**Predictive Modelling**

Student Name: Suneel Kumar Pentapalli

Date: 10/July/2023



Contents

[***Linear Regression*** 3](#_Toc140068866)

[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. 4](#_Toc140068867)

[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. 18](#_Toc140068868)

[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. 20](#_Toc140068869)

[1.4 Inference: Basis on these predictions, what are the business insights and recommendations. 23](#_Toc140068870)

[***Logistic Regression, LDA and CART*** 24](#_Toc140068871)

[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. 24](#_Toc140068872)

[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. 29](#_Toc140068873)

[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. 36](#_Toc140068874)

[2.4 Inference: Basis on these predictions, what are the insights and recommendations. 43](#_Toc140068875)

[Please explain and summarise the various steps performed in this project. There should be proper business interpretation and actionable insights present. 45](#_Toc140068876)

# Linear Regression

**Problem 1**: Linear Regression

The comp-activ databases is a collection of a computer systems activity measures .  
The data was collected from a Sun Sparc station 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/90006/files/8391643/download?verifier=8SKH1mjNZTwZAhrIIXTUsmVBS4qOwddXjORSwsxf&wrap=1)

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

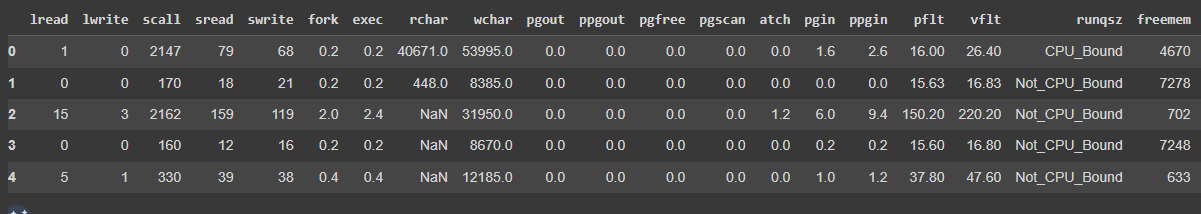
* lread - Reads (transfers per second ) between system memory and user memory
* lwrite - writes (transfers per second) between system memory and user memory
* scall - Number of system calls of all types per second
* sread - Number of system read calls per second .
* swrite - Number of system write calls 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.

The given dataset "compactiv.xlsx" has

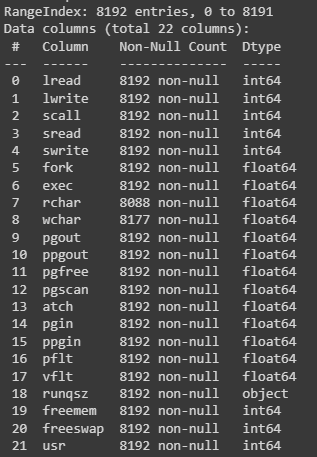
* 8192 records and 22 columns.
* 1.2% and 0.18% Nulls in rchar and wchar respectively.
* No duplicate records.
* Of all 22 columns, only runqsz is of object type and rest all are of numeric type.

Below is the sample data,

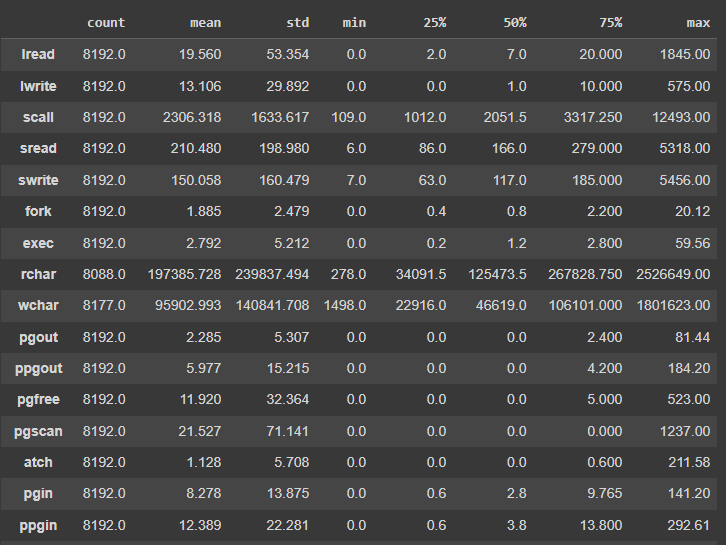
A screenshot of a number

Description automatically generated

**Data types and column names:**



**5 Point summary of compactiv.xlsx:**



A screenshot of a graph

Description automatically generated

As per above summary,

1. values in each columns seems to be of different scales.
2. 13 attributes have values starting with 0.
3. there is considerable difference between mean and median for almost all the attributes which indicates skewness in data.

