|  |  |
| --- | --- |
|  |  |
| PREDICTIVE MODELLING PROJECT |  |
|  |  |
|  | Submitted by,VIDYA V |
|  | PGPDSBA.O.2023.B 06.08.2023 |

**CONTENTS**

|  |  |
| --- | --- |
| Case 1: Compactiv Data- Linear Regression | 4 |
| 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 |
| 1. 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. | 11 |
| 1. 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. | 12 |
| 1. Inference: Basis on these predictions, what are the business insights and recommendations. | 12 |
| Case 2: Contraceptive method data- Logistic Regression, LDA, CART | 15 |
| 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. | 15 |
| 1. 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. | 20 |
| 1. 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. | 23 |
| 1. Inference: Basis on these predictions, what are the insights and recommendations. | 26 |

**List of Figures**

|  |  |
| --- | --- |
| **Name** | **Page No.** |
| **Fig 1.1 Dataset Info** | 4 |
| **Fig.1.2. Data Description** | 5 |
| **Fig.1.3. Univariate Analysis- Numerical Columns** | 6-9 |
| **Fig.1.4. Univariate Analysis- Categorical Columns** | 9 |
| **Fig.1.5. Heatmap of fields** | 10 |
| **Fig.1.6. Multivariate Analysis** | 11 |
| **Fig.2.1. Info of Dataset** | 14 |
| **Fig.2.2. Data description** | 14 |
| **Fig.2.3. Univariate Analysis- Numerical Columns** | 15 |
| **Fig.2.4. Univariate Analysis- Categorical columns** | 16 |
| **Fig.2.5. Multivariate Analysis** | 17-18 |
| **Fig.2.6. Regularized Decision Tree (Pruned)** | 20 |
| **Fig.2.7. Logistic Regression scores** | 21 |
| **Fig.2.8. LDA scores** | 21 |
| **Fig.2.9. Regularized Decision Tree scores** | 22 |
| **Fig.2.10. Accuracy and scores comparison across models for Train and test Data** | 23 |
| **Fig.2.11. Confusion Matrix comparisons** | 23 |
| **Fig.2.13. ROC Curves across models for train and test data** | 24 |
| **Fig.2.14. AUC scores comparison** | 25 |

**List of Figures**

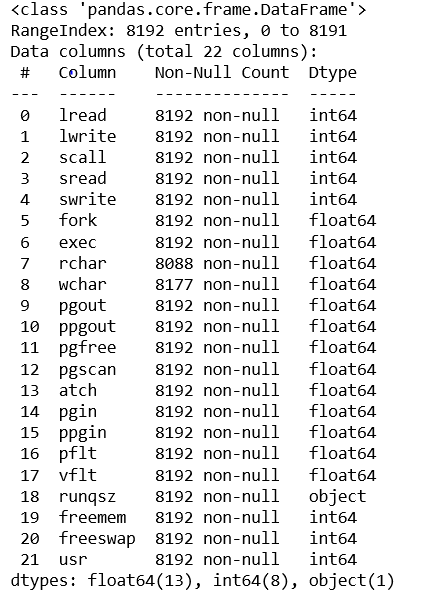
|  |  |
| --- | --- |
| **Name** | **Page No.** |
| **Table.1.1 Linear Regression- Performance Comparison** | 12 |
| **Table 2.1. Type I and II Error comparison across models** | 24 |

# Case 1: Compactiv Data- Linear Regression

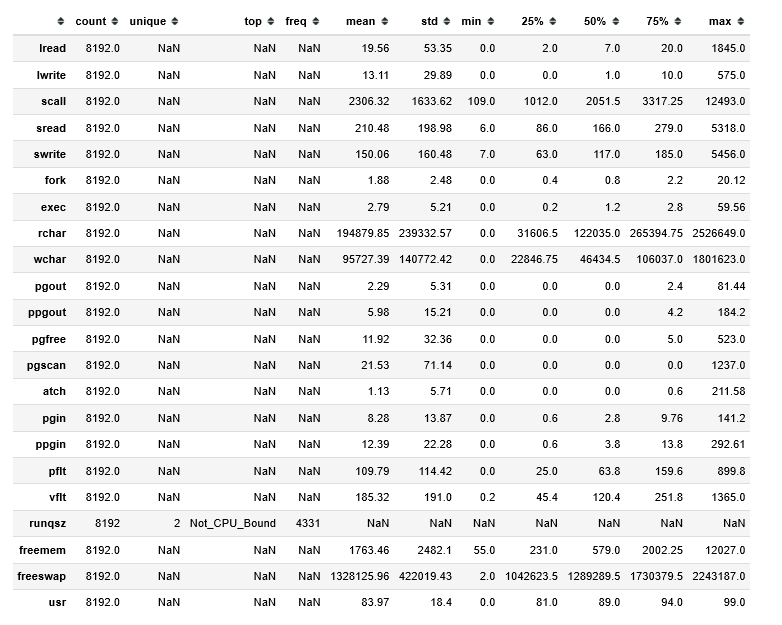
The comp-activ databases is a collection of a computer systems activity measures .  
The data was collected from a Sun Sparcstation 20/712 with 128 Mbytes of memory running in a multi-user university department. Users would typically be doing a large variety of tasks ranging from accessing the internet,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

