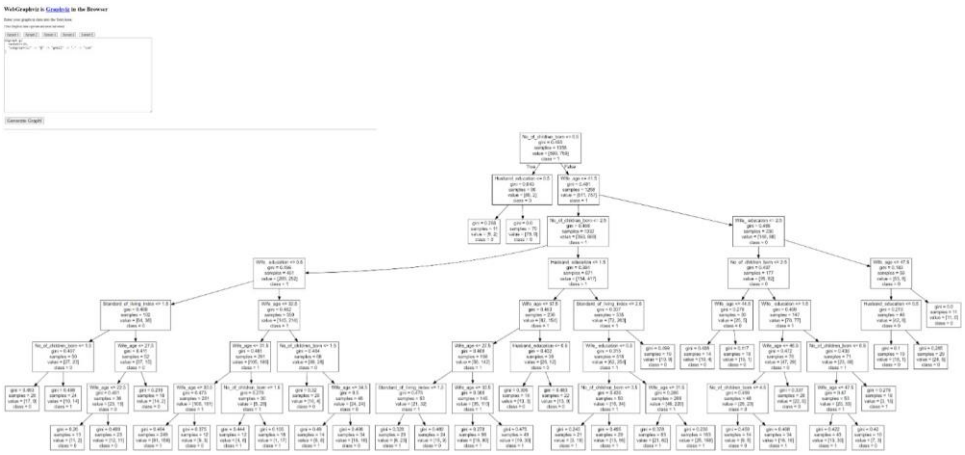
We have split the data into train and test and perform logistic regression, cart and linear discrimating analysis on the dataset.

Cart: while performing cart, we have take the independent variable contraceptive_method_used variable.

```
X_train (1358, 9)
X_test (30, 9)
train_labels (1358,)
test_labels (30,)
```

	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	0	1	3.0	1	0	2	0	0
1	45.0	3	1	10.0	1	0	3	2	0
2	43.0	0	1	7.0	1	0	3	2	0
3	42.0	1	0	9.0	1	0	3	0	0
4	36.0	1	1	8.0	1	0	3	1	0

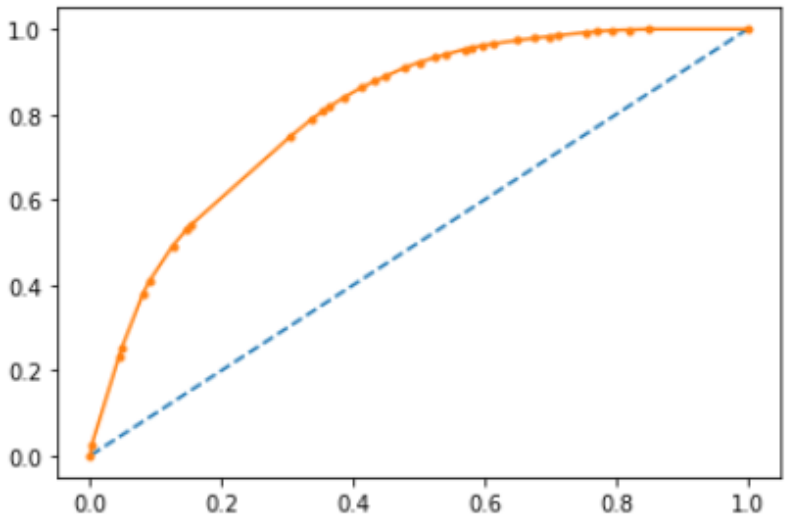
This is decision tree model of the dataset.



	Imp
Wife_age	0.323742
Wife_ education	0.093350
Husband_education	0.059879
No_of_children_born	0.217454
Wife_religion	0.030811
Wife_Working	0.062365
Husband_Occupation	0.086679
Standard_of_living_index	0.106234
Media_exposure	0.019487

