Murali B

PREDICTIVE MODELLING

**Contents : Page Number:**

**PROBLEM 1:**

**Table :**

**1.1 Top 5 data 2  
1.2 Bottom 5 data 2  
1.3 Check Info 3  
1.4 Describe Data 3  
1.5 Correlation 6**

**OUTPUT :**

**1.5 Shape of the data 4  
1.6 Check columns 4  
1.7 Check for Null values 4  
1.8 Check for duplicates 5  
1.9 Check for datatypes 5  
1.31 Check for Null values 19  
1.32 Treating Null values 20  
1.33 Check for Null values 21  
1.34 Check for Duplicate values 21  
1.37 Check for datatypes 23  
1.38 Rows and cols of test, train data 24  
1.39 Logistic Regression 25  
1.40 R-square of Train data 25  
1.41 R-square of test data 25  
1.42 Predictions on test data 25  
1.43 Prediction on train data 25  
1.44 RMSE of test data 25  
1.45 RMSE of train data 25  
1.46 Regression model co-efficient 25  
1.47 Model score of test data 26  
1.48 Model score of train data 26  
1.49 Regression model Intercept 26**

**FIGURES:**

**Univariate Analysis**

**1.11 7-8  
1.12 8  
1.13 9  
1.14 9  
1.15 10  
1.16 10**

**Bivariate Analysis**

**1.17 11  
1.18 12  
1.19 12  
1.20 12  
1.21 13  
1.22 13  
1.24 14  
1.25 15  
1.26 15  
1.27 16  
1.28 16  
1.29 17**

**Multivariate Analysis**

**1.30 18  
1.35 Before treating Outliers 22  
1.36 After Treating outliers 23  
1.50 Statsmodel 27**

**PROBLEM 2 :**

**Contents: Page number**

**2.1 Confusion matrix of train data 52  
2.8 Confusion matrix of test data 55**

**Figure:**

**2.2 Confusion matrix of test data 52  
2.3 Confusion matrix of test data 53  
2.4 Logistic Regression RUC curve-train 53  
2.5 Logistic Regression RUC curve-test 54  
2.6 LDA RUC curve-train 56  
2.16 LDA RUC curve-test 57  
2.17 CART RUC curve-train 57  
2.18 CART RUC curve-test 58  
2.19 Decision tree 59**

**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.

**Dataset for Problem 1:**[**compactiv.xlsx**](https://olympus.mygreatlearning.com/courses/81615/files/6885948/download?verifier=y4L3YBUgAO6XU52sk3CCwCaCcxoVXsuXTxrbiojD&wrap=1)

DATA DICTIONARY:  
-----------------------  
System measures used:

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 second  
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 might 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 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.**

**Top 5 data** :  
Graphical user interface, application

Description automatically generated  
Table 1.1   
  
**Bottom 5 data :**   
Graphical user interface, text, application

Description automatically generated  
Table 1.2

**Checking the Info :**   
Table

Description automatically generated  
Table 1.3

We can see 22 columns and in which 13 columns are float datatypes , 8 columns are Integer datatypes and 1 object column.

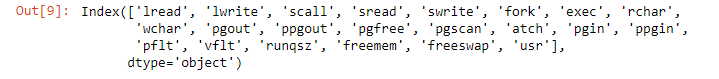
**Describe the data :**   
Graphical user interface

Description automatically generated with medium confidence  
Table 1.4  
  
It describes the 5 point summary such as count , mean , standard deviation , minimum value , 25th percentile , 50th percentile , 75th percentile and maximum value details.

**Shape of the data :**

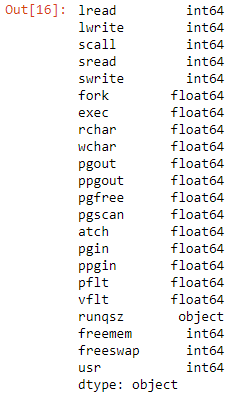
  
Output 1.5

We can see 22 columns with 8192 observations.