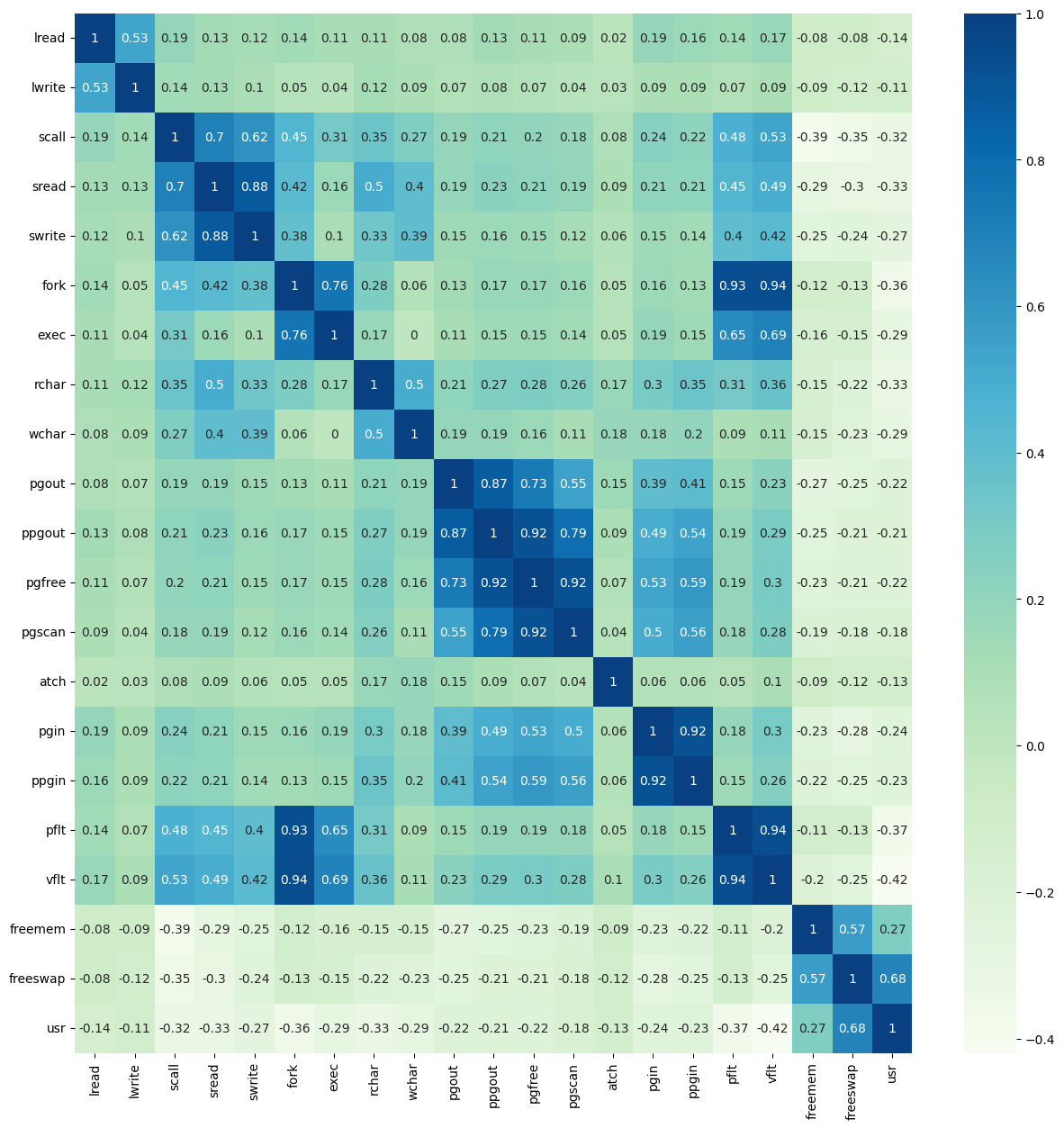
**Correlation Plot:**

From the correlation plot, we can see that,

There is high amount of positive correlation between fork-pfit, fork-vfit, vfit-pfit, pgin-ppgin, pgscan-pgfree, ppgout,pgfree, sread-swrite etc.,

There is high amount of negative correlation between usr-vfit,usr-pfit etc.,

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.



**Univariate Analysis:**

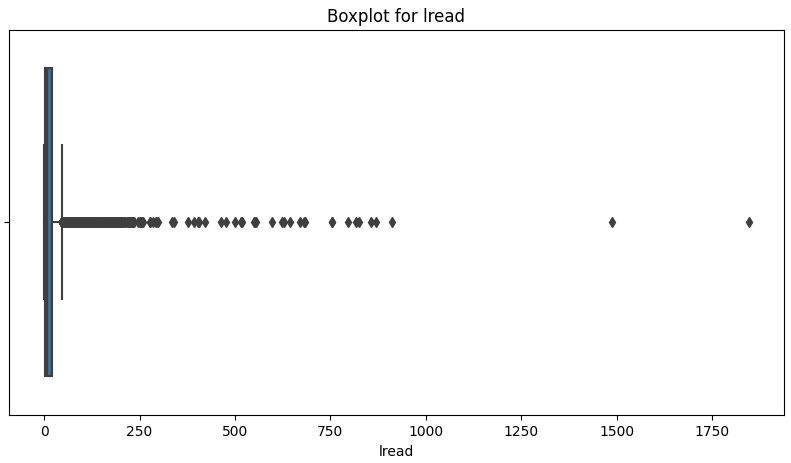
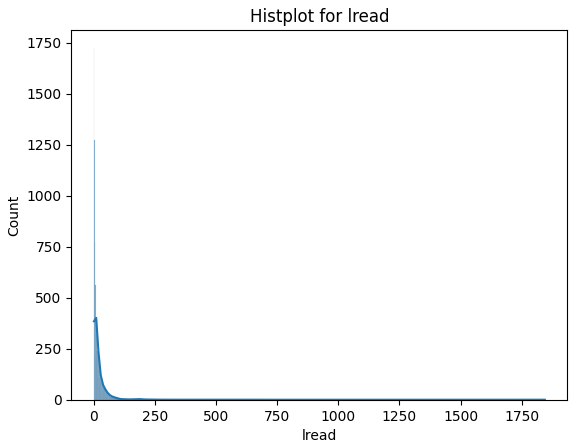
**Univariate analysis** is the simplest form of analyzing data. “Uni” means “one”, so in other words your data has only one variable. It doesn’t deal with causes or relationships and it’s major purpose is to describe; It takes data, summarizes that data and finds patterns in the data.

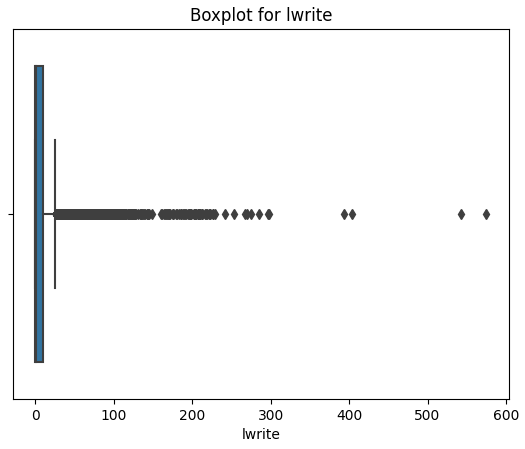
Below is the visual representation of univariate analysis of all the numerical attributes,

Graph on the left side is a box plot, the ponts in the box plot represents outliers.

Similarly graph to the right shows a histplot for visualizing distribution of data in the corresponding column. Higher Skewness in this graph represents outliers.

As per both the graphs, it is evident that most of the attributes are highly skewed. Also distribution of data indicates that mode lies near lower numbers.

A graph with a line

Description automatically generated

A graph of a box plot

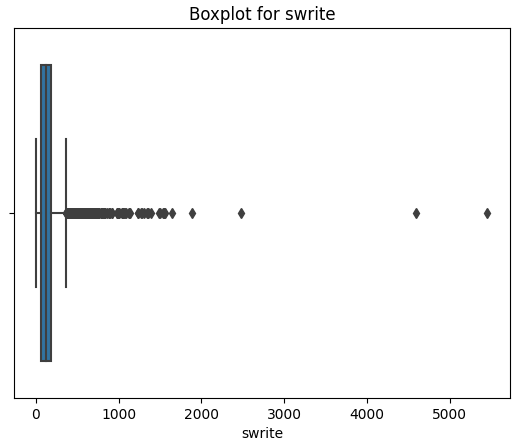
Description automatically generatedA graph of a scallop

