**Business Report**

Week 05 – Linear Regression, Logistic Regression, LDA, CART

Table of Contents

Contents

[Problem 01: 1](#_Toc119265678)

[Executive Summary 1](#_Toc119265679)

[Introduction 1](#_Toc119265680)

[Data Description 2](#_Toc119265681)

[Sample of the dataset: 2](#_Toc119265682)

[Q1: Exploratory Data Analysis 3](#_Toc119265683)

[Let us check the types of variables in the data frame. 3](#_Toc119265684)

[Let us check for the some of the statistics that can describe the data better . 3](#_Toc119265685)

[Q2: Check for missing values in the dataset: 4](#_Toc119265686)

[Treating the missing values in the dataset: 5](#_Toc119265687)

[Importance of the rows with zeros: 5](#_Toc119265688)

[Checking for Outliers: 6](#_Toc119265689)

[Treating the Outliers: 6](#_Toc119265690)

[Correlation Plot 7](#_Toc119265691)

[PairPlot 8](#_Toc119265692)

[Q3 & Q4: Linear Regression Analysis: 9](#_Toc119265693)

[Splitting the data and building the Linear Regression Model 9](#_Toc119265694)

[First Linear Fit Model: 9](#_Toc119265695)

[Second Linear Regression Model: 12](#_Toc119265696)

[Third Linear Regression Model: 13](#_Toc119265697)

[Conclusion: 15](#_Toc119265698)

[Linear Regression Equation for the above problem: 15](#_Toc119265699)

[Problem 02: 16](#_Toc119265700)

[Executive Summary 16](#_Toc119265701)

[Introduction 16](#_Toc119265702)

[Sample of the dataset: 16](#_Toc119265703)

[Q1: Exploratory Data Analysis 16](#_Toc119265704)

[Let us check the types of variables in the data frame. 16](#_Toc119265705)

[Let us check for the some of the statistics that can describe the data better . 17](#_Toc119265706)

[Check for missing values in the dataset: 17](#_Toc119265707)

[Treating the missing values in the dataset: 17](#_Toc119265708)

[Checking for Outliers: 18](#_Toc119265709)

[Treating the Outliers: 18](#_Toc119265710)

[Correlation Plot 19](#_Toc119265711)

[PairPlot 20](#_Toc119265712)

[Bivariate Analysis: 20](#_Toc119265713)

[Q2: Logistic Regression: 21](#_Toc119265714)

[Splitting the data and building the Logistic Regression Model and Confusion Matrix 21](#_Toc119265715)

[Confusion Matrix: 21](#_Toc119265716)

[Calculating some important Metrics: 21](#_Toc119265717)

[Q2 & Q3: Linear Discriminant Analysis (LDA) 22](#_Toc119265718)

## Problem 01:

## 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

The purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset using central tendency and other parameters. The data consists of 22 columns and 8192 rows. Out of 22 columns, one of the columns is “usr” (Portion of time (%) that CPUs run in user mode). In this case we analyse how each attribute affects the system to be in 'usr' mode using a list of system attributes and find out a linear equation to build a model to predict 'usr'

## Data Description

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

### Sample of the dataset:

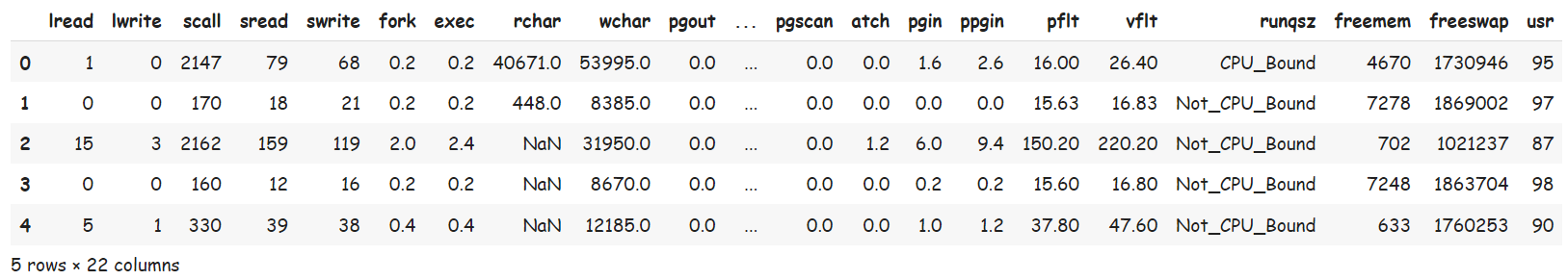
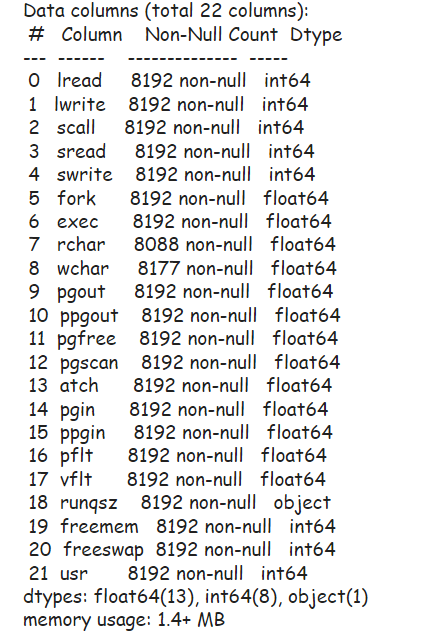


Table 1. Dataset Sample

Dataset has of 22 columns and 8192 rows.

## Q1: Exploratory Data Analysis

### Let us check the types of variables in the data frame.



There are total of 22 columns and 8192 rows in the dataset. Out of 22, 13 columns are of float type, 8 columns are of int type and rest 1 is of object type.