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.**



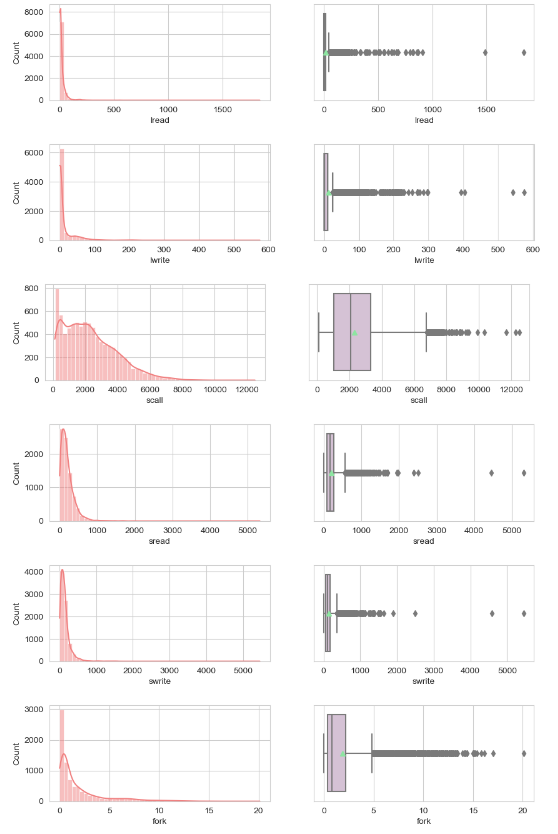
**Fig.1.1 Dataset info**

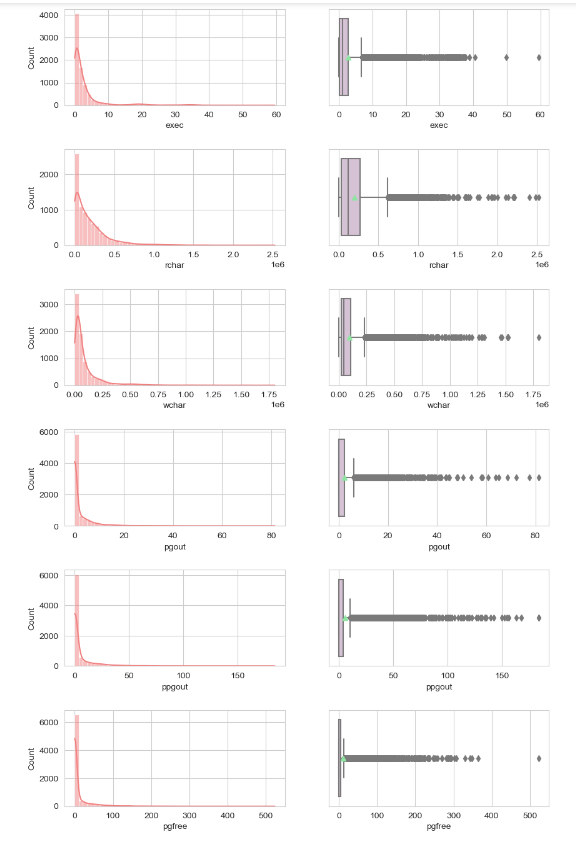


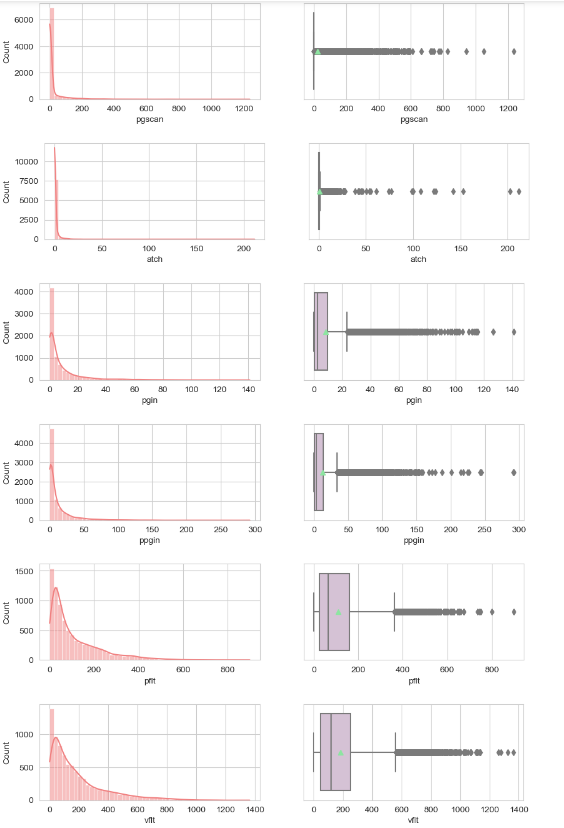
**Fig.1.2. Data Description**

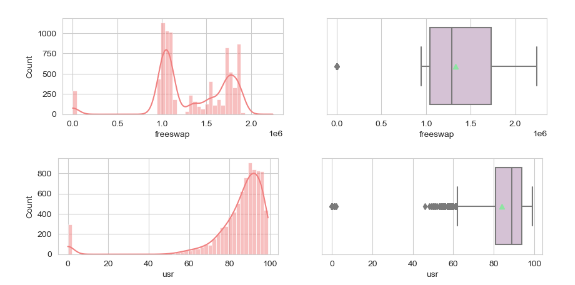
**Observations:**

* There are 8192 datapoints and 22 fields in the given dataset
* All the fields are numerical, except for 'runqsz' which has two categories- CPU bound and Non-CPU-bound
* The field 'usr', which is the percentage of time the CPU runs in user mode, is the target variable for the linear regression
* The fields 'rchar' and 'wchar' have NaN values. These represent the number of characters transferred during system read and write calls.
* Based on domain knowledge, it is plausible that either of these can be 0, but not both 0 for a given datapoint. So, for further analysis, we choose to impute with 0. As the number is very limited when compared to the depth of the dataset (119 in 8192), the choice is justified.
* There are no bad values in the dataset
* Some fields have outliers
* There are no duplicate values
  1. **EDA**