Description automatically generated

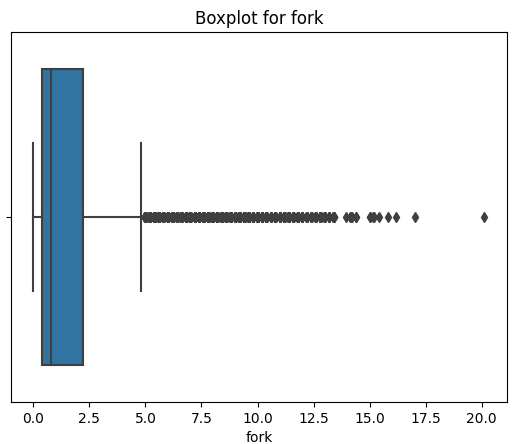
A graph with a box plot

Description automatically generatedA graph of a number of blue lines

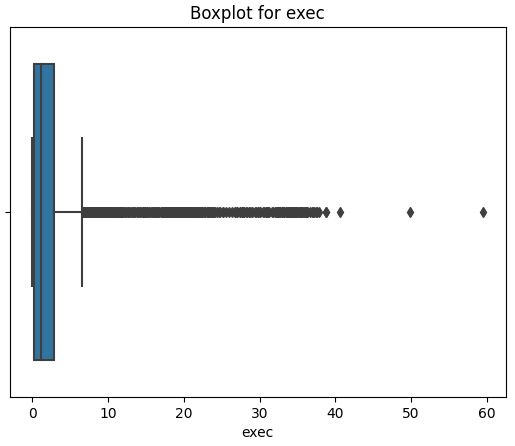
Description automatically generated

A graph with a blue line

Description automatically generated

 A graph of a number of objects

Description automatically generated

 A graph with blue lines

Description automatically generated

A graph of a box plot

Description automatically generatedA graph of a number of blue lines

Description automatically generated

A graph of a box plot

Description automatically generatedA graph of a graph

Description automatically generated

A graph with a blue bar and black dots

Description automatically generatedA graph with numbers and lines

Description automatically generated

A graph with a box plot

Description automatically generatedA graph with numbers and lines

Description automatically generated

A graph of a box plot

Description automatically generatedA graph with numbers and lines

Description automatically generated

A graph with a line graph

Description automatically generatedA graph with a line

Description automatically generated

A graph with a number of dots

Description automatically generatedA graph with numbers and lines

Description automatically generated

A graph with a box plot

Description automatically generatedA graph of a number of points

Description automatically generated

A graph of a box plot

Description automatically generatedA graph with a line

Description automatically generated

A graph with a box plot

Description automatically generatedA graph of a graph

Description automatically generated

A graph with a blue rectangular bar

Description automatically generatedA graph of a number of blue lines

Description automatically generated

A graph of a box plot

Description automatically generatedA graph of a number of freemen

Description automatically generated

A diagram of a box plot

Description automatically generatedA graph of a graph

Description automatically generated

A graph with a blue rectangular object

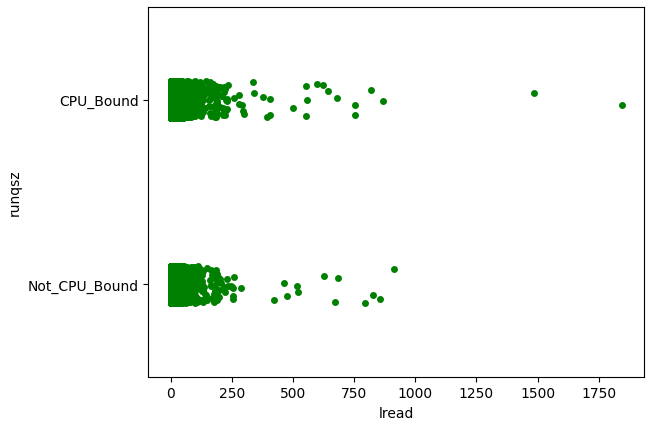
Description automatically generatedA graph of a number of blue lines

Description automatically generated

**Bivariate Analysis:**

In bivariate analysis we study distribution and patterns between two variables as shown below,

Here we have used stripplot to look at the datapoints against each runqsz type.  
Stripplot of lread against runqsz shows more lread(Reads (transfers per second ) between system memory and user memory) in CPU\_Bound systems than in Not\_CPU\_Bound systems.



Stripplot between usr(Portion of time (%) that cpus run in user mode) and runqsz shows that portion of systems with CPU\_Bound are staying idle.

A graph with blue dots

Description automatically generated

**Multivariate Analysis:**

Multivariate analysis is a set of techniques to **study data with more than one variable** and their relationships

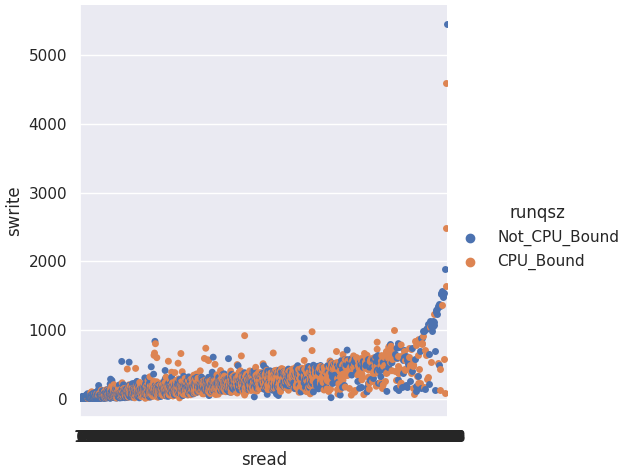
Below graph shows the correlation of fork and usr against runqsz.

fork and user are varying inversely due to their negative correlation.

A graph of different colored dots

Description automatically generated

attributes sread and lread are almost evenly distributed and also varying in same direction due to high correlation.



### 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.

**Handling Null Values:**

Fetching null percentages in null columns.



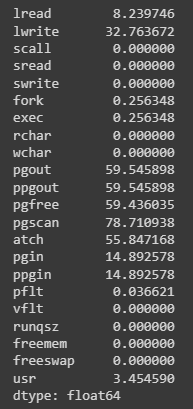
Null percentage in above columns is very less which corresponds to around 1.2% and 0.18% in rchar and wchar respectively.

Replaced all the null values with the corresponding attribute mean values.



**Checking for zeroes:**

Below are the percentages of zeroes in each column,



six attributes have zeroes more than 50%, but there were no rules available on how each attribute is related or calculated. Hence we are not changing/imputing those zeroes.

As already shown in boxplots in 1.1,

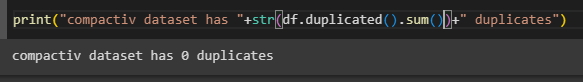
Outliers are present in all the attributes(lread, lwrite, scall, sread, swrite, fork, exec, rchar, wchar, pgout, ppgout, pgfree, pgscan, atch, pgin, ppgin, pfit, vfit, freemem, freeswap, usr)

Only freeswap columns has less outliers

But as these outliers seems to be actual values, we will proceed with further steps without treating them.

**Duplicates:**

As shown below the given dataset has no duplicates.



### 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 String values in runqsz:**

As seen above, the only object type column in the dataset is runqsz and it has values 'CPU\_Bound', 'Not\_CPU\_Bound' only.



Before building our linear regression model its better to change the object datatype to int by decoding the values to 0 and 1 as shown below.

Note: Here we have hardcoded value “Not\_CPU\_Bound” of type object to the value 1 of type int, similary modified “CPU\_Bound” to 0 of int type.