### Let us check for the some of the statistics that can describe the data better .

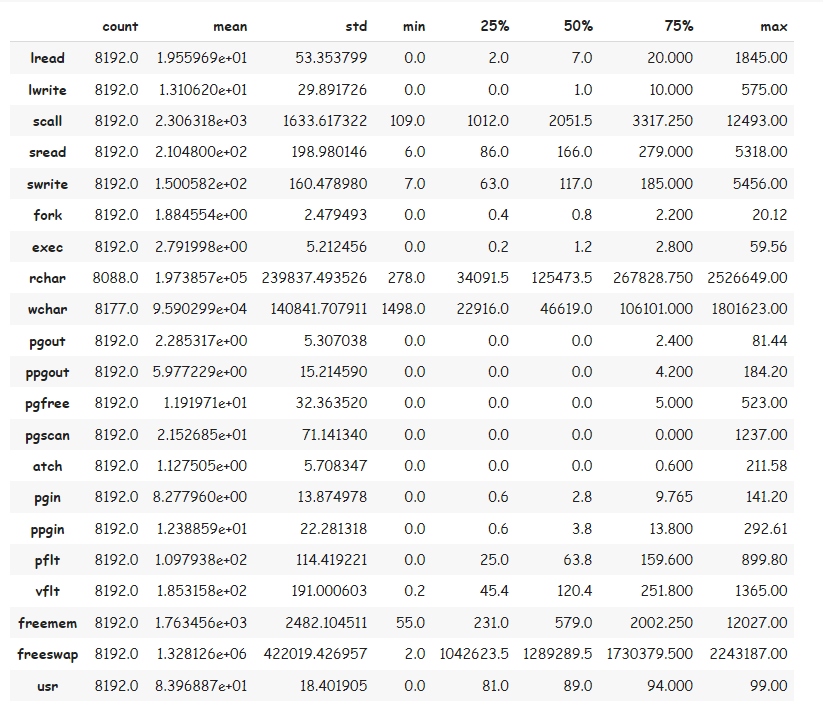


Table 2. Statistics of the given dataset

## Q2: Check for missing values in the dataset:

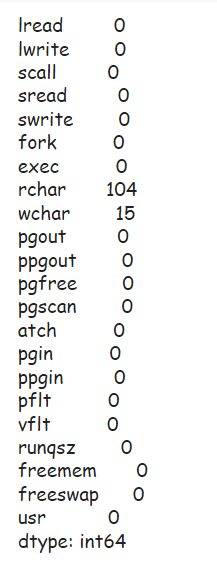


Fig 1. Dataset Sample before missing value treatment

From the above results we can see that there is missing value only in the wchar and rchar columns present in the dataset.

### Treating the missing values in the dataset:

We can consider treating the missing values with the median of the column as the median is not sensitive to the outliers if any present unlike the mean which is susceptible to outliers

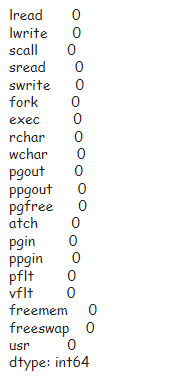


Fig 2. Dataset Sample after missing value treatment

### Importance of the rows with zeros:

When analysed the data, we can observe that many rows had zero values. But instead of dropping these rows it was decided to be retained for the below reasons:

1. Even if some values in the row were 0, the other values were present. This might affect the model or decrease the predictive power if the row in major attribute is dropped.

2. The values with 0 are large enough to be dropped and will decrease the accuracy and precision when linear equation without majority data is dropped.

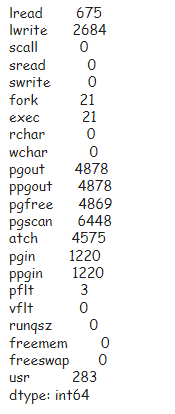


Fig 3. Count of Columns in Dataset with zeros

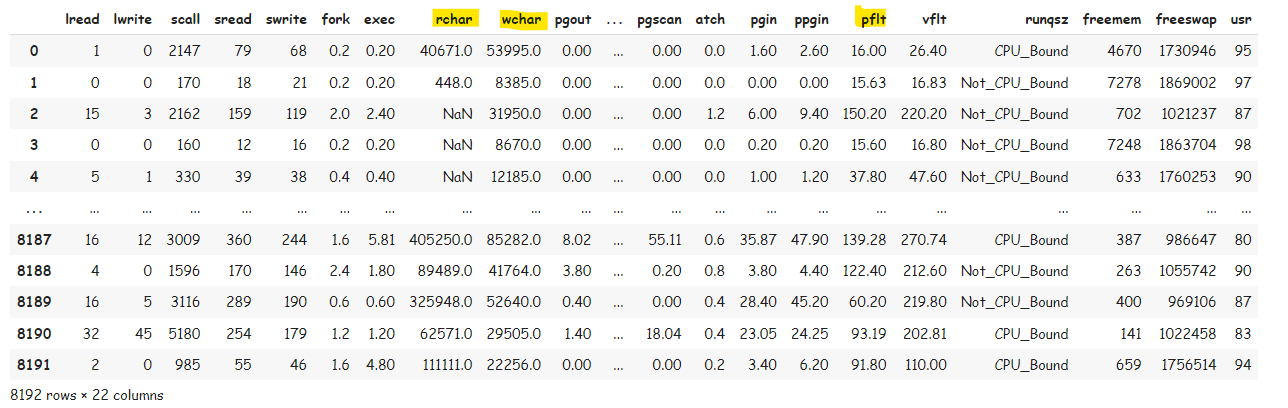


Fig 4. Columns in Dataset with zeros

## Checking for Outliers:

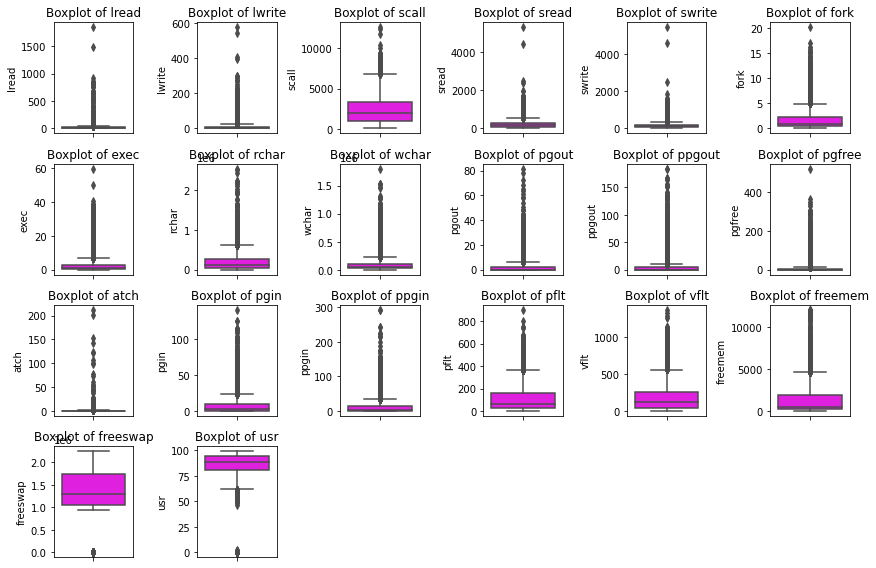


Fig 5. Detection of Outliers in the Dataset

### Treating the Outliers:

Its important to treat the Outliers before we proceed to Linear Regression as the algorithm is highly sensitive to the outliers and will significantly cause errors in the prediction, we are about to make using the linear equation.