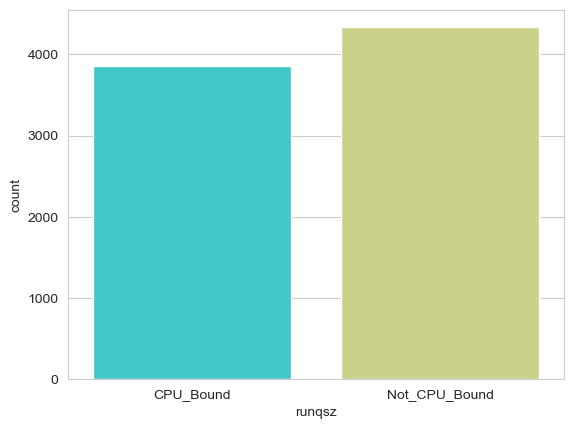




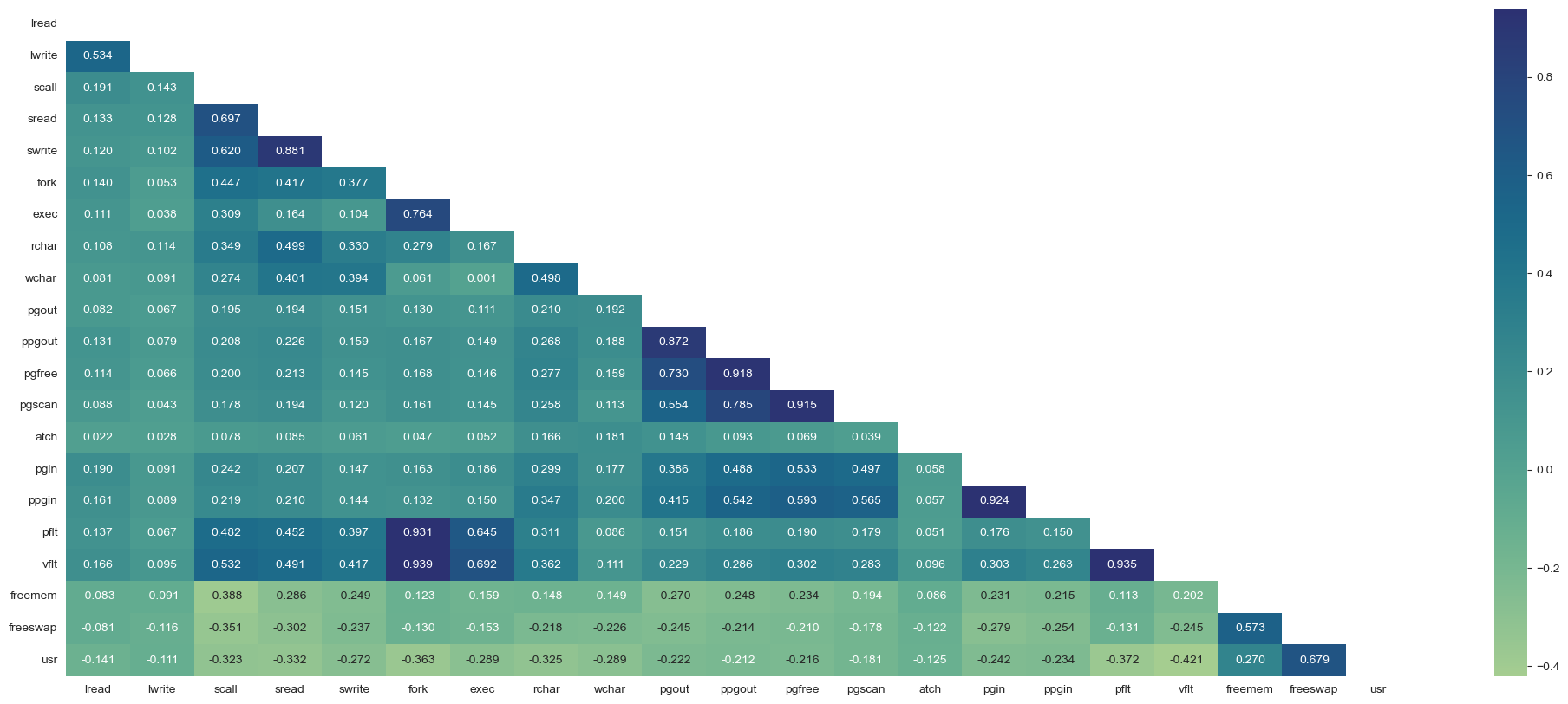




**Fig.1.3 Univariate Analysis- Numerical Columns**

****

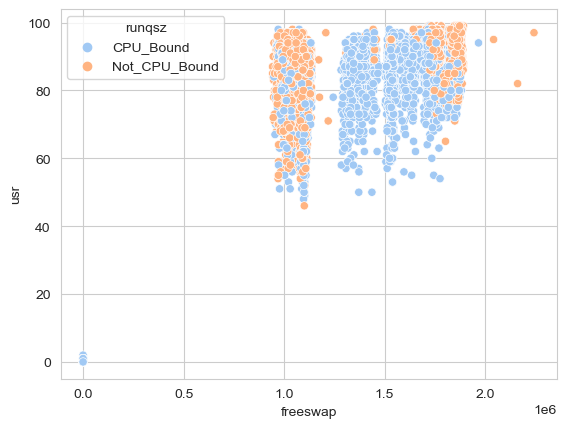
**Fig.1.4. Univariate Analysis- Categorical Column**

****

**Fig.1.5 Heatmap of fields**

**Observations:**

* Univariate analysis of numerical fields:
  + All fields have outliers and hence are skewed
  + The 'freeswap' field has a bimodal distribution
* Univariate analysis of categorical field:
  + There are more instances of 'Not\_CPU\_Bound' than 'CPU\_Bound'
* Bivariate Analysis:
  + Few fields like pflt, vflt and fork, pgout and ppgout, pgscan and pgfree exhibit a very strong correlation
  + However, none of the fields have a very strong correlation to the target variable- usr

****

**Fig.1.6. Multivariate Analysis**

1. **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.**

**Observations:**

* The fields 'rchar' and 'wchar' have NaN values. These represent the number of characters transferred during system read and write calls.
* Based on domain knowledge, it is plausible that either of these can be 0, but not both 0 for a given datapoint. So, on further exploration we find that atleast one of these is non zero for all datapoints. So, as such, there is no missing data and hence are imputed with 0
* There are no bad values in the dataset
* Some fields have outliers and have been treated with IQR approach
* There are no duplicate values
* Feature engineering:
  + The features of the given dataset were combined as per the formulae below to derive new features:
    - Total\_io: Total\_IO\_Activity = lread + lwrite
    - Total\_disk\_io : Total\_Disk\_IO = sread + swrite
    - Total\_pg: Total\_Page\_Activities = pgout + ppgout + pgfree + pgscan + atch + pgin + ppgin + pflt + vflt
    - Total\_proc: Total\_Process\_Activities = fork + exec
    - Disk\_mem\_usage: Total\_Disk\_Memory\_Usage = freemem + freeswap
    - Total\_scalls: Total\_System\_Calls = scall + fork + exec
    - Char\_total: Total\_Characters\_Transferred = rchar + wchar
    - Io\_scall\_ratio: Total\_io/Total\_scalls
    - Total\_pg\_faults= pflt+vflt

1. **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.**

For the sake of analysis- 2 different approaches were adopted, based on which 2 different models were built

**Model 1**: Without feature Engineering

**Model 2:** With Feature Engineering

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Performance Metric | MODEL 1- No Feature Engineering | | | | | | MODEL 2- With Feature engineering | | | | | |
| **Scikit learn Linear Regression** | | **Stats models OLS** | | **Post-VIF drop** | | **Scikit learn Linear Regression** | | **Stats models OLS** | | **Post-VIF drop** | |
| **Train** | **Test** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| R square/Score | 0.786 | 0.793 | 0.786 | 0.797 | 0.963 | 0.967 | 0.779 | 0.785 | 0.779 | 0.788 | 0.961 | 0.963 |
| Adj R square | - | - | 0.785 | 0.795 | 0.963 | 0.967 | - | - | 0.778 | 0.787 | 0.961 | 0.963 |
| RMSE | 4.546 | 4.357 | 219 | 19 |  |  | 4.620 | 4.436 | 21 | 19 |  |  |

**Table.1.1 Linear Regression- Performance Comparison**

**Observations**:

* Comparison between Models:
  + Model 1, i.e., the model without feature engineering performs better in all aspects
* Comparison Between Methods:
  + Sklearn linear Regressions have better RMSE for both the models, but statsmodels OLS have better R square values
* Variance Influence factor analysis:
  + For both the models, after analyzing the most factor that most influences variance and dropping, the R square and adjusted R square values made a significant jump.
  + The difference of the R square and adjusted R square values before and after dropping the factor of most influence was 0.18.
  + However, beyond this, the r square and adjusted r square values showed no improvement for subsequent dropping of the most influential factors.

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

**Linear Equation- sklearn**

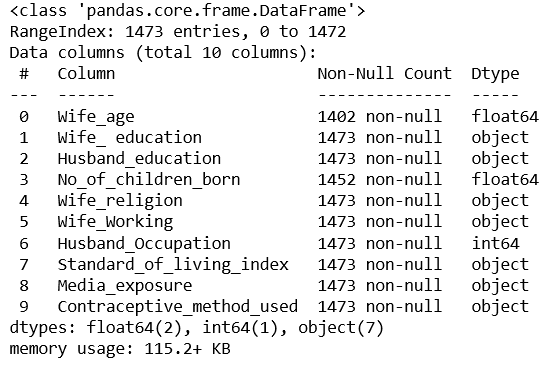
usr = (-0.05420 \* lread) + (0.04506 \* lwrite) - (0.00076 \* scall) + (0.00228 \* sread) - (0.00495 \* swrite) - (0.14707 \* fork) - (0.23054 \* exec) - (0.00000 \* rchar) - (0.00000 \* wchar) - (0.45885 \* pgout) + (0.03462 \* ppgout) + (0.04065 \* pgfree) + (0.00000 \* pgscan) + (0.51061 \* atch) + (0.00700 \* pgin) - (0.05910 \* ppgin) - (0.03145 \* pflt) - (0.00639 \* vflt) - (0.00053 \* runqsz) + (0.00001 \* freemem) + (1.84976 \* freeswap)