**Check the columns :**   
  
Output 1.6  
  
**Checking for Null values :**   
Table

Description automatically generated  
Output 1.7  
  
We can see 104 null values in rchar and wchar fields and no other null values present in other fields.   
  
**Checking for duplicate values:**   
  
Output 1.8

So we can see no duplicate values present in the dataset.

**Checking for datatypes :** **Output 1.9**It displays the datatypes of each fields.

**Check for correlation :** Graphical user interface, application, table, Excel

Description automatically generated **Table 1.10**

**Univariate Analysis :** Calendar

Description automatically generatedChart, waterfall chart

Description automatically generated **Fig 1.11**

The histogram displays the univariate analysis of each, and all datasets present in the dataframe. We can see linear representation in each graph.

Chart, bar chart

Description automatically generated  
Fig 1.12  
Here in runqsz field, we can see two unique values such as CPU\_Bound and Not\_CPU\_Bound. We can see the above plot representation using countplot.

Chart

Description automatically generated with low confidence  
Fig 1.13  
  
So using countplot, we can see the pictorial graph representation of lread component and the data seems to be very compressed from 0 to 1000.

Graphical user interface

Description automatically generated with medium confidence  
Fig 1.14

The above plot using countplot, we can see that lwrite component data is very minimal found within the range between 1500.

Chart, histogram

Description automatically generated  
Fig 1.15

Here scall data is scattered across all various points between the count of 0 to 8 and the data are plotted with the help of countplot.   
  
Chart, histogram

Description automatically generated  
Fig 1.16

Here sread data is represented across all various points linearly between the count of 0 to 40 and the data are plotted with the help of countplot.

**Bivariate Analysis :**

Chart

Description automatically generated **Fig 1.17**

Here we have taken lread values and represented through the help of histplot and boxplot. From the graph, we can see datapoints lying between 0 to1250.

Chart

Description automatically generated **Fig 1.18**

In this, lwrite data points is being represented using histplot and boxplot and we can see data ranging from 0 to 2750.

Chart, histogram

Description automatically generated  
Fig 1.19

Here we can see clear datapoints of scall datapoints through boxplot and histplot. The data range lies between 0 to 8000.   
  
Chart

Description automatically generated  
Fig 1.20  
  
Here we can see the representation of datapoints of sread datapoints through boxplot and histplot. The data range lies between 0 to 500.

Chart, scatter chart

Description automatically generated  
Fig 1.21

Through the help of scatterplot, we can see the relationship of lread and lwrite components and more datapoints are lying between 0 to 1000.

Chart, scatter chart

Description automatically generated  
Fig 1.22

With the help of scatterplot, we can see the relationship of scall and sread components and more datapoints are lying between 0 to 10000.

Chart, scatter chart

Description automatically generated   
Fig 1.23

Using the help of scatterplot, we can see the relationship of swrite and fork components and more datapoints are lying between 0 to 2000.

Chart, scatter chart

Description automatically generated  
Fig 1.24

Through the help of scatterplot, we can see the relationship of exec and rchar components and more datapoints are lying between 0 to 40.

A picture containing graphical user interface

Description automatically generated  
Fig 1.25  
  
From the above boxplot, we can see discrete representation of scall values across sread values in a linear way. It covers most part of the all the datapoints.

A picture containing chart

Description automatically generated   
Fig 1.26

From the above boxplot, we can see curve type representation of lread values across lwrite values. The datapoints are plotted in a curve fashioned way across each values.

Chart

Description automatically generated  
Fig 1.27

Here we can see the datapoints are hugely crowded when plotted across exec and rchar values with the help of boxplot.

Chart

Description automatically generated  
Fig 1.28

From the above boxplot, we can see the data are roughly organized and scattered across all the data values when plotted swrite again fork.

**Heatmap :** Chart

Description automatically generated with medium confidence **Fig 1.29**

The above heatmap depicts that strong correlation relationship between each datasets and we can see strong relationship ranging within 0.6 to 0.94.

**Multivariate Analysis :** A picture containing scatter chart

Description automatically generatedTable

Description automatically generated with low confidence **Fig 1.30**

The above plot is represented using the help of pairplot which shows multivariate analysis of each and every components present in the dataset.