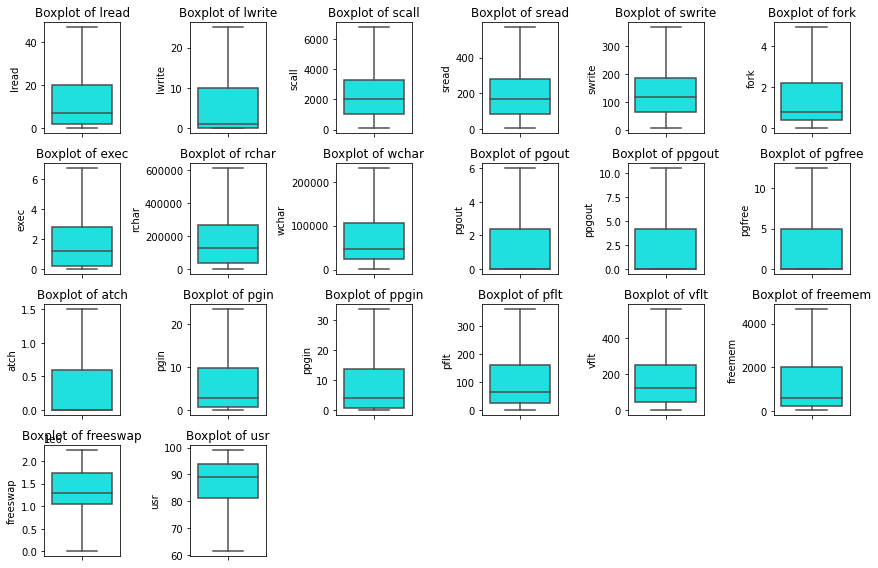


Fig 6. Outliers in the dataset are now treated

## Correlation Plot



Fig.7 – Correlation Heatmap

From the correlation plot, we can see that various attributes of the car are highly correlated to each other. Correlation values near to 1 or -1 are highly positively correlated and highly negatively correlated respectively. Correlation values near to 0 are not correlated to each other.

## PairPlot

Pairplot shows the relationship between the variables in the form of scatterplot and the distribution of the variable in the form of histogram.

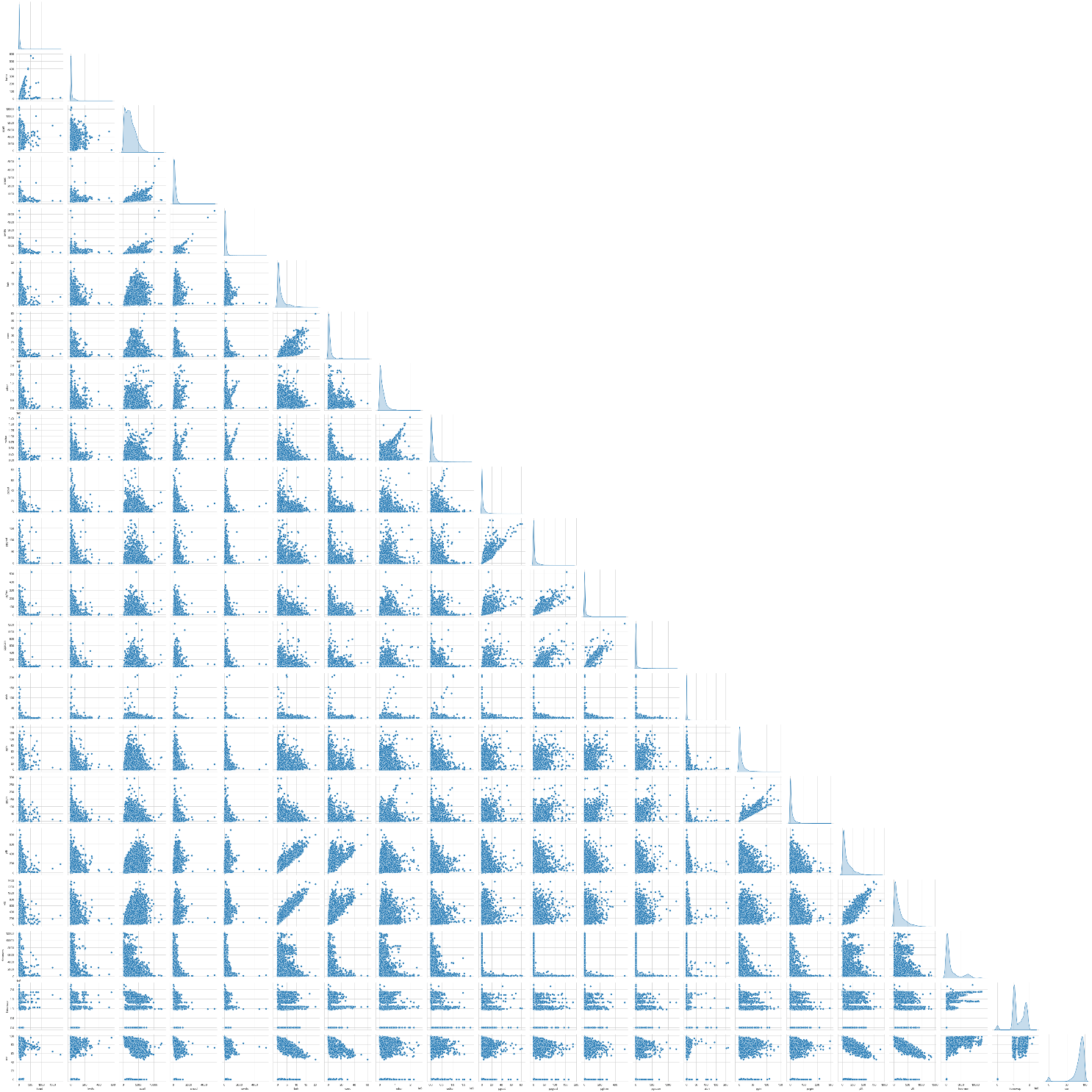


Fig.8 – Pairplot

## Q3 & Q4: Linear Regression Analysis:

### Splitting the data and building the Linear Regression Model

As now we have pre-processed the data and treated the data for outliers, we are good to start the Linear Regression.