**Observations and Inferences:**

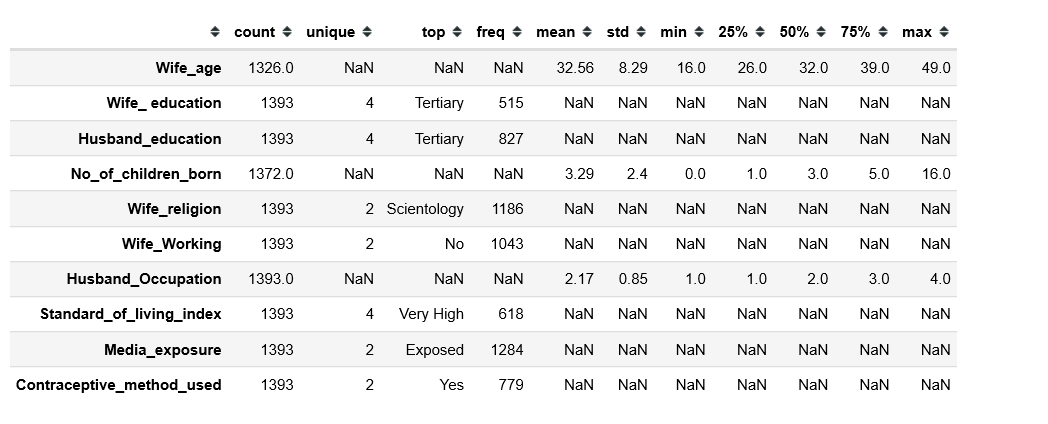
* Based on the above equation, it can be inferred that the variables- freeswap, atch, and pgout contribute most to the percentage of time spent in usr mode, and hence are the most significant
* Variables with positive coefficients (e.g., `lwrite`, `sread`, `ppgout`, `pgfree`, `atch`, `pgin`, `freemem`, `freeswap`) have a positive impact on the target variable `usr`. An increase in these variables is associated with an increase in the portion of time that CPUs run in user mode.
* Variables with negative coefficients (e.g., `lread`, `scall`, `swrite`, `fork`, `exec`, `pgout`, `ppgin`, `pflt`, `vflt`, `runqsz`) have a negative impact on the target variable `usr`. An increase in these variables is associated with a decrease in the portion of time that CPUs run in user mode.
* Some input variables have zero coefficients (`rchar`, `wchar`, `pgscan`). This indicates that changes in these variables do not significantly impact the target variable.

# Case 2: Contraceptive dataset- Logistic Regression, LDA and CART

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.**



**Fig.2.1. Info of dataset**

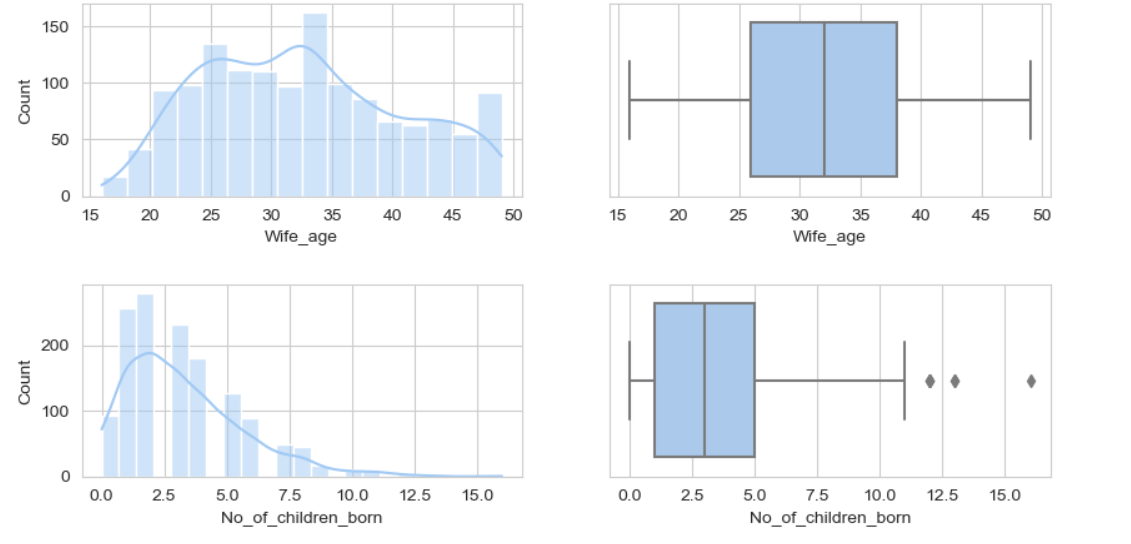


**Fig.2.2 Data description**

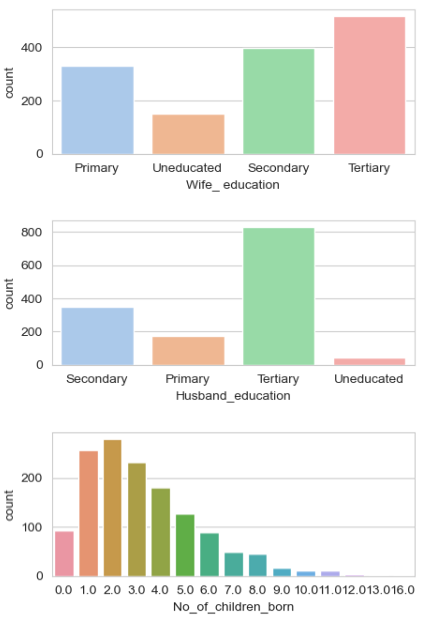
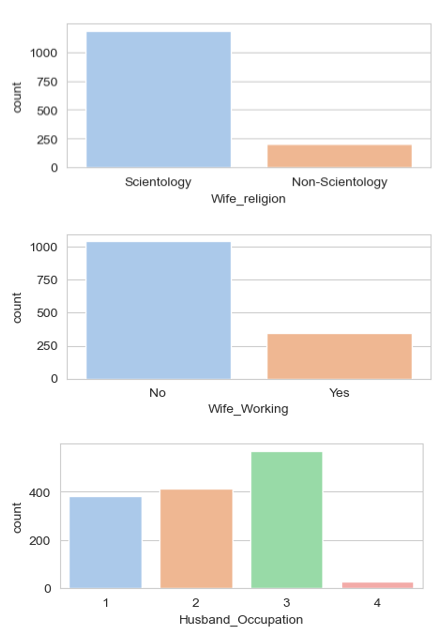
**Observations:**

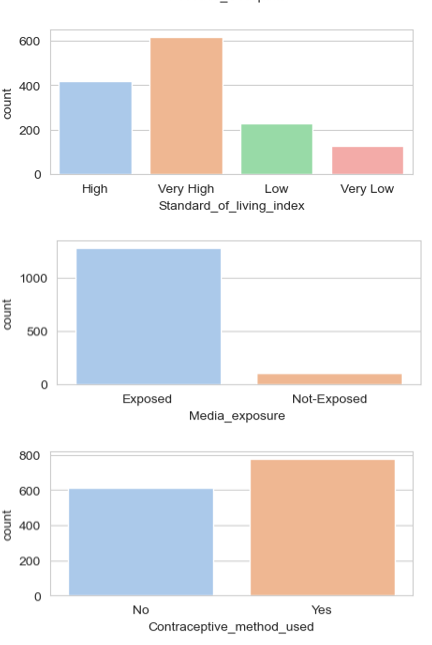
* Wife\_age, No\_of\_children\_born contains null values
* Of these, it is possible for the Number of children to be 0, hence need not be imputed
* There are blank values present in No\_of\_children\_born. These are imputed with mode of the field
* There were 80 duplicate records, which were removed.
* No\_of\_children field has outliers, but not very significant