**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.**

**Checking for Null values:**

Table

Description automatically generated  
Output 1.31

We can see there are 104 null values present in rchar and 15 null values present in wchar.

**Treating Null values :**

So we have imputed the Null values with the help of mean values.

After treating the Null values then checking for Null values :

Table

Description automatically generated  
Output 1.32

We can still 3 Null values in rchar.

**Treating Null Values :**   
  
We have filled the Null values with 0 in rchar.

**Checking for Null values:**   
Table

Description automatically generated  
Output 1.33

Now, we can see no null values present in the dataset.

**Checking for duplicates :**

 **Output 1.34**

We can see no duplicate values present in the dataframe.

**Checking for outliers :** Chart

Description automatically generated

Chart

Description automatically generatedChart, box and whisker chart

Description automatically generated **Fig 1.35**

We can see outliers present in almost all the datasets present in the dataframe.

So we have treated outliers and checking for outliers :

Chart, bar chart, box and whisker chart

Description automatically generated

Chart, waterfall chart

Description automatically generated Chart

Description automatically generated  
Fig 1.36

We have tretaed the outliers with the help of boxplot and we can see now no outliers present in the dataset.

**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.**

**Encode the data (having String values) for Modelling :**

So we have converted the runqsz field object datatype to int datatype.

Checking for datatypes after converting:   
Table

Description automatically generated  
Output 1.37  
Now, we can see runqsz has been converted to int datatype.

**Split the data into train and test (70:30) :**

Text

Description automatically generated **Output 1.38**

The above screenshot shows the number of rows present in the test and train set.

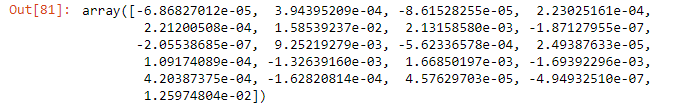
**Apply LinearRegression :**   
  
Output 1.39

**R square on training data :**   
  
Output 1.40

**R square on test data :** **Output 1.41**

**Predictions on Test data :** **Output 1.42**

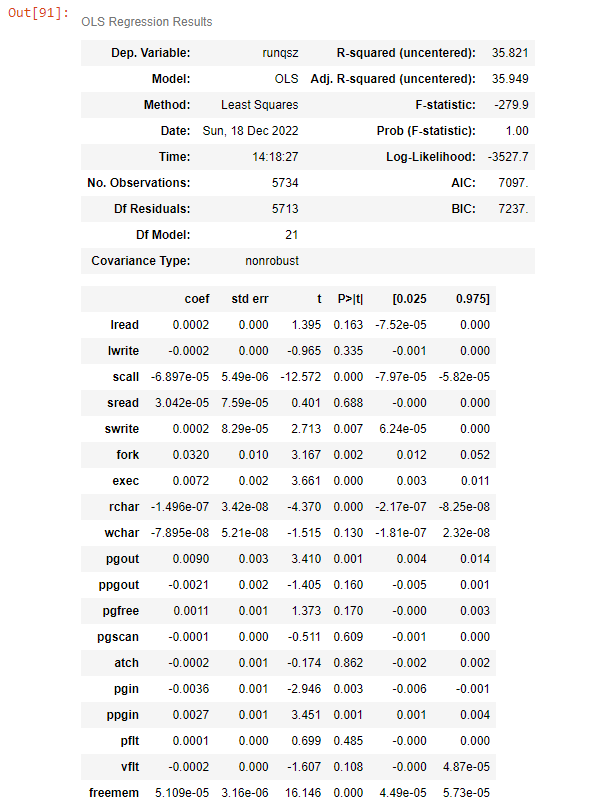
**Predictions on Train data :** **Output 1.43  
  
Root Mean Squared Error of Test data :** **Output 1.44  
  
Root Mean Squared Error of train data :**

**Output 1.45  
  
Regression model coefficient :**

**Output 1.46**

**Model score on Test data :** **Output 1.47**

**Model score on Train data :** **Output 1.48  
  
regression model intercept :** **Output 1.49**

**Perform checks for significant variables using appropriate method form statsmodel :**Graphical user interface, text

Description automatically generated with medium confidence **Fig 1.50**We can see R Square is computed without centering (uncentered) since the model does not contain a constant. Standard Errors assume that the covariance matrix of the errors is correctly specified. The condition number is large ,2.42e+06. This might indicate that there are strong multicollinearity or other numerical problems.