Let’s first split the given data frame into test and train in the ratio 70:30

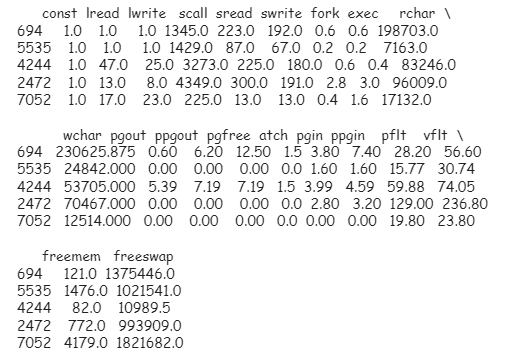


Fig.9 – Glimpse of Train Data

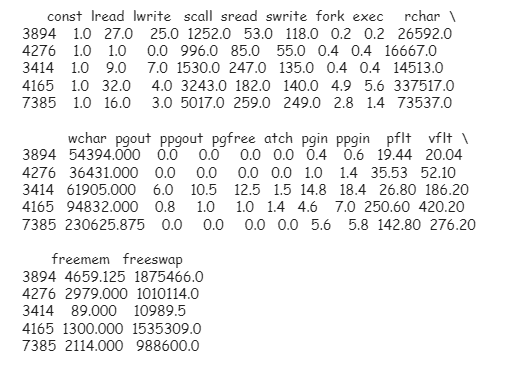


Fig. 10 – Glimpse of Test Data

### First Linear Fit Model:

Considering the above test and train data, let’s build our first Linear Regression Model.

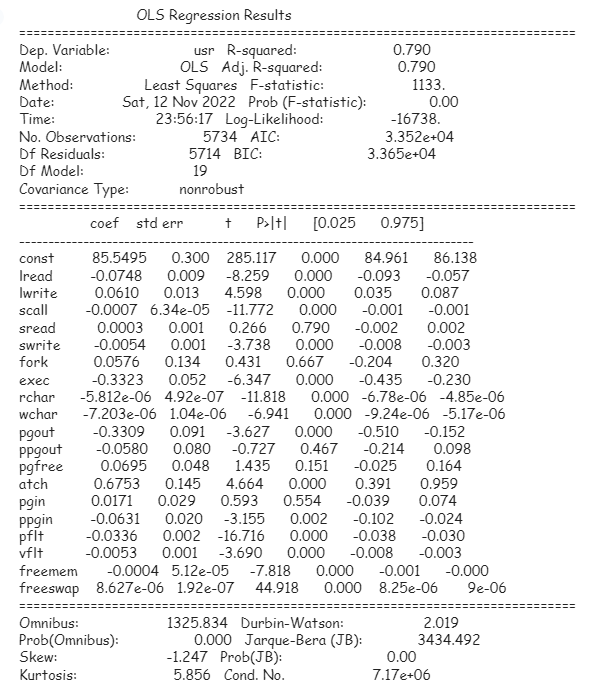


Table 3: First Linear Regression Table

#### Interpretation of R-squared

The R-squared value tells us that our model can explain 79% of the variance in the training set.

#### Interpretation of Coefficients

* The coefficients tell us how one unit change in X can affect Y.
* The sign of the coefficient indicates if the relationship is positive or negative.
* Multicollinearity occurs when predictor variables in a regression model are correlated. This correlation is a problem because predictor variables should be independent. If the collinearity between variables is high, we might not be able to trust the p-values to identify independent variables that are statistically significant.
* When we have multicollinearity in the linear model, the coefficients that the model suggests are unreliable.

#### Interpretation of p-values (P > |t|)

For each predictor variable there is a null hypothesis and alternate hypothesis.

* Null hypothesis : Predictor variable is not significant
* Alternate hypothesis : Predictor variable is significant
* (P > |t|) gives the p-value for each predictor variable to check the null hypothesis.
* If the level of significance is set to 5% (0.05), the p-values greater than 0.05 would indicate that the corresponding predictor variables are not significant.
* However, due to the presence of multicollinearity in our data, the p-values will also change.
* We need to ensure that there is no multicollinearity in order to interpret the p-values.

#### Multicollinearity

* If VIF is 1, then there is no correlation among the kth predictor and the remaining predictor variables, and hence, the variance of beta k is not inflated at all.
* If VIF exceeds 5, we say there is moderate VIF, and if it is 10 or exceeding 10, it shows signs of high multi-collinearity.
* The purpose of the analysis should dictate which threshold to use.

#### Checking the VIF of the predictors:

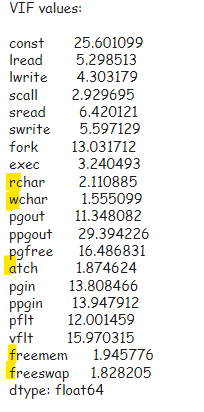


Fig. 11– Checking VIF for predictors 1

* + Multicollinearity affects only the specific independent variables that are correlated. Therefore, in this case, we can trust the p-values of freeswap, freemem, atch, wchar and rchar.
  + 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.

#### Treating Multicollinearity:

Considering the above figure of the VIF of the predictors, we are now going to drop the attributes one by one and observe if there are any difference in the R-Squared value (0.79 that is obtained in the First Model)

After dropping each value in the above cases, we could see decrease in R-Squared values as below:

lread - 0.02

pflt - 0.01

scall - 0.005

exec - 0.001

So, let's try dropping the columns other than the above mentioned columns

### Second Linear Regression Model:

After dropping all the columns in the dataset except lread, pflt, scall, exec, Let’s see the how the model performs.