**2.1. EDA**



**Fig.2.3. Univariate analysis- Numerical columns**

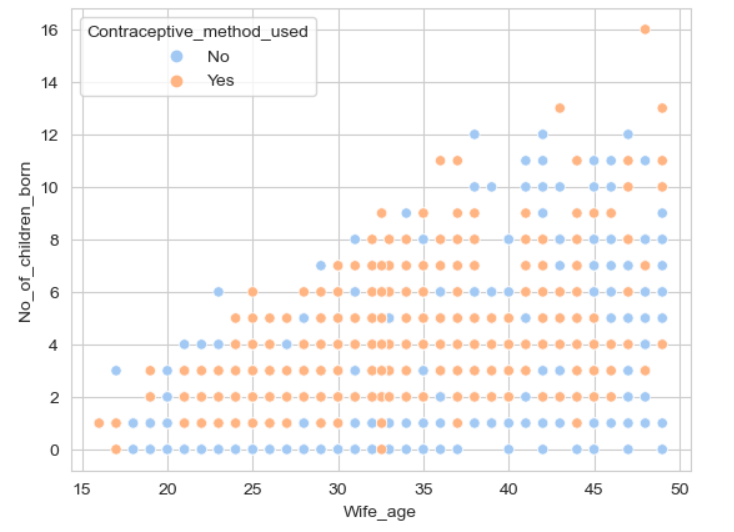
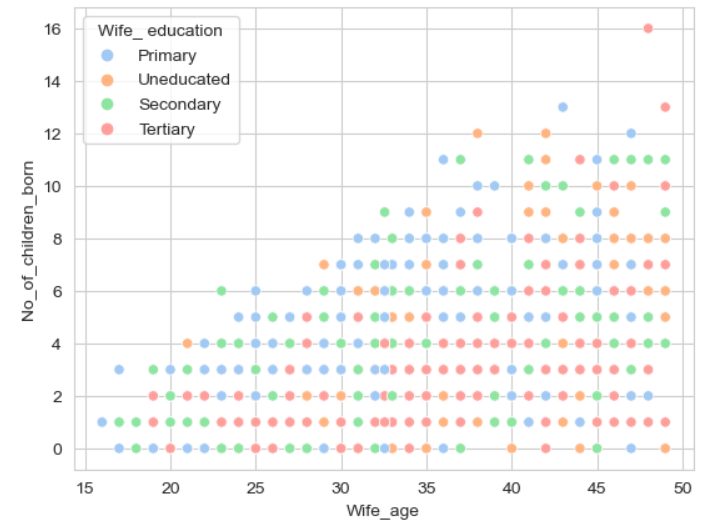
 

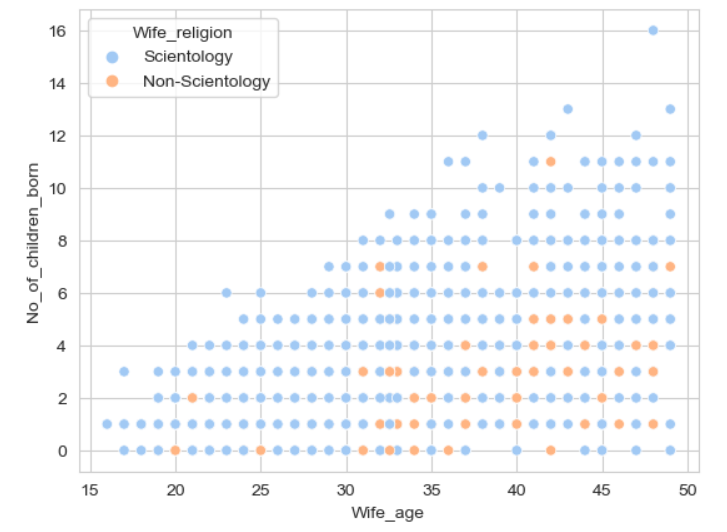


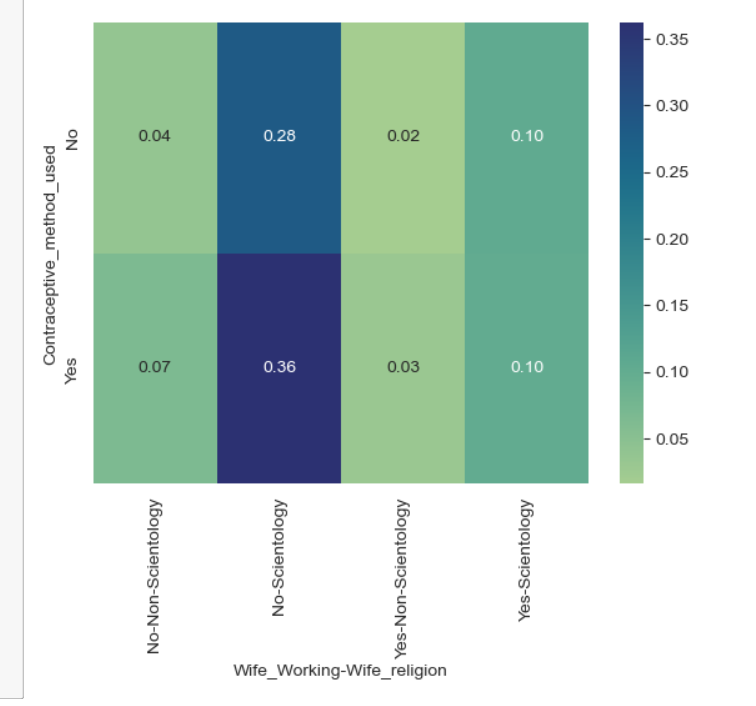
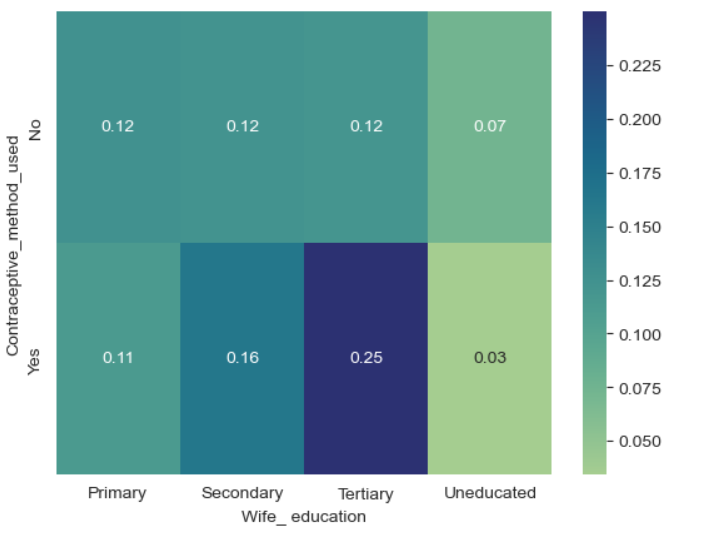
**Fig.2.4. Univariate analysis- categorical columns**