**Variation of independent variable :**   
Output 1.51  
  
**Coefficient determination R square on Train set :**   
Output 1.52

**Coefficient determination R square on Test set :**  **Output 1.53**

**VIF output :**Text

Description automatically generated **Output 1.54**

**We can see test data model produces good results compared to train data.**

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

**Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present.**

**From all the analysis made, we can see prediction of test data is more than train data and R-square, RMSE , model score and co-efficient values of test data is more so training data is less efficient when compared so we can choose test data and based on the test data outcome we can modify some components in test data to yield even more between results than train data.**

**Here, we did exploratory data analysis , checked and treated null values, duplicate values and outliers. We encode the data for modelling and did split of data into train and test 70:30. We performed statsmodel check and checked the performance of tran and test model through raquare, RMSE and Adj Rsquare. We can conclude that test data model produces good results compared to train data in all the performance metrics.**

**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.**

**Dataset for Problem 2: Contraceptive\_method\_dataset.xlsx**

**Data Dictionary:**

**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=verlow, 2, 3, 4=high**

**9. Media exposure (binary) Good, Not good**

**10. Contraceptive method used (class attribute) No,Yes**

**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.**

**Checking the top 5 data :**Graphical user interface

Description automatically generated **Table 1.55**

**Checking the bottom 5 data :** A screenshot of a computer

Description automatically generated with medium confidence **Table 1.56**

**Checking the Info :**Text

Description automatically generated **Output 1.57**

**We can see that there are 2 datasets having float datatype and 1 integer datatype and 7 object datatypes.**

**Describe the data :** Table

Description automatically generated **Table 1.58**

**Here, we can see data description of the dataframe such as count, mean , standard deviation , minimum , 25th percentile, 50th percentile, 75th percentile , maximum values.**

**Shape of the data :** **Output 1.59  
We can see there are 10 columns with 1473 observations.**

**Checking for Null values :**

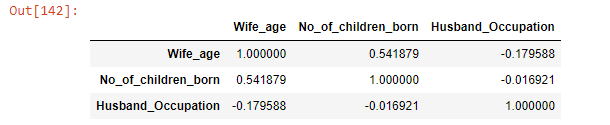
Text

Description automatically generated **Output 1.60**

**We can see 71 Null values in Wife\_age dataset and 21 Null values present in No\_of\_children\_born dataset.**

**Checking for datatypes :**Text

Description automatically generated **Output 1.61**

**Checking for correlation :**  **Output 1.62**

**We can see three datasets correlation details.**

**Treating Null values**

**So we have treated Null values with mean values and let’s check for Null values :**

Text

Description automatically generated **Output 1.63**

**We still 21 Null values present in No\_of\_children\_born dataset.**

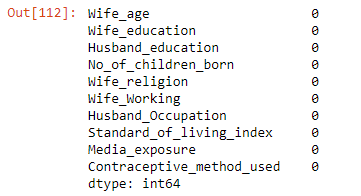
**Renaming the column names and checking the columns details :**Word

Description automatically generated with medium confidence **Output 1.64**

**We have renamed column names of Media\_exposure and Wife\_education.**

**Treating Null values:**

**So we have treated remaining Null values with the help of filling it with 0 and checking for Null values:**

 **Output 1.65**

**Now, we can see no Null values present in the dataset.**

**Checking for duplicates :**Table

Description automatically generated **Output 1.66  
We can see 80 duplicate values.**