A screenshot of a phone

Description automatically generated

Now we have out dataset full of numerical and categorical values.

**Splitting the dataset in to test and train data:**

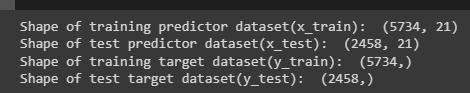
Before splitting the data and test and train sets, lets divide our features into predictor/features and target splits as shown below.

A screen shot of a computer code

Description automatically generated

Now that we have created two datasets with all predictors in one dataframe(x) and target variable(usr) in one data frame(y)

Now we will split x and y into x\_train,x\_test and y\_train and y\_test dataframes with test sample size of 30% and train sample size of 70%



Now we have created a model and fitted the training values in the model.

Here are some of the metrics of model on training and test datasets,

**R2 Value: The most common interpretation of r-squared is how well the regression model explains observed data. For example, an r-squared of 60% reveals that 60% of the variability observed in the target variable is explained by the regression model. Higher the r-squared value, better the model is. This means, our model is almost 64% accurate.**

The coefficient of determination R^2 of the prediction on Train set 0.6363661657685671

The coefficient of determination R^2 of the prediction on Test set 0.6468867961739946

**RMSE: Root mean square error shows us the deviation of predicted values from actual values. Lesser the deviation, better the model.**

The Root Mean Square Error (RMSE) of the model for training set is: 10.991906841867554

The Root Mean Square Error (RMSE) of the model for testing set is: 11.168605637051579

**Adjusted R2 Value:** **Adjusted R-squared is a modified version of R-squared that adjusts for the number of predictors in a regression model. R-squared measures the proportion of the variation in the dependent variable explained by the independent variables for a linear regression model. Adjusted R-squared compares the explanatory power of regression models that contain different numbers of predictors. It increases only if the new term improves the model more than would be expected by chance.**

Adjusted R square of the model for training set is: 0.6350292766721279

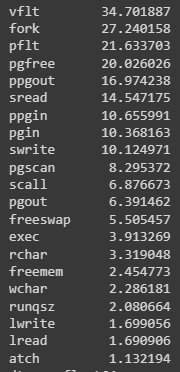
Adjusted R square of the model for testing set is: 0.643842716830667

As shown above, our model is only 64% accurate, lets explore more and find out if it can be improved,

Variance Inflation Factor(VIF):

This metric shows us the multicollinearity in predictors.

Generally, a VIF above 3 indicates that multicollinearity might exist, and further investigation is required. When VIF is higher than 10, there is significant multicollinearity that needs to be corrected.



Upon verifying by recreating the model removing each column at a time from the above list didn’t really increase our model accuracy.

Hence lets proceed further and check for the feature importance by checking the pvalue of each attribute and remove the variables that are not adding any value to the model.

Used ols method to create model and fetch the below data,

A screenshot of a computer screen

Description automatically generated

Values in the columns “P>|t|” represents p value, if its less than assumed significance level(0.05) then that means the attribute is important and we shouldn’t delete it. If its greater than 0.05 then we can try deleting the column and calculate if its impacting our R2 value or RMSE.

Upon repeating above steps, we have identified that the columns 'lwrite','sread','swrite','exec','pgscan','pgin','ppgin' are adding no values to the model and can be removed.

Final model score still remains the same. This means the given accuracy is due to the nature of data.

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

Final model score still remains the same. This means the given accuracy is due to the nature of data.

Linear equation to calculate “usr” variable is as follows,

usr = (-0.02)\*lread+(0.001)\*scall+(-2.072)\*fork+(-0.0)\*rchar+(-0.0)\*wchar+(-0.24)\*pgout+(0.106)\*ppgout+(-0.05)\*pgfree+(-0.051)\*atch+(-0.039)\*pflt+(0.023)\*vflt+(7.749)\*runqsz+(-0.002)\*freemem+(0.0)\*freeswap+43.589

Note that out of 22 columns given, only above columns are identified as important and they alone were able to predict the value of usr variable with 63% accuracy.

Though the variables rchar, wchar and freeswap have very less coefficient values, we cannot deleted those as dropping those columns is reducing the our model accuracy and increasing RSME value. Hence we choose not to drop them.

Basically linear regression model assumes to perform better when the data has below characteristics,

**1. Linear relationship:** There exists a linear relationship between the independent variable, x, and the dependent variable, y.

**2. Independence:**The residuals are independent. In particular, there is no correlation between consecutive residuals in time series data.

**3. Homoscedasticity:**The residuals have constant variance at every level of x.

**4. Normality:**The residuals of the model are normally distributed.

But unfortunately our dataset has violated all the above, Hence this might have caused our low model score.

# 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**](https://olympus.mygreatlearning.com/courses/90006/files/8391642/download?verifier=yeDVnXhdejRL06kVU3l2itYOIFSDTrVMUEbk9epZ&wrap=1)

**Data Dictionary:**

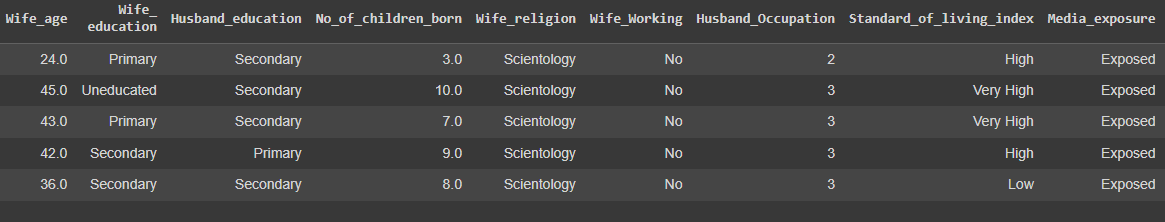
1. Wife's age (numerical)  
2. Wife's education (categorical) 1=uneducated, 2, 3, 4=tertiary  
3. Husband's education (categorical) 1=uneducated, 2, 3, 4=tertiary  
4. Number of children ever born (numerical)  
5. Wife's religion (binary) Non-Scientology, Scientology  
6. Wife's now working? (binary) 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.

The given dataset "compactiv.xlsx" has

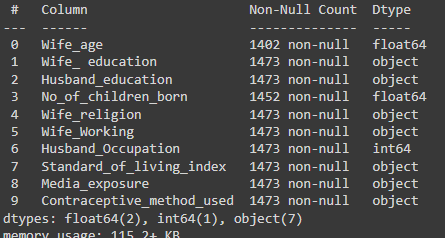
* 1473 records and 10 columns.
* 71 nulls in column Wife\_age and 21 nulls in No\_of\_children\_born.
* 80 duplicate records.
* Of all 10 columns, Wife\_age, N.