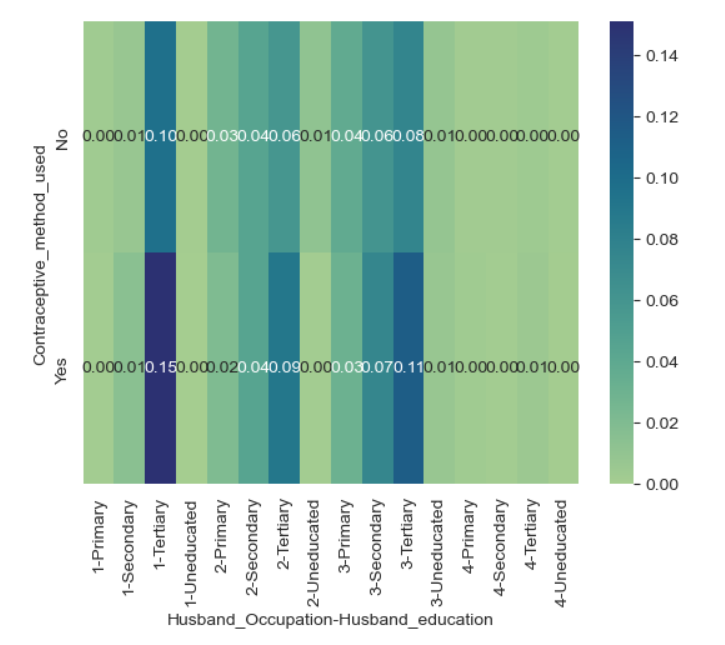
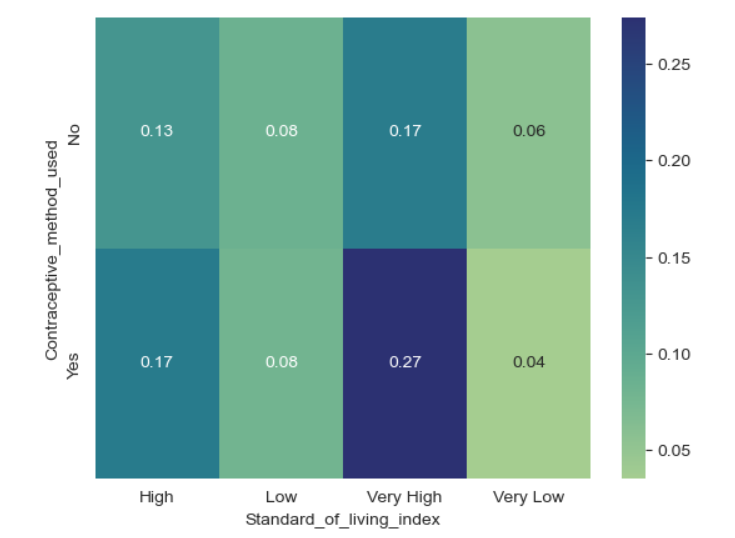
**Observations:**

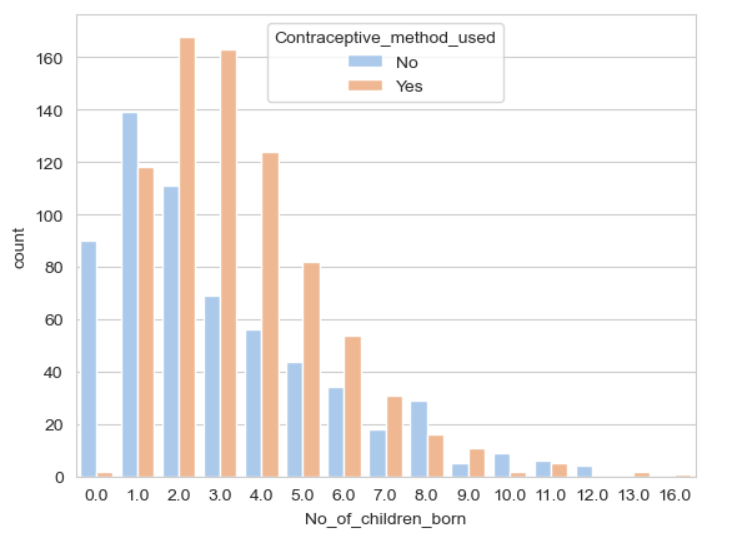
* The age of wife is almost normally distributed
* There are a few outliers in the Number\_of\_children\_born field
* Most women are educated, with the highest number of women having completed teritiary education
* The husband\_education field also has maximum values in teritiary education field, with only a very low count of uneducated males.
* Most have upto 4 children. However, the data also shows people having more than 10 children. These might be genuine, or bad values.
* Most women adopt scientology as religion
* Most of the women are ot working
* A majority of the husbands have occupation -3
* This data has high counts of people having high or very high standard of living index
* Most of these people are media exposed
* most of these people have used contraceptive methods





**Fig.2.5. Multivariate Analysis**

**Observations:**

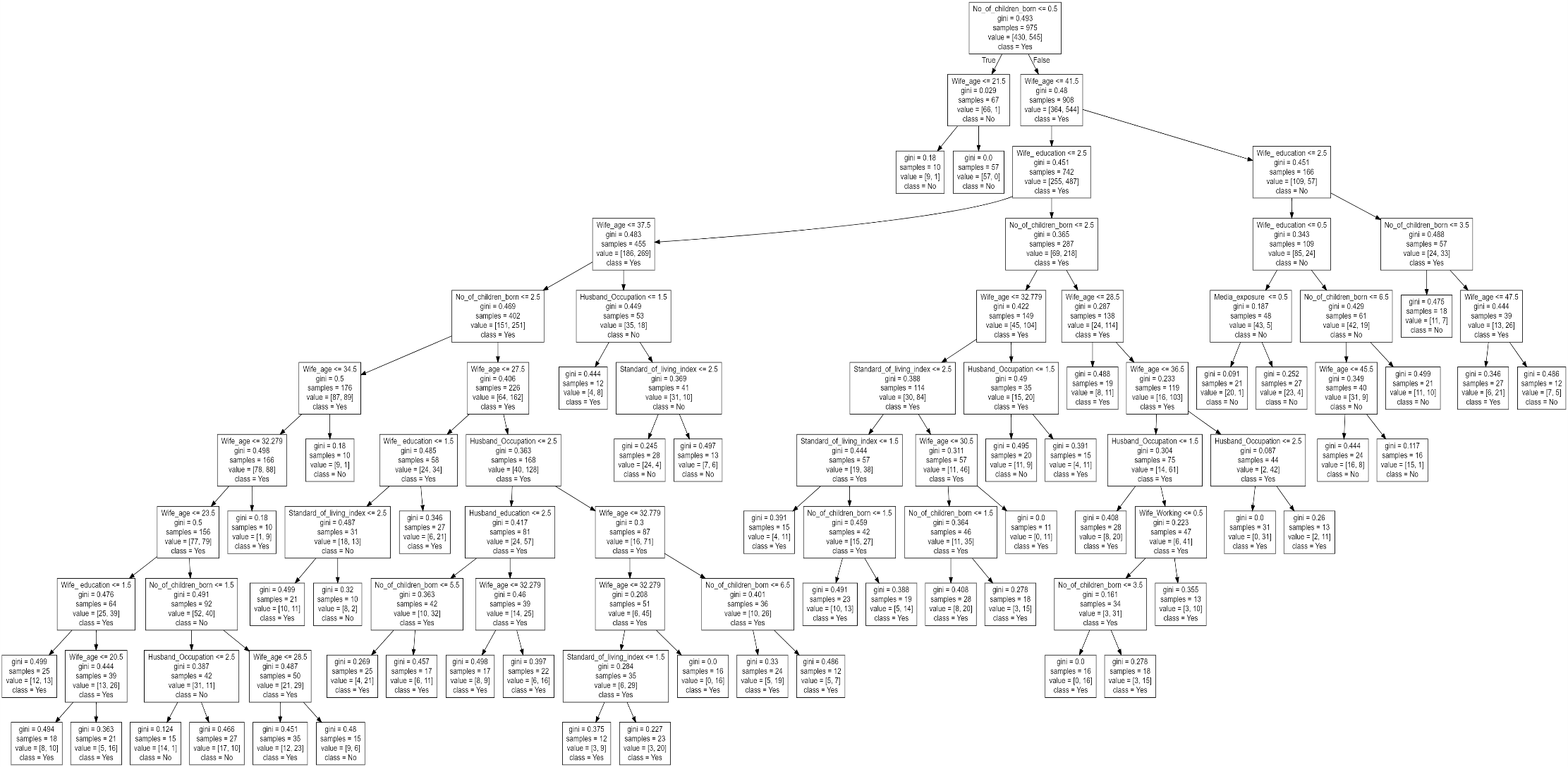
* No discernable patterns emerge from the scatter plots
* Some minor correlation detected between the status and usage of contraceptives