**Treating duplicate values :** **Output 1.67**

**So no duplicate values present in the dataset.**

**Checking for Outliers :**Chart, histogram

Description automatically generated **Fig 1.68**

**We can see outliers present in No\_of\_children\_born dataset.**

**Treating outliers and checking for Outliers:** Chart, box and whisker chart

Description automatically generated **Fig 1.69**

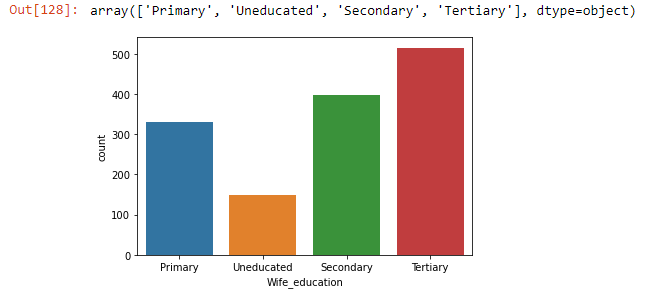
**We have treated the outliers with the help of boxplot and we can see no outliers present in the dataset.**

**Univariate Analysis :** Chart, histogram

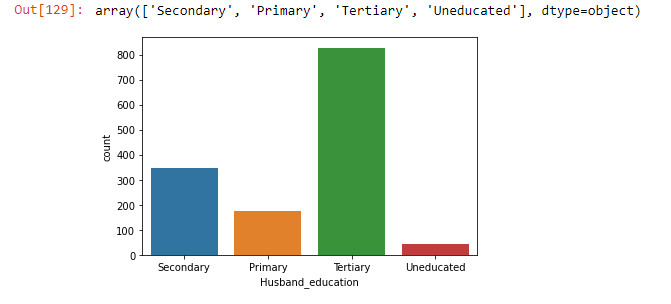
Description automatically generated Chart

Description automatically generated  
Fig 1.70

We can see univariate analysis of Wife\_age , No\_of\_Children\_born , Husband\_occupation with the help of histplot and there is a linear represnetation in each datasets.

**Bivariate Analysis :**  **Fig 1.71**

We can see 4 unique values present in Wife\_education dataset such as Primary , Uneducated , Secondary and Tertiary and we can see Tertairy values is more in the dataset.

  
Fig 1.72

We can see 4 unique values present in Husband\_education dataset such as Primary , Uneducated , Secondary and Tertiary and we can see Tertairy values is more in the dataset.

Chart, bar chart

Description automatically generated  
Fig 1.73

We can see Scientology values is more than Non-Scientology values in Wife\_religion dataset.

Chart, bar chart

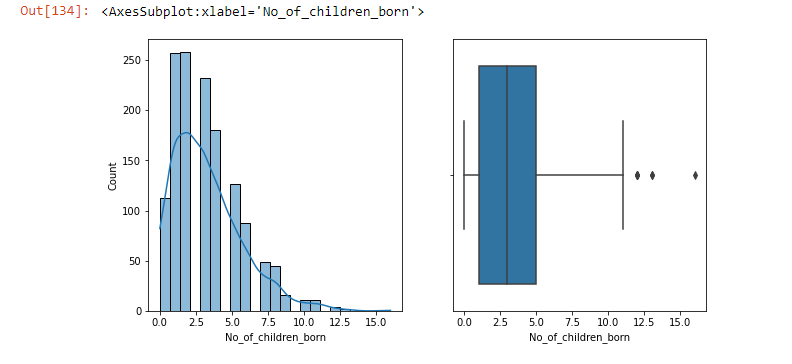
Description automatically generated  
Fig 1.74

We can see working of wife numbers is more than wife who are working in Wife\_Working dataset.