Below is the sample datao\_of\_children\_born and Husband\_Occupation are of numeric type and remaining 7 columns are of object type,

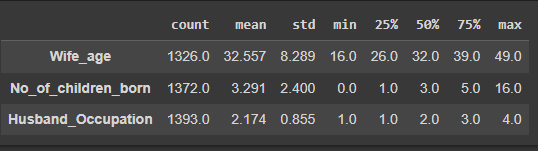
A screenshot of a computer

Description automatically generated

**Datatypes and Column names:**



**Five point summary after removing duplicates:**



As per above summary,

Minimum age of the wife from the given data set is 16 and maximum is 49. Max age seems legible but min age 16 is concerning.

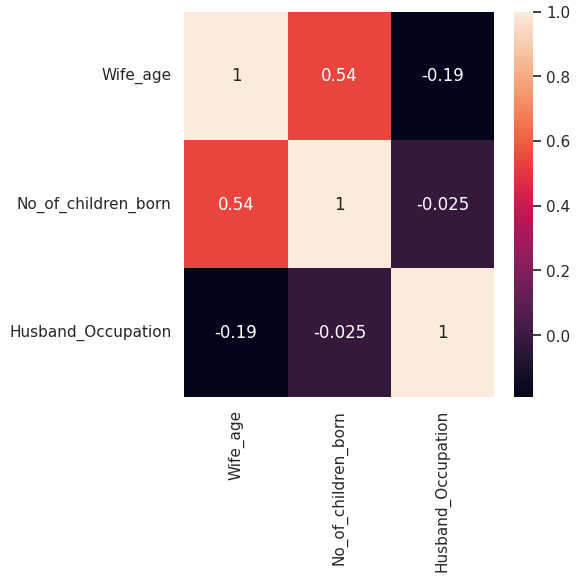
Max Number of children is 16, which might be incorrect given the age of wife. This could be a outlier data.

**Correlation Plot:**

From the correlation plot, we can see that,

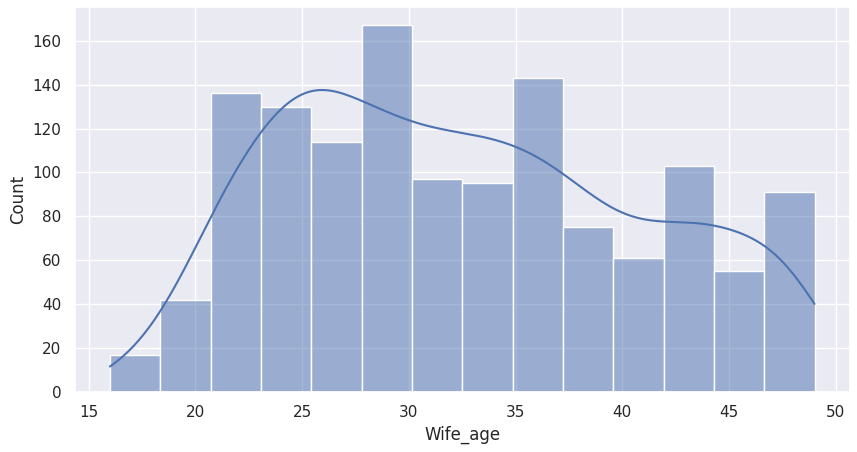
There is a positive correlation between wife age and number of children which is obvious. Remaining columns don’t have much correlation.

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.



**Univariate Analysis:**

Checking Distribution of Wife\_age, No\_of\_children\_born and Husband\_Occupation



Values in Wife\_age column are almost uniformly distributed, with mode value between 25 to 30.

A graph with a line going up

Description automatically generated

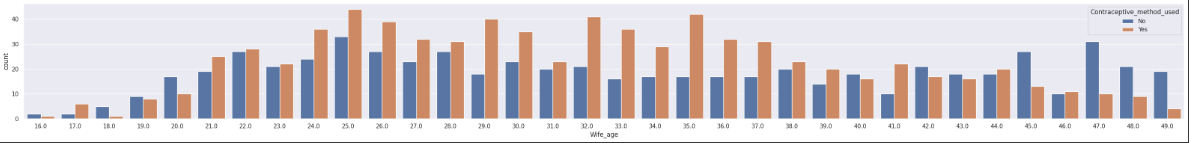
When we look at distribution of No\_of\_children\_born, it is evident that the abnormal child count near 16 is a outlier and its causing skewness in data. Most of the people are giving births to minimum 2 kids.

A graph with blue lines

Description automatically generated

From the distribution of Husband\_Occupation, it shows that the given dataset has more records with occupation type 3 and less records with occupation type 4.

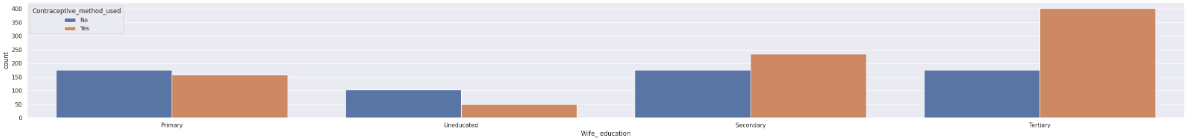
**Bivariate Analysis:**



Above graph shows counts of females who did take contraceptives and who didn’t among different age groups. Age groups 25 to 35 seems to have taken more people who took contraceptives.