1. **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.**

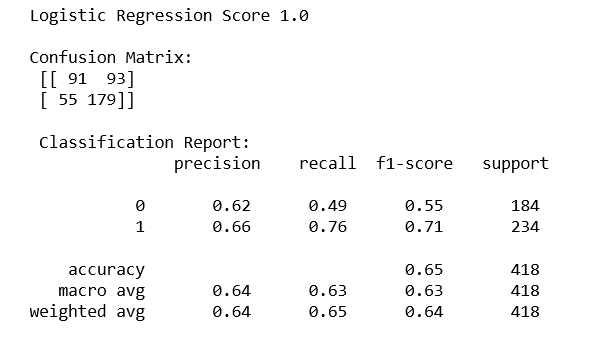
**The following equations are obtained after the application of logistic regression and LDA respectively:**

**log(contraceptive usage)=**-0.425 - 0.077 \* Wife\_age + 0.567 \* Wife\_education - 0.048 \* Husband\_education + 0.308 \* No\_of\_children\_born - 0.265 \* Wife\_religion - 0.065 \* Wife\_working + 0.069 \* Husband\_occupation + 0.158 \* Standard\_of\_living\_index + 0.314 \* Media\_exposure

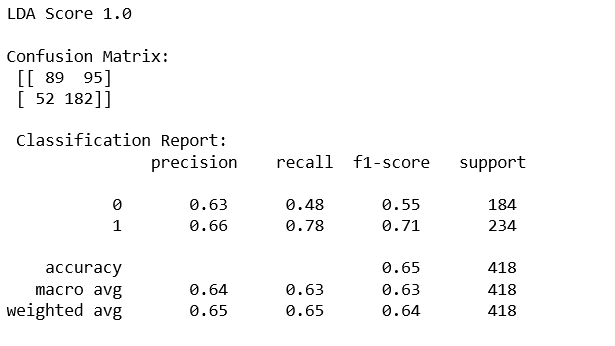
**contraceptive\_usage=** 0.286- 0.075 \* Wife\_age + 0.572 \* Wife\_education - 0.057 \* Husband\_education + 0.297 \* No\_of\_children\_born - 0.287 \* Wife\_religion - 0.071 \* Wife\_working + 0.069 \* Husband\_occupation + 0.158 \* Standard\_of\_living\_index + 0.299\* Media\_exposure



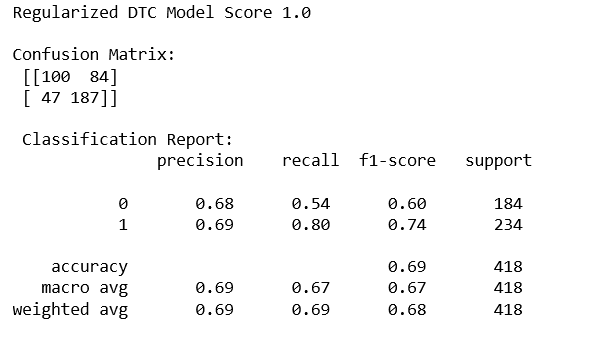
**Fig 2.6. Regularized Decision Tree(Pruned)**



**Fig.2.7. Logistic regression scores**



**Fig. 2.8. LDA scores**

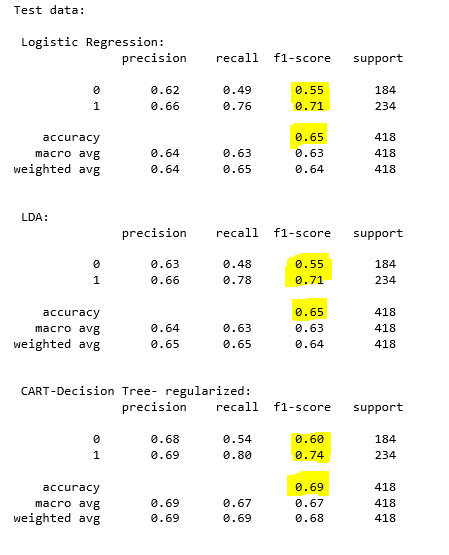
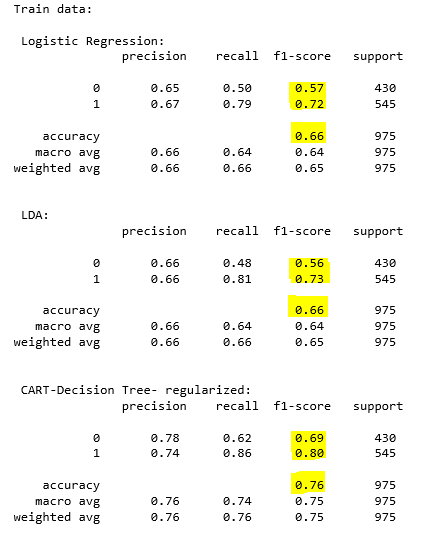


**Fig.2.9. Regularized Decision Tree scores**

**Observations- Regularized Decision Tree:**

* The output of the model is a pruned tree
* There are 10 levels
* 1 feature has a coefficient of 0- Wife\_religion, and 2 have coefficients close to 0- Wife\_working, Media\_exposure.

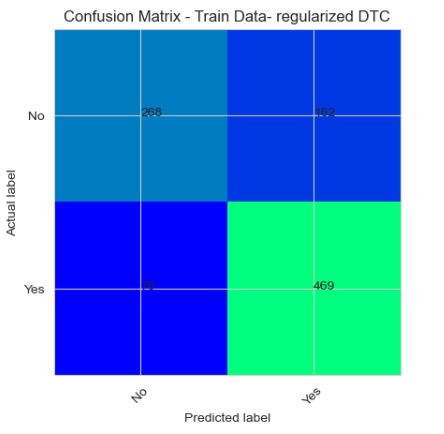
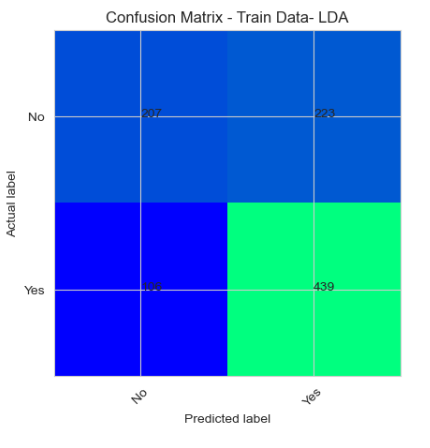
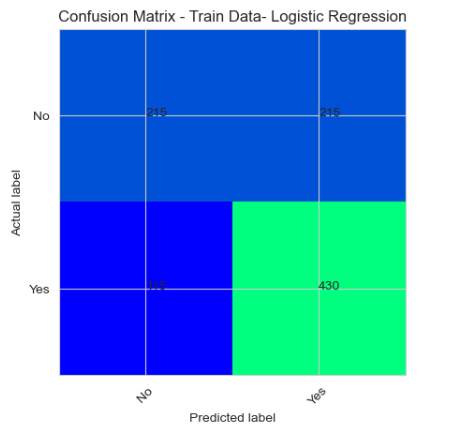
1. **Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC\_AUC score for each model Final Model: Compare Both the models and write inference which model is best/optimized.**