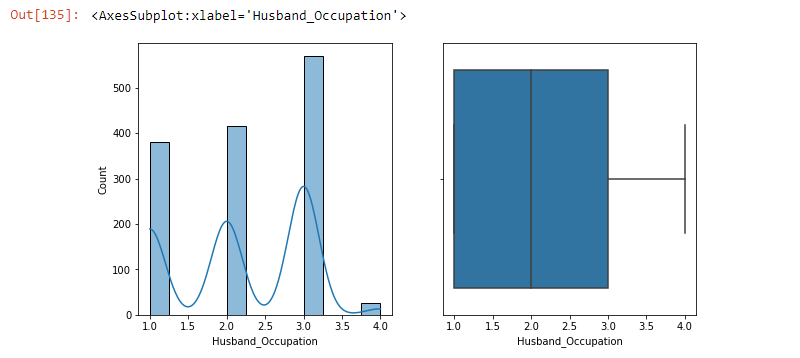
Chart, histogram

Description automatically generated  
Fig 1.75

With the help of boxplot and histplot, we can see representation of Wife\_age as upside curve down with non-linear represenation.

  
Fig 1.76

Here, we can see the values are steep raised initially and then falls apart when we see the datapoints shown in No\_of\_Children\_born using histplot.

  
Fig 1.77

We see here curve shaped linear representation data formation of Husband\_Occupation datasets with the help of histplot.

Chart

Description automatically generated with medium confidence  
Fig 1.78

Using scatterplot, we can see straight line relationship between Wife\_age and Wife\_education using scatterplot.

Chart

Description automatically generated  
Fig 1.79

Here we can see, linear way of data representation between Husband\_occupation and No\_of\_children\_born through scatterplot.

Chart, scatter chart

Description automatically generated  
Fig 1.80

The datapoints are not properly segregated when plotted Husband\_occupation across standard\_of\_living\_index using scatterplot.

Chart, box and whisker chart

Description automatically generated  
Fig 1.81

Using boxplot, we can see relationship between Wife\_age and Wife\_education and Uneducated values is more.

Chart, box and whisker chart

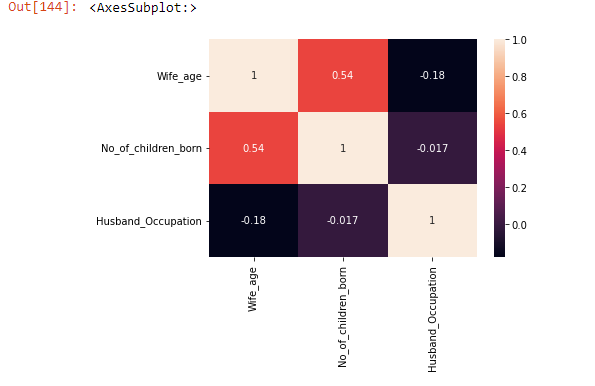
Description automatically generated  
Fig 1.82

Through the help of boxplot, we can see relationship between Husband\_education and No\_of\_children\_born and Uneducated values is more.

Chart, bar chart

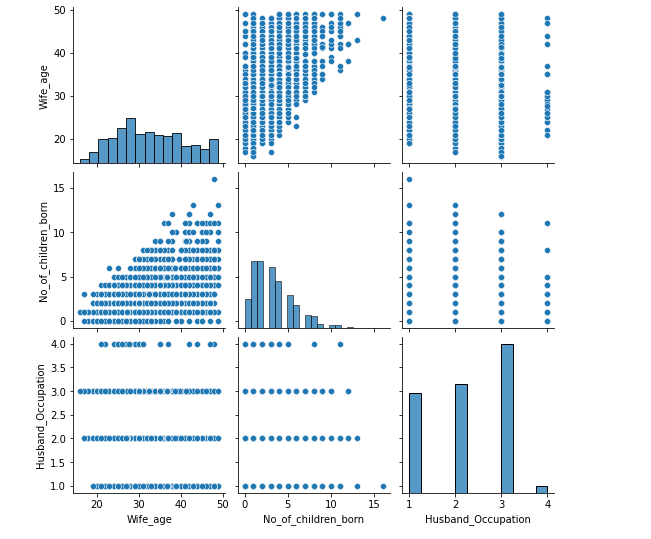
Description automatically generated  
Fig 1.83

From the above screenshot, we can see when plotted Husband\_Occupation along with Standard\_of\_living\_index there shows values ranging between 2.0 to 3.0 and are equally segregrated.