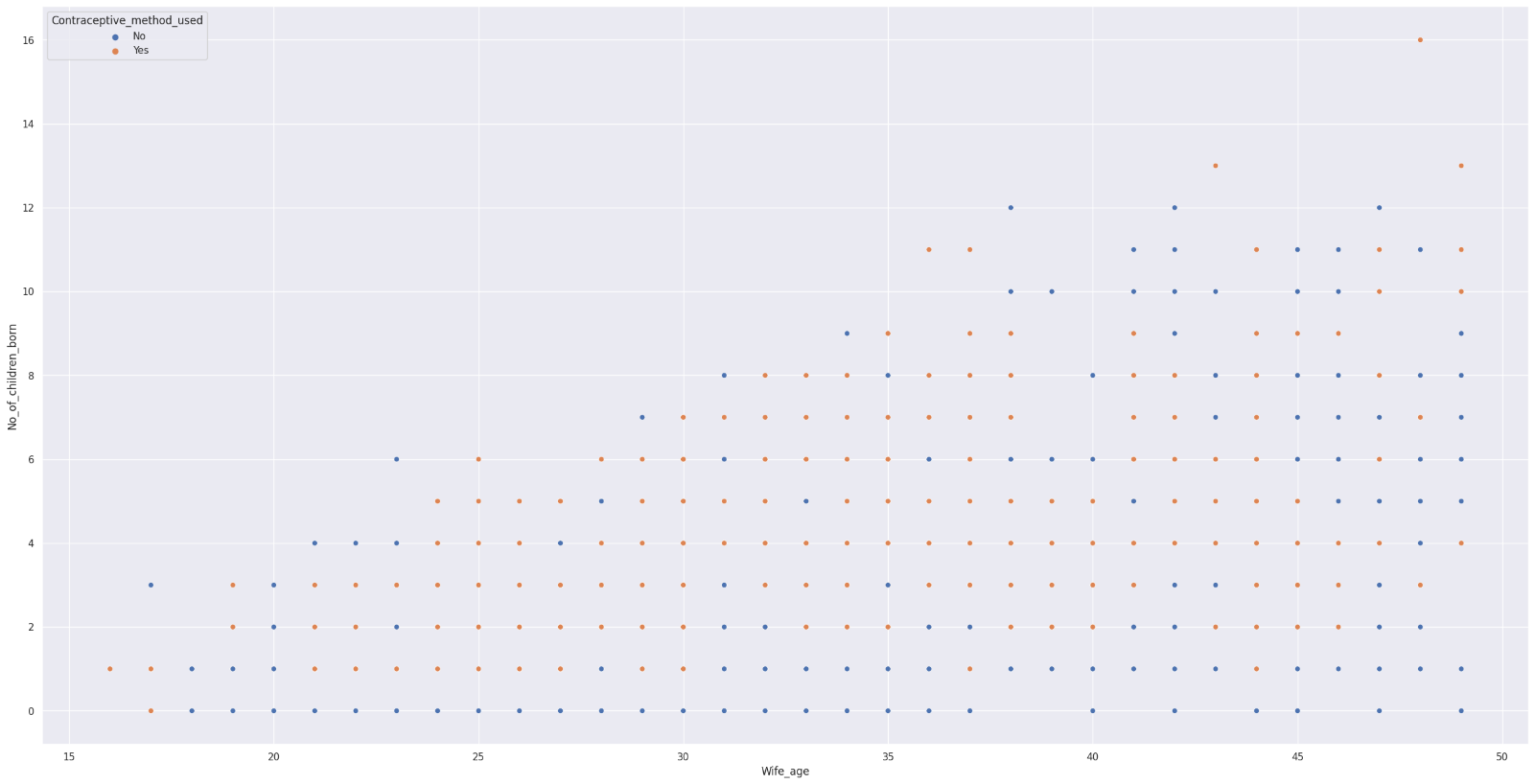


Above chart shows the distribution of females who took/didn’t took contraceptives among working and not working wives. From the given data, female who are not working seems to be taking more contraceptives.



Similarly above graph shows distribution of contraceptives among different wife\_education categories. Females with education type Tertiary have taken more contraceptives.

**Multivariate Analysis:**



Above is the visual representation of change of contraceptives used in accordance with age and number of children.

There is an increase in females with contraceptives use with increase in number of children.

### 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.

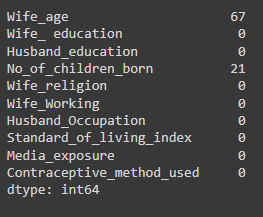
Before splitting our data into predictor and predicted set,

Lets treat the data by removing nulls that we observed in Wife\_age and No\_of\_children\_born.

Treat outliers present in No\_of\_children\_born.

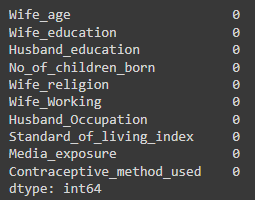
**Handling Nulls:**

Below are the counts on nulls in each column,

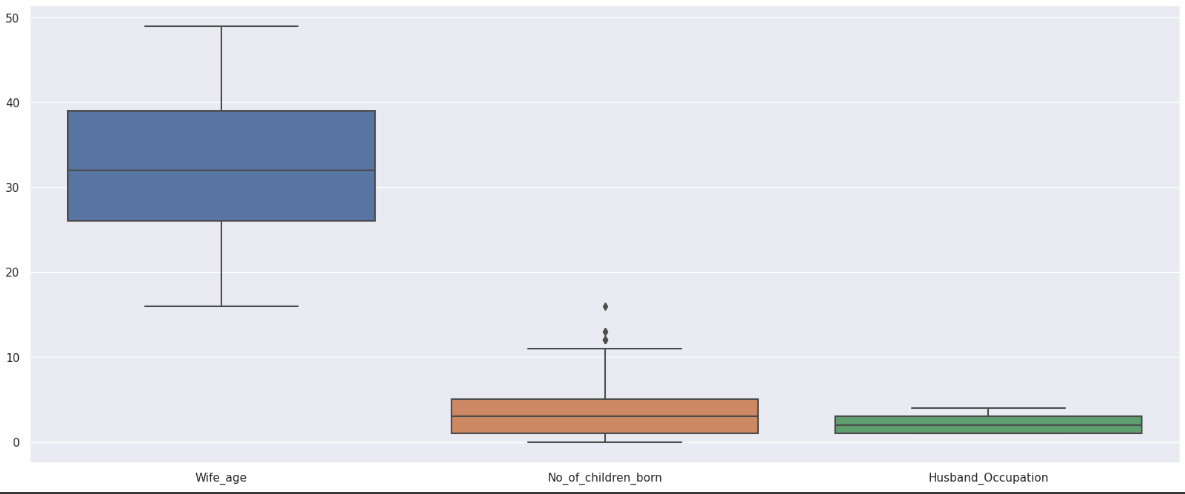


Imputed nulls in wife\_age column with mean age of the column and the nulls in No\_of\_children\_born are replaced with zeroes.

Here is the count of nulls after treatment.



**Handling Outliers:**

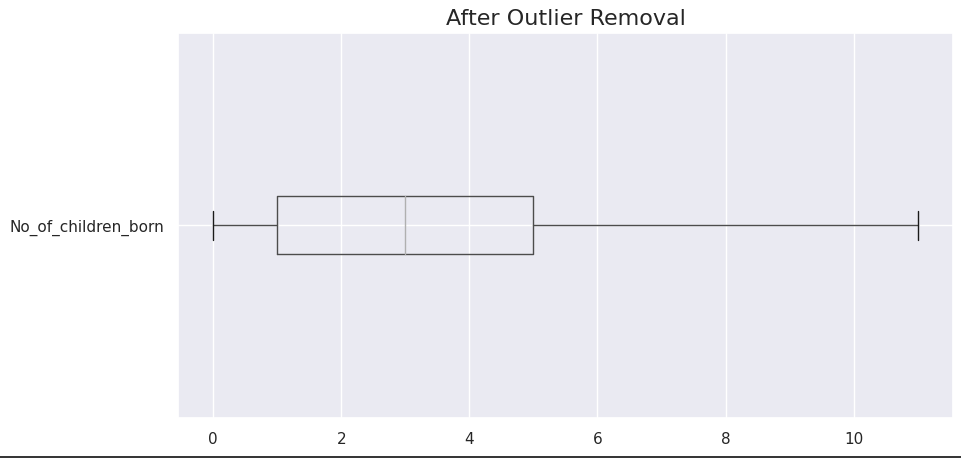


As we have already observed some radical values in No\_of\_children\_born column like number of children as more than 8. We will treat them using Inter Quartile Range method.

This method says any value less than (Q1-1.5IQR) or value greater than (Q3+1.5IQR) is treated as outlier and we can replace them with 25th and 75th percentiles of the total columns values respectively.