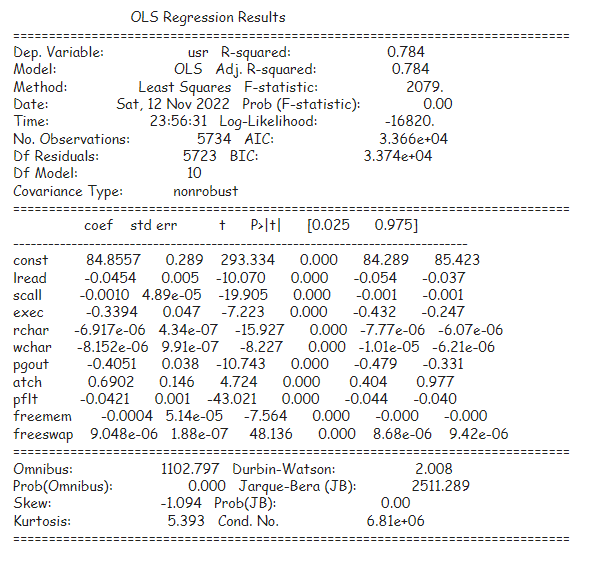


Table 4: Second Linear Regression Table

#### Checking for Multicollinearity:

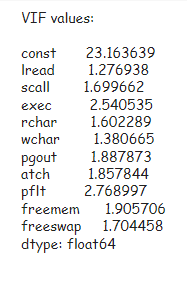


Fig. 12– Checking VIF for predictors 2

From the above observation, we can find that the multicollinearity is now significantly reduced. However, it is still present in the data making the Data unreliable to decide based on the p-values.

Out of the columns lread, pflt, scall, exec:

* Exec is the least affecting R Value predictor giving the difference of 0.001. So let’s try dropping it and building another model.

### Third Linear Regression Model:

* Let’s drop the exec (least affecting predictor out of 4 potential predictors) and try checking multicollinearity again.

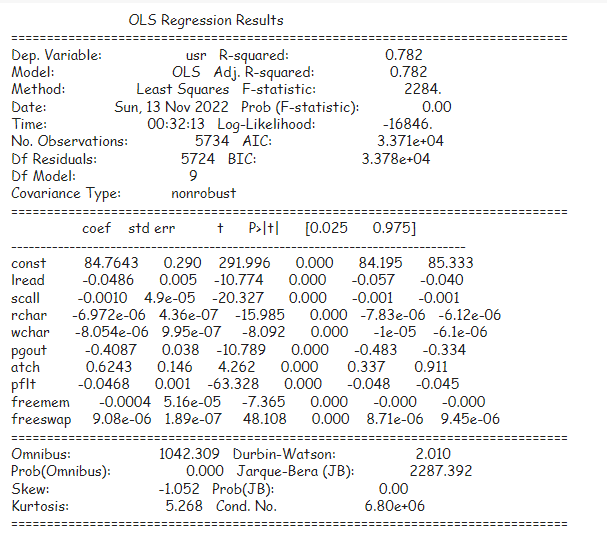


Table 5: Third Linear Regression Table

#### Checking for Multicollinearity:

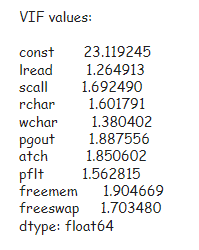


Fig. 13– Checking VIF for predictors 3

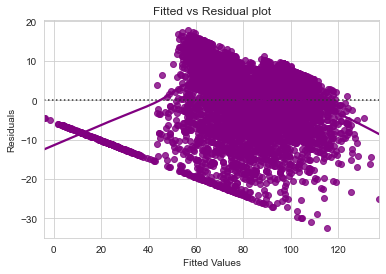
We do not have multicollinearity in our data anymore, the p-values of the coefficients have become reliable.

The Third Model is more reliable as there is no more multicollinearities.

Further, we have to check if the third model satisfies the certain condition to make it an optimally performing algorithm.

#### Let us recreate the data frame with actual, fitted, and residual values:

Now that we have recreated the Data Frame, Lets try plotting the residual plot to see the distribution of the data points



#### Normality Testing:

The above data points upon transformation are plotted to see if they satisfy the normality.

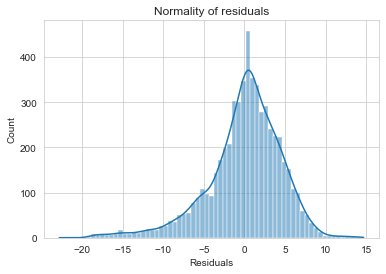


Fig. 14– Normality Testing

#### QQ-Plot

The QQ plot of residuals can be used to visually check the normality assumption. The normal probability plot of residuals should approximately follow a straight line.

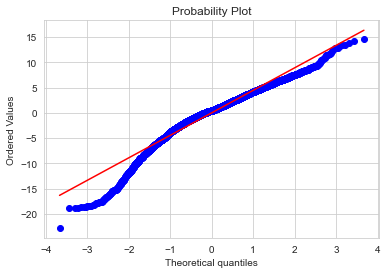


Fig. 15– QQ Plot Testing

Most of the points are lying on the straight line in QQ plot

#### Shapiro-Wilk Test:

The Shapiro-Wilk test can also be used for checking the normality. The null and alternate hypotheses of the test are as follows:

1. Null hypothesis - Data is normally distributed.
2. Alternate hypothesis - Data is not normally distributed.



Fig. 17– Shapiro p-value

Since p-value < 0.05, the residuals are not normal as per shapiro test. Strictly speaking - the residuals are not normal. However, as an approximation, we might be willing to accept this distribution as close to being normal

#### Homoscedasticity:

The null and alternate hypotheses of the goldfeldquandt test are as follows:

Null hypothesis : Residuals are homoscedastic

Alternate hypothesis : Residuals have heteroscedasticity



Fig. 17– GoldFeldquandt p-value

Since p-value > 0.05 we can say that the residuals are homoscedastic.

## Conclusion:

From the data exploration and Linear Regression, we can conclude the below:

1. The Third Linear Regression we built would be the most appropriate model, as there is no more or very little multicollinearity.
2. It has satisfied all the conditions to perform optimally.
3. There is not a big drop in the prediction when compared to the first model proving that the attributed dropped did not have a great predicting ability.