  
Fig 1.84

From the above heatmap, we can see there is strong relationship between Wife\_age and No\_of\_children\_born ranging between values 0.54.

**Multivariate Analysis :**

 **Fig 1.85**

**We can see multivariate analysis of wife\_age , no\_of\_children\_born and husband\_occupation using pairplot.**

**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.**

**Encode the data (having string values) for Modelling :**

**We have converted the object datatypes to float and int datatypes.**

Text

Description automatically generated **Output 1.86**

**Data Split: Split the data into train and test (70:30).**

**LogisticRegression :**

 **Output 1.87**

 **Output 1.88**

**LDA :**

 **Output 1.89**

**Coefficient values :**  **Output 1.90**

**Intercept values:**

 **Output 1.91** **Output 1.92**

**Probabilities :**Graphical user interface, text, application

Description automatically generated **Table 1.93**

**CART :**

 **Output 1.94**

 **Output 1.95**

**Important Features :**

Chart, funnel chart

Description automatically generated **Fig 1.96**

**From the above plot, we can see Wife\_age fetaure has more important values.**

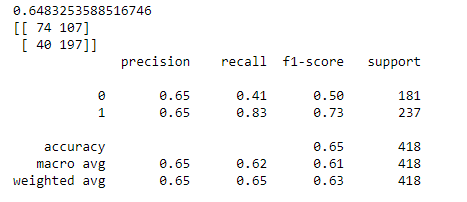
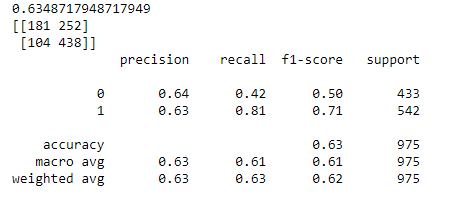
**2.3 Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.**

**Logistic Regression :**

**Accuracy score of Test data :** **Output 1.97**

**Accuracy score of Training data :**

 **Output 1.98**

**Confusion Matrix of test data:** **Output 1.99  
  
  
Confusion Matrix of train data:** **Output 2.1**

**Confusion matrix of train data :**Chart, treemap chart

Description automatically generated **Fig 2.2**

**Confusion matrix of test data :**

Chart, treemap chart

Description automatically generated **Fig 2.3**

**AUC values and ROC curve of train data :**Chart, line chart

Description automatically generated **Fig 2.4**

**AUC values and ROC curve of test data :**

Chart, line chart

Description automatically generated **Fig 2.5**

**LDA :**

**Accuracy score of test data :** **Output 2.6**

**Accuracy score of train data :** **Output 2.7**

**Confusion matrix of test data :**

Table

Description automatically generated **Output 2.8**

**Confusion matrix of train data :**

Table

Description automatically generated **Output 2.9**

**AUC score and ROC curve of train data:**Chart, line chart

Description automatically generated **Fig 2.10**

**AUC score and ROC curve of test data:**Chart, line chart

Description automatically generated **Fig 2.11**

**CART :   
  
Accuracy score of test data :** **Output 2.12**

**Accuracy of train data :** **Output 2.13**

**Confusion matrix of test data :**

Table

Description automatically generated **Output 2.14**

**Confusion matrix of train data:**

Table

Description automatically generated **Output 2.15**

**AUC score and ROC curve of train data:**Chart, line chart

Description automatically generated **Fig 2.16**

**AUC score and ROC curve of test data:**Chart, line chart

Description automatically generated **Fig 2.17**

**Decision Tree:** Diagram, engineering drawing

Description automatically generated  
Fig 2.18

Cart model is best and optimised model based on the various techniques tried on both test and train data.

**2.4 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.**

**We have done various descriptive analysis and drawn many infrences and build a model using Logistic regression , LDA and CART. We have treated Null values , checked for duplicate values and treated outliers. We have split the data to 70:30 , and did many data ingestion and performance metrics.**

**CART model is best suited since its prediction values, accuarcy values , confusion matrix values and ROC curve and AUC score values are better than other 2 models so we can still cutsomize the model in CART based on the business requirement.**