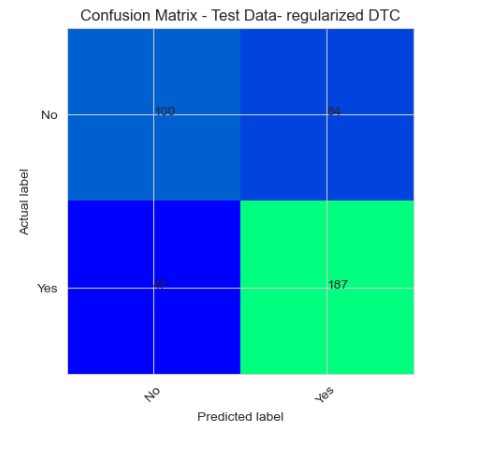
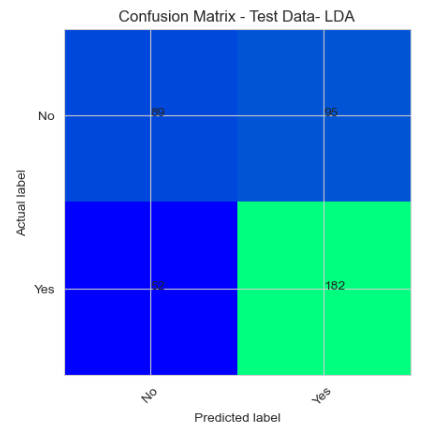
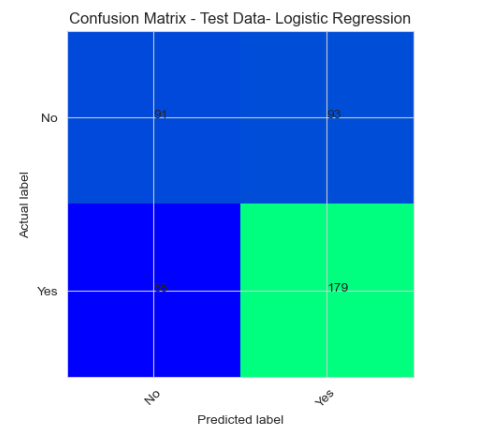


**Fig.2.10. Accuracy and scores comparison across models for Train and test Data**

**Observations:**

* Train Data:
  + CART is better in terms of accuracy- 0.76 against 0.66 for LDA and Logistic Regression
  + CART also gives better F1 scores for both the classes (0.69 and 0.80) as against LDA (0.56 and 0.73) and Logistic regression (0.57 and 0.72)
* Test Data:
  + CART is better in terms of accuracy- 0.69 against 0.65 for LDA and Logistic Regression
  + CART also gives better F1 scores for both the classes (0.60 and 0.74) as against LDA and Logistic regression (0.55 and 0.71)





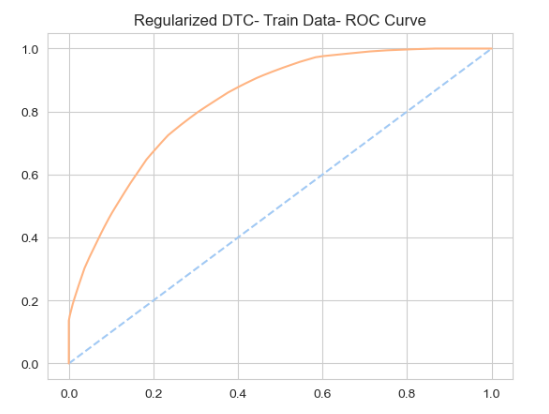
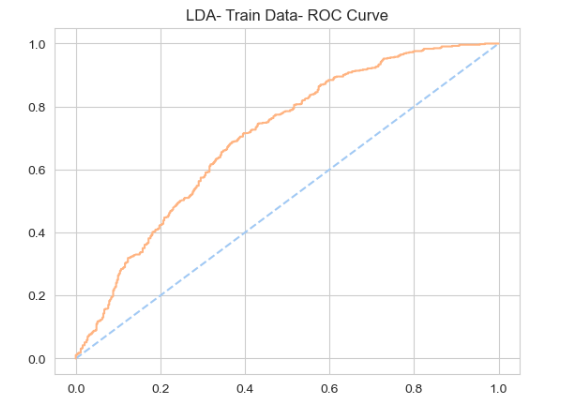
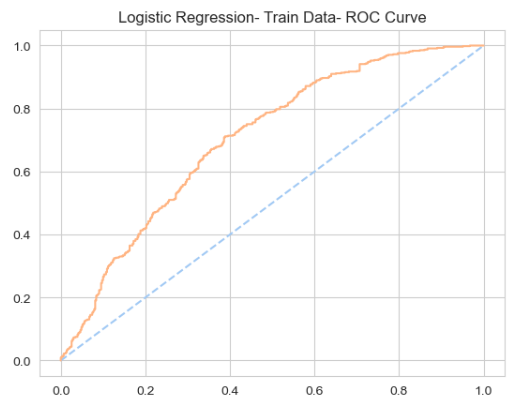
**Fig.2.11. Confusion Matrix comparisons**

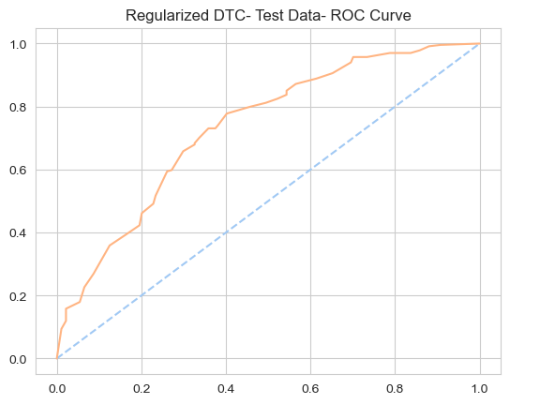
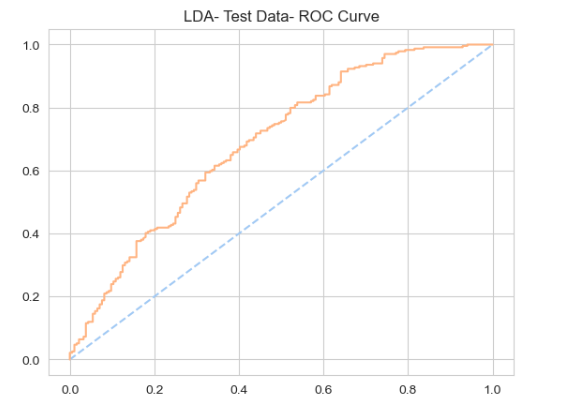
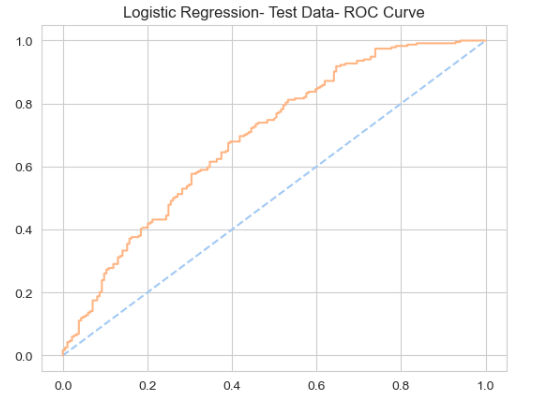
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Logistic Regression | | LDA | | CART- Regularized | |
| **Train** | **Test** | **Train** | **Test** | **Train** | **Test** |
| Type I Error | **215** | **93** | **223** | **96** | **162** | **84** |
| Type II Error | **115** | **55** | **106** | **52** | **76** | **47** |

**Table 2.1. Type I and II Error comparison across models**

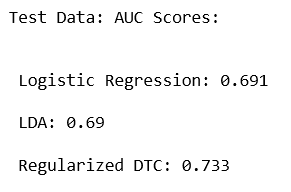