Where IQR = 75th percentile of (No\_of\_children\_born) – 25th percentile of (No\_of\_children\_born)

After removing outliers, below is the new boxplot, where we can observe that there are no dots/datapoints outside our right whisker.



Now that we have removed irregularities in our data,

We also need decode the categorical values in the columns as shown below, so that we can pass the numerical values to our model.

{

'Wife\_education': {'Primary': 0, 'Uneducated': 1, 'Secondary': 2, 'Tertiary': 3},

'Husband\_education': {'Secondary': 0, 'Primary': 1, 'Tertiary': 2, 'Uneducated': 3},

'Wife\_religion': {'Scientology': 0, 'Non-Scientology': 1},

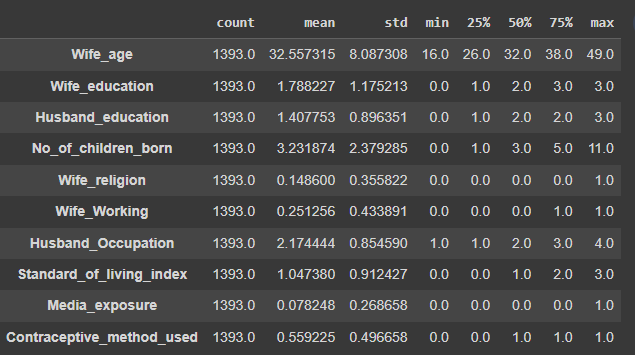
'Wife\_Working': {'No': 0, 'Yes': 1},

'Standard\_of\_living\_index': {'High': 0, 'Very High': 1, 'Low': 2, 'Very Low': 3},

'Media\_exposure': {'Exposed': 0, 'Not-Exposed': 1},

'Contraceptive\_method\_used': {'No': 0, 'Yes': 1}}

Here is the summary of values in each attribute after decoding,



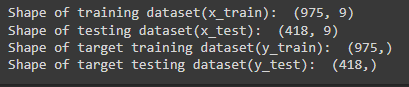
Now we split data into

predictor variables(x) with columns ['Wife\_age', 'Wife\_education', 'Husband\_education', 'No\_of\_children\_born', 'Wife\_religion', 'Wife\_Working', 'Husband\_Occupation', 'Standard\_of\_living\_index', 'Media\_exposure'] and

predicted dataset(y) with column [‘Contraceptive\_method\_used’]

Now we can split our data into training and testing datasets in 70:30 ratio respectively. Also we used stratify method while splitting to ensure the distribution of target variable remains almost similary in training and test datasets.

Here are the record counts and column counts after splitting data. Training data has 975 records and testing data has 418 records.



**Building Logistic Regression(LR) Model:**

Upon building logistic regression model and post passing our training set for fitting, these are the scores for training and testing datasets.



Below graph shows the magnitude of each coefficient for the linear equation we can use from this model.

A graph with blue rectangles

Description automatically generated

**Building Linear Discriminant Analysis(LDA) Model:**

Upon building LDA model and post passing our training set for fitting, these are the scores for training and testing datasets.



For the linear equation that LDA model has predicted, below are the values for each coefficient.

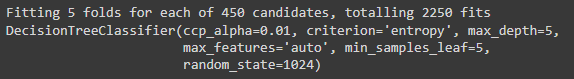
A graph with a bar chart

Description automatically generated

**Building Classification and Regression Tree(CART) Model:**

Instead of directly building of CART model, lets find out the best possible parameters to pass for into our DecisionClassifier method using GridSeachCV algorithm.

The output we got is as follows,

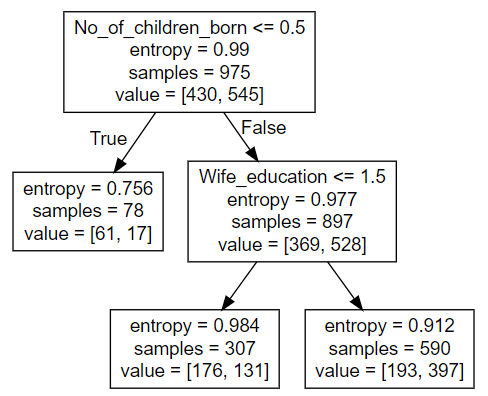


Upon building CARTmodel and post passing our training set for fitting with the above parameters, these are the scores for training and testing datasets.

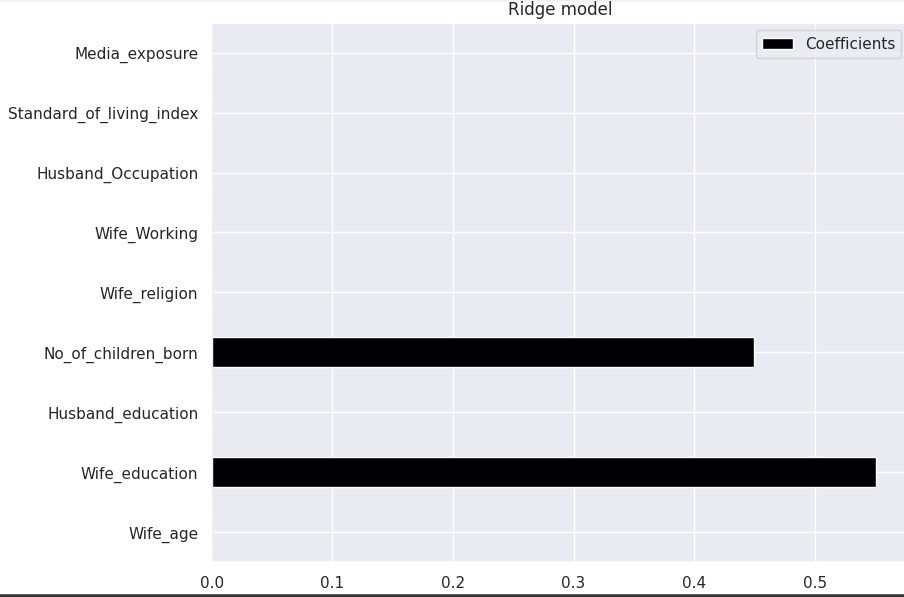
A close up of a text

Description automatically generated

Below is the decision tree used by the model to predict contraceptive use,



As this is a decision tree model, unlike regression model we do not get any linear equation. Hence we wont get coefficients as above, instead we can see the features with utmost importance for this model, they are as follows,



This model just used two columns to predict the target variable, we will see how each of this model is performing in below sections.