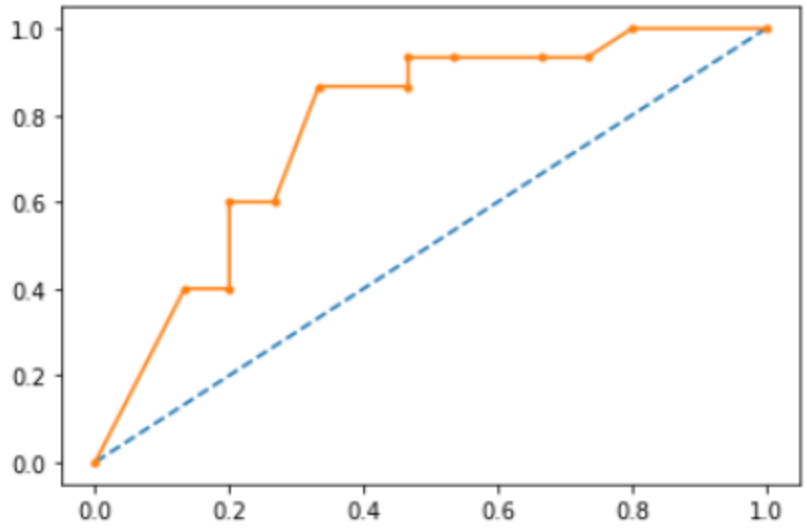
AUC for train data

AUC: 0.805



AUC for test data

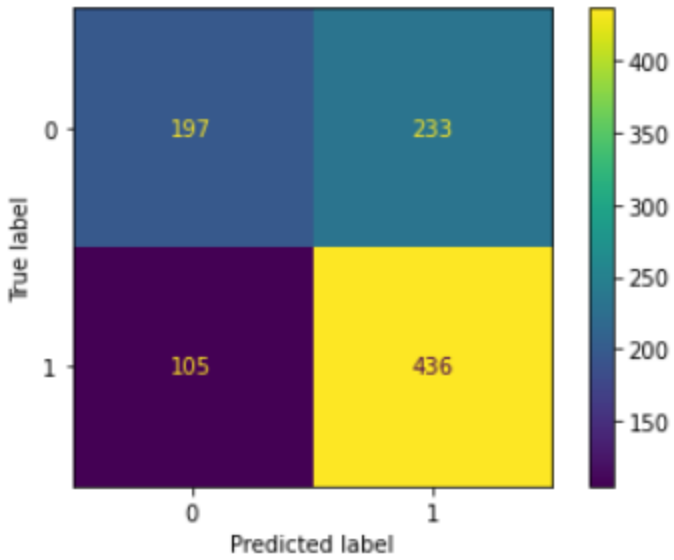
AUC: 0.771



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: The confusion matrix and classification report for testing data: The precision is having 64% and recall is having 46% for variable contraceptive_method_used for label 0.

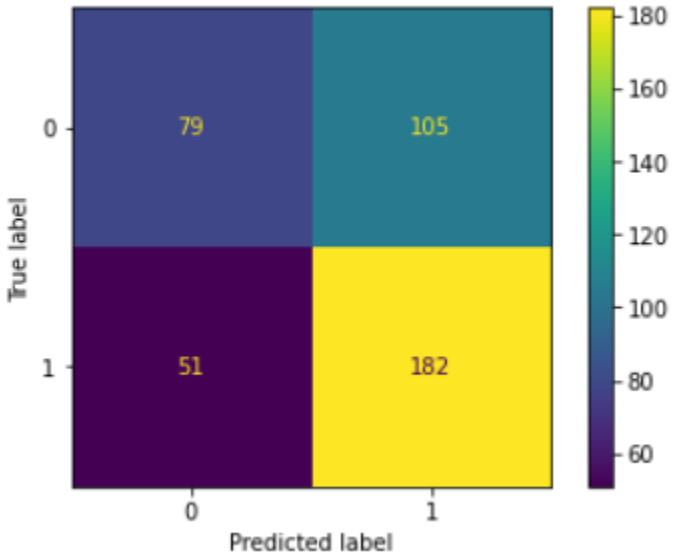
And for the label 1, it's precision is 65% and recall is having 80%.



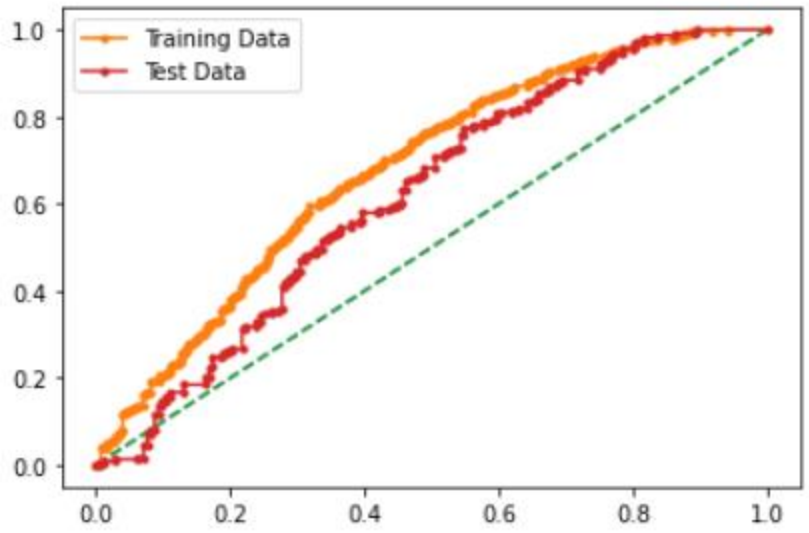
The confusion matrix and classification report for test data.

The precision is having 61% and recall is having 45% for variable contraceptive_method_used for label 0.

And for the label 1, it's precision is 64% and recall is having 77%.



AUC for the training data: 0.681
AUC for the test data: 0.622



Confusion matrix and classification report of the testing data.

```
[[ 82 102]
 [ 53 180]]
              precision    recall  f1-score   support

           0       0.61      0.45      0.51       184
           1       0.64      0.77      0.70       233

    accuracy                           0.63       417
   macro avg       0.62      0.61      0.61       417
weighted avg       0.62      0.63      0.62       417
```

Confusion matrix and classification report of the training data.

```
[[197 233]
 [110 431]]
              precision    recall  f1-score   support

           0       0.64      0.46      0.53       430
           1       0.65      0.80      0.72       541

    accuracy                           0.65       971
   macro avg       0.65      0.63      0.62       971
weighted avg       0.65      0.65      0.64       971
```