### Linear Regression Equation for the above problem:

usr = 81.35707237532267 + -0.056543116960659515 \* ( lread ) + -1.723278484920159e-07 \* ( scall ) + -7.938122103793608e-06 \* ( rchar ) + -9.421634547877954e-06 \* ( wchar ) + -0.3773284761121021 \* ( pgout ) + 0.47365049070696297 \* ( atch ) + -0.00012979650738554826 \* ( pflt ) + -0.00029022865795560256 \* ( freemem ) + 9.535265807228392e-06 \* ( freeswap )

## Problem 02:

## Executive Summary

The Republic of Indonesia Ministry of Health 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 purpose of this whole exercise is to explore the dataset. Do the exploratory data analysis. Explore the dataset using central tendency and other parameters. The data consists of 10 columns and 1473 rows.

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

System measures used:

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

### Sample of the dataset:

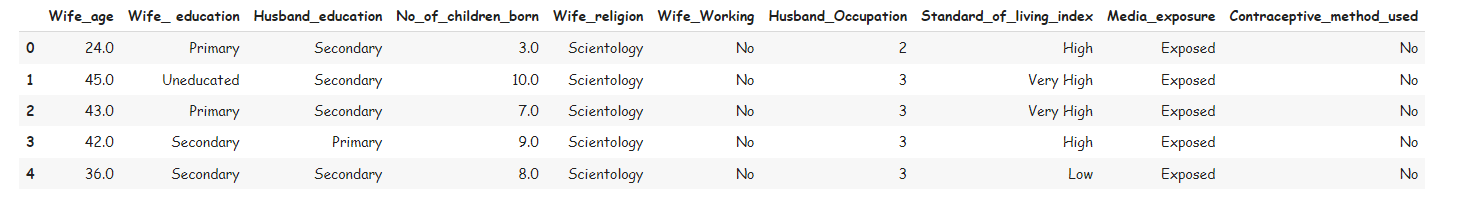
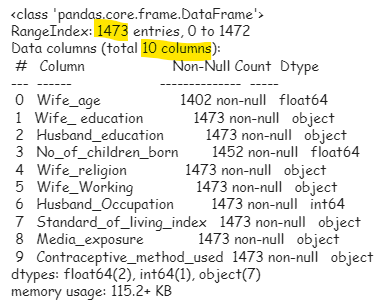


Table 2.1. Dataset Sample

Dataset has 10 columns and 1473 rows.

## Q1: Exploratory Data Analysis

### Let us check the types of variables in the data frame.



There are total of 10 columns and 1473 rows in the dataset. Out of 10, 2 columns are of float type and rest 8 is of object type.

### Let us check for the some of the statistics that can describe the data better .

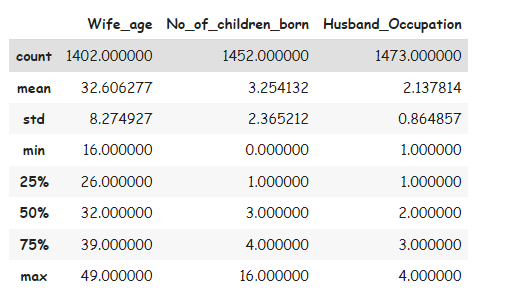


Table 2.2. Statistics of the given dataset

## Check for missing values in the dataset:

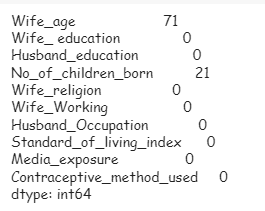


Fig 2.1. Dataset Sample before missing value treatment

From the above results we can see that there are missing value only in the Wife\_age and No\_of\_children\_born columns present in the dataset.

### Treating the missing values in the dataset:

We can consider treating the missing values with the median of the column as the median is not sensitive to the outliers if any present unlike the mean which is susceptible to outliers

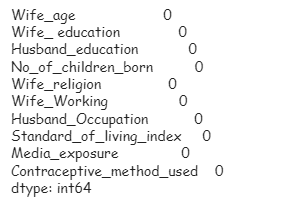


Fig 2.2. Dataset Sample after missing value treatment

## Checking for Outliers:

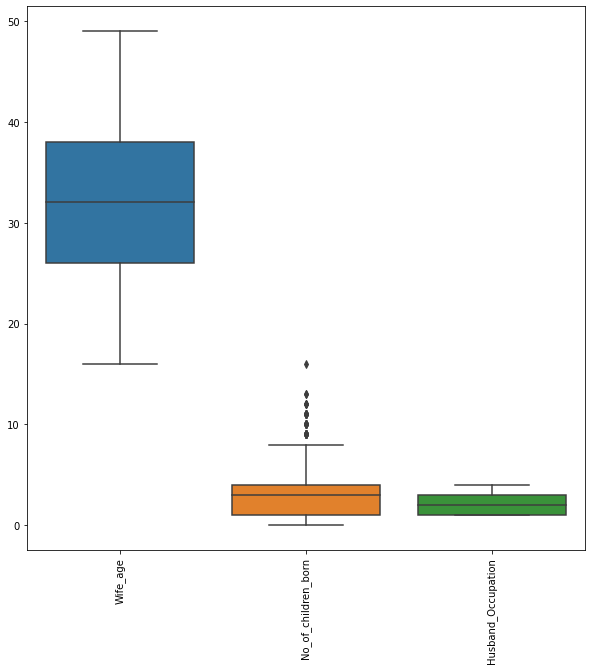


Fig 2.3. Detection of Outliers in the Dataset

### Treating the Outliers:

It’s important to treat the Outliers before we proceed to LDA or CART or Logistic regression as all these algorithms are highly sensitive to the outliers and will significantly cause errors in the prediction, we are about to make using the LDE or CART tree.