### 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:**

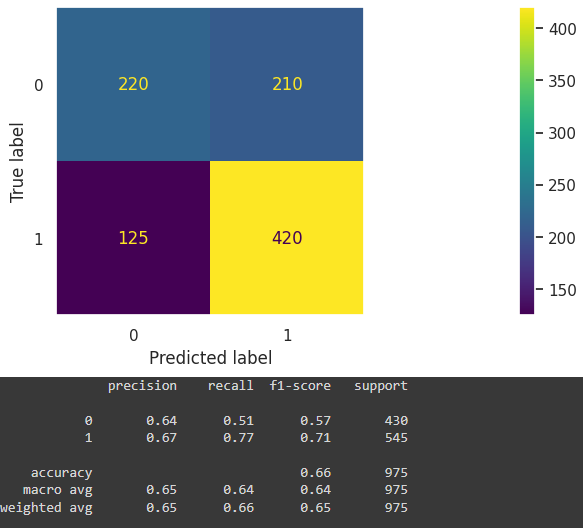
AUC and ROC for the training data(69.7) AUC and ROC for the testing data(65.9)

A graph with a line

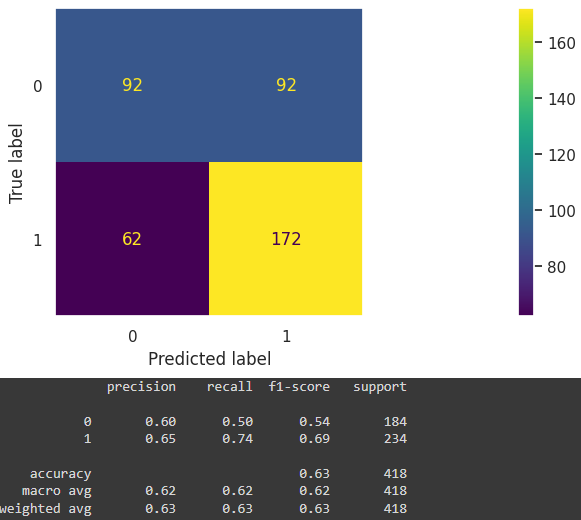
Description automatically generated A graph with a line

Description automatically generated

Confusion Matrix and Classification Report for training Data

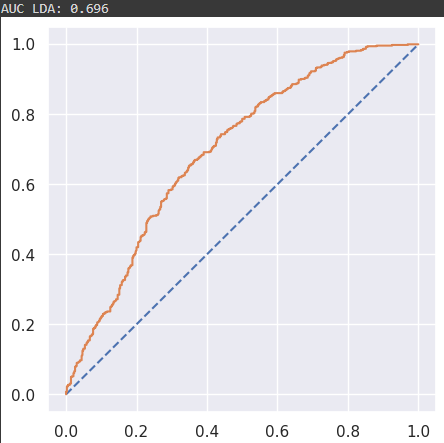


Confusion Matrix and Classification Report for testing Data



**Linear Discriminant Analysis(LDA):**

AUC and ROC for the training data(69.6) AUC and ROC for the testing data(65.8)

 A graph with a line

Description automatically generated

Confusion Matrix and Classification Report for training Data

A screenshot of a graph

Description automatically generated

Confusion Matrix and Classification Report for testing Data

A screenshot of a graph

Description automatically generated

**Classification and Regression Tree(CART):**

AUC and ROC for the training data(65.0) AUC and ROC for the testing data(63.7)

A graph with a line

Description automatically generated A graph with a line

Description automatically generated

Confusion Matrix and Classification Report for training Data

A screenshot of a graph

Description automatically generated

Confusion Matrix and Classification Report for testing Data

A screenshot of a graph

Description automatically generated

**Comparison of score from all the models:**



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

**Logistic Regression:**

Inferences- from Logistic Regression Model :

Contraceptive method used : **0 indicates No, 1 indicated Yes**

For predicting that person do not use contraceptive method (Label-0 ):

Precision (60%) – 60% of females predicted are not using contraceptives out of all females that were predicted as not using contraceptives.

Recall (50%) – Out of all the females who didnt take contraceptives, 50% of females have been predicted correctly .

For predicting that a person takes contraceptive(Label-1 ):

Precision (65%) – Of all the females predicted to use contraceptives, 65% of females actually use contraceptives.

**Recall (74%) – Out of all the females actually taking contraceptives , 74% of females have been predicted correctly .**

Linear Equation from Logistic Regression to find Contraceptive\_method\_used:

Contraceptive\_method\_used = 1.038 + Wife\_age\*(-0.068) + Wife\_education\*(0.452) + Husband\_education\*(-0.114) + No\_of\_children\_born\*(0.248) + Wife\_religion\*(0.697) + Wife\_Working\*(-0.273) + Husband\_Occupation\*(0.076) + Standard\_of\_living\_index\*(-0.185) + Media\_exposure\*(-0.571)

**Linear Discriminant Analysis:**

Inferences- from Linear Discriminant Analsyis(LDA) :

Contraceptive method used : 0 indicates No, 1 indicated Yes

For predicting that person do not use contraceptive method (Label-0 ):

Precision (58%) – 58% of females predicted are not using contraceptives out of all females that were predicted as not using contraceptives.

Recall (47%) – Out of all the females who didnt take contraceptives, 47% of females have been predicted correctly .

For predicting that a person takes cotraceptive(Label-1 ):

Precision (64%) – Of all the females predicted to use contraceptives, 64% of females actually use contraceptives.

**Recall (74%) – Out of all the females actually taking contraceptives , 74% of females have been predicted correctly .**

Linear Equation from Linear Discriminant Analysis to find Contraceptive\_method\_used:

Contraceptive\_method\_used = 1.26 + Wife\_age\*(-0.072) + Wife\_education\*(0.461) + Husband\_education\*(-0.13) + No\_of\_children\_born\*(0.254) + Wife\_religion\*(0.732) + Wife\_Working\*(-0.291) + Husband\_Occupation\*(0.053) + Standard\_of\_living\_index\*(-0.201) + Media\_exposure\*(-0.61)

**CART:**

Contraceptive method used : 0 indicates No, 1 indicated Yes

For predicting that person do not use contraceptive method (Label-0 ):

Precision (59%) – 59% of females predicted are not using contraceptives out of all females that were predicted as not using contraceptives.

Recall (53%) – Out of all the females who didnt take contraceptives, 53% of females have been predicted correctly .

For predicting that a person takes cotraceptive(Label-1 ):

Precision (66%) – Of all the females predicted to use contraceptives, 66% of females actually use contraceptives.

**Recall (71%) – Out of all the females actually taking contraceptives , 71% of females have been predicted correctly .**

As per CART, only Wife\_education and No\_of\_children\_born are enough to predict the target column.

**Conclusion:**



**Though LR and LDA models are giving 74% Recall value for contraceptive taken(Yes), as per above graph, LDA model score has been reduced in testing set. CART model gives us the Recall value of 71% which is relatively less.**

**Hence it is suggested to use LR model for prediction for this dataset.**

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

Below are the steps performed while doing the predictions in this project,

1. Data Collection
2. Exploratory Data Analysis
3. Data Cleansing
   1. Null handling
   2. Outlier Removal
   3. Duplicate record Handling
   4. Datatype management
4. Data Split into training and testing sets
5. Model Building
6. Testing Model Predictions
7. Checking and Comparing model performance obtained by passing different parameters each time, also used below to quantify performance
   1. AUC
   2. ROC
   3. RSME
   4. R2 value
   5. Adjusted R2 value
8. Documenting Inferences