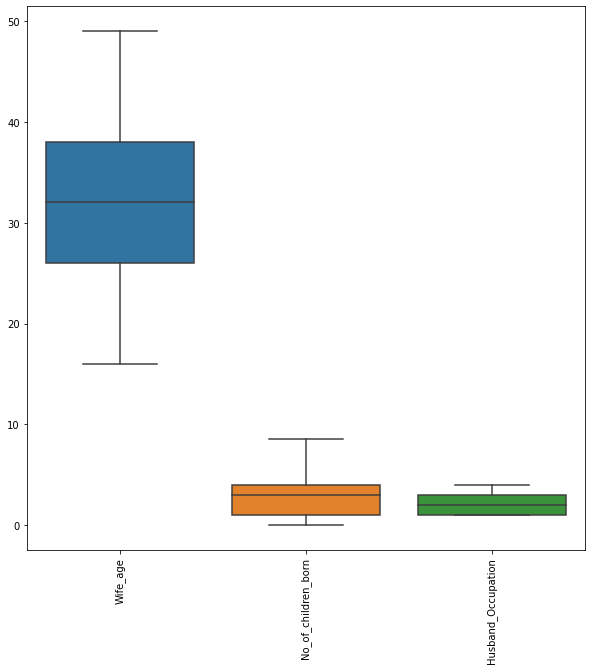
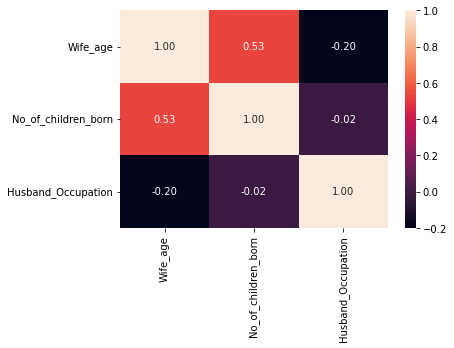


Fig 2.4. Outliers in the dataset are now treated

## Correlation Plot

 Fig.2.5 – Correlation Heatmap

From the correlation plot, we can see that various attributes of the car are highly correlated to each other. Correlation values near to 1 or -1 are highly positively correlated and highly negatively correlated respectively. Correlation values near to 0 are not correlated to each other.

## PairPlot

Pairplot shows the relationship between the variables in the form of scatterplot and the distribution of the variable in the form of histogram.

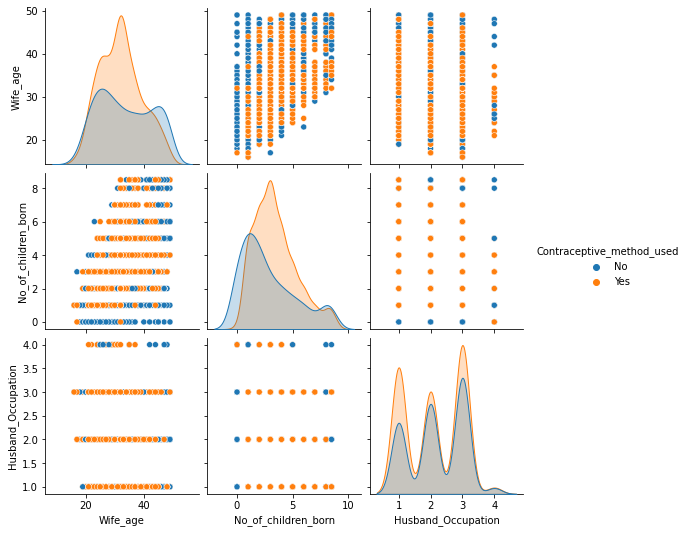


Fig.2.6 – Pairplot

## Bivariate Analysis:

Let’s try to understand the behaviour of the variables with respect to the target variable (Contraceptive\_method\_used) in the given data frame

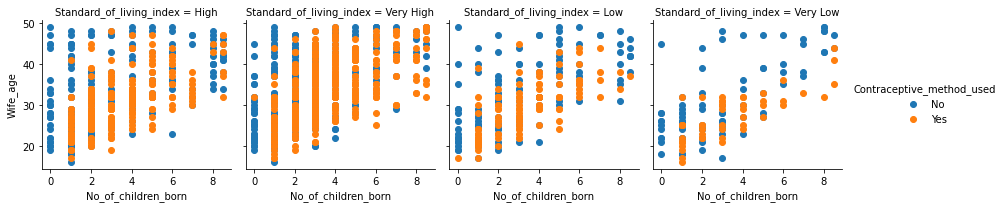


Fig.2.7 – Scatter Plot

## Q2: Logistic Regression:

### Splitting the data and building the Logistic Regression Model and Confusion Matrix

As now we have pre-processed the data and treated the data for outliers, we are good to start the Linear Regression.

Let’s first split the given data frame into test and train in the ratio 70:30

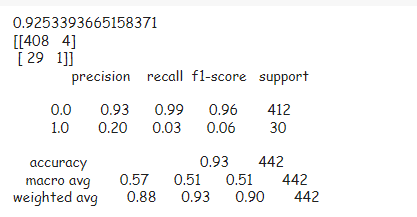


Fig.2.8 – Prediction of Logistic Regression

### Confusion Matrix:

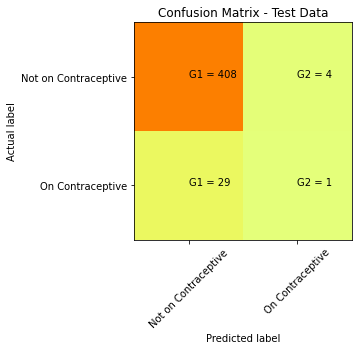


Fig.2.9 – Confusion Matrix

### Calculating some important Metrics:

Based on the above predicted and actual values, let us calculate the metrics such as Accuracy, Precision, Specificity, Sensitivity and F1 Score



#### Accuracy: 82%

#### Precision: 89%

#### Sensitivity: 90%

#### Specificity: 12%

## Q2 & Q3: Linear Discriminant Analysis (LDA)

First of all, the percentage of the target variable needs to be calculated to just give us a overall data before starting our Analysis.



Fig.2.10 – Percentage of target variable

### Building the New Data Frames X and Y for Analysis

* After Encoding and re-arranging the given data set for the ease of analysis, the Data frame is processed for the next step
* The data is now split into X and Y with respect to our target variable (ContraceptiveUsed). Such that X contains all the columns expect target variable (ContraceptiveUsed) and Y containing only (ContraceptiveUsed).

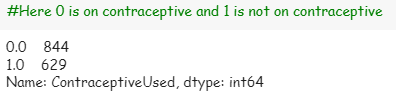


Fig.2.11 – Number of target variable after splitting of data into X and Y

### Confusion Matrix:

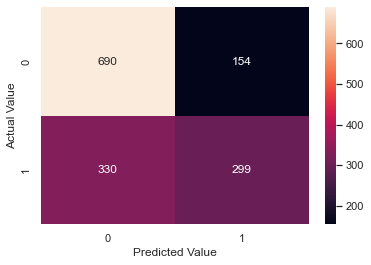


Fig.2.12 – Confusion Matrix

1020 rows classified as 0 (On Contraceptive) and 453 rows classified as 1 (Not On Contraceptive)

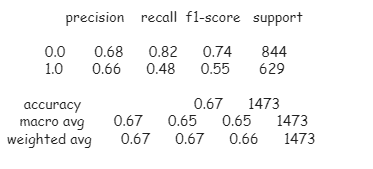


Fig.2.13 – Prediction of LDA

### LDA Analysis and Equation:

1. From the above mentioned results, the accuracy of the model is 67% which is much lower than the Logistic Regression.
2. The coefficients of the LDA are calculated and established as below mentioned

[0.08778769, -0.33817804, 0.07923715, 0.32977892, -0.09758563,-0.17989079, -0.13346181, -0.57494048, 0.98568596]



1. The equation for the LDA:

LDF=-2.031+ X1(0.087) + X2(-0.338) + X3(0.079) + X4(0.329) + X5(-0.097) + X6(-0.179) + X7(-0.133) + X8(-0.574) + X9\*0.985

1. So, from the above equation the following things can be summarized as:

* the Coeff of X9 predictor is largest in magnitude thus it helps in discriminating the target the best
* The Coeff of X5 predictor is smallest in magnitude thus it helps in discriminating the target the least.
* All the DS can be computed for each row using the above f(x) which will aid in classification

## Q2 & Q3: CART Analysis:

* Similar to LDA, after Encoding and re-arranging the given data set for the ease of analysis, the Data frame is processed for the next step
* The data is now split into X and Y with respect to our target variable (ContraceptiveUsed). Such that X contains all the columns expect target variable (ContraceptiveUsed) and Y containing only (ContraceptiveUsed).

### Establishing the importance of variables:

To understand the effect of each column or attribute on the target variable we perform a feature importance test before actually performing the CART Analysis

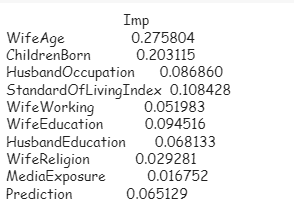


Fig.2.14 – Importance of Variables

### Building and Regularizing the Decision tree

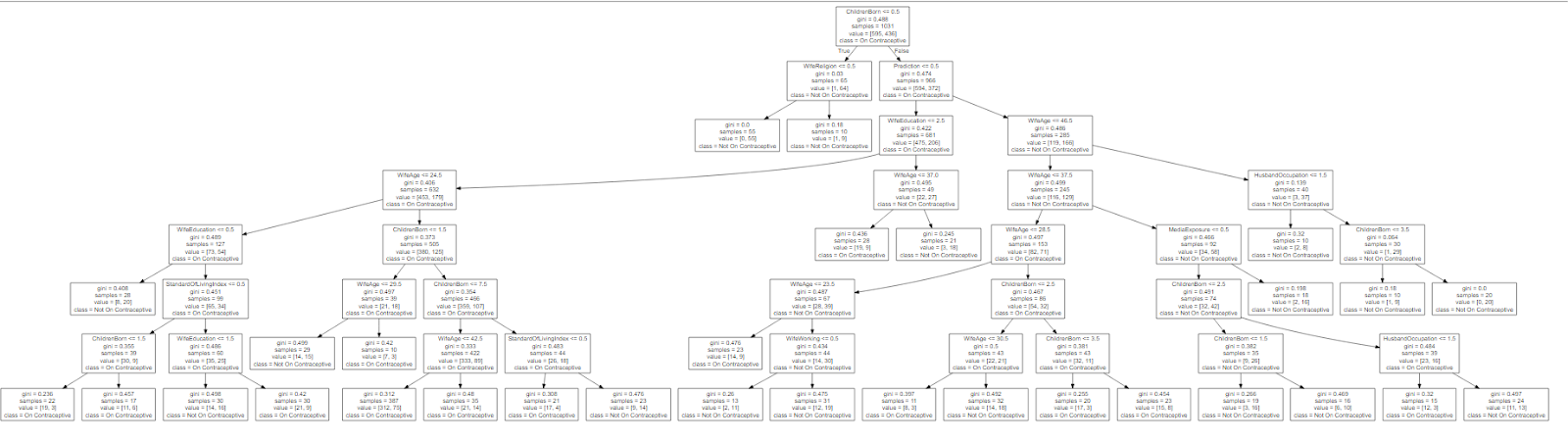
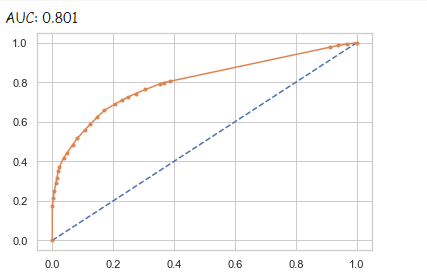


Fig.2.14 – Regularized CART Analysis Tree

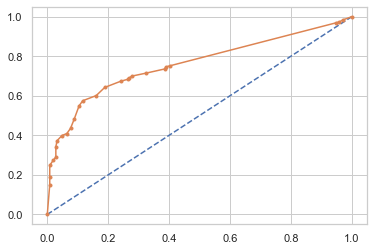
### AUC Value for training data:



AUC curve value for the training set is 80%. So let us know plot the AUC for test data and see the actual performance of the Decision tree.

### AUC Value for test data:





The regularized decision tree is 75% accurately or optimally working on the test data set.

This is the main reason for regularizing the decision tree before running on the test data to avoid the over-fitting of the tree and thereby decreasing its efficiency.

## Conclusion:

After the thorough analysis of the dataset, we can conclude the following:

* The use of contraceptive in the population is majorly affected or majorly used by women of older age and the ones with many numbers of children.
* Secondly comes the standard of living. This is another major aspect to affect the Contraceptive used.
* In a very low and low Standard of living the use of contraceptive is likely to be avoided when compared to high and very high classes.
* Next up comes the education level of the couples, Husband’s Occupation and then working wife.

## Suggestions:

* The government can come up with the health camps and programmes to provide the affordable contraceptives to the lower and very lower classes individuals.
* We may consider on educating on the alternatives of contraceptives those who are not very comfortable with the use of contraceptives
* Further, education on the contraceptives among uneducated and poorly educated is much needed to improve their knowledge and usage of the contraceptive.
* Individuals with more than a couple of children needs to be much concentrated and encouraged to use contraceptives which in order would help in the standard of their and children’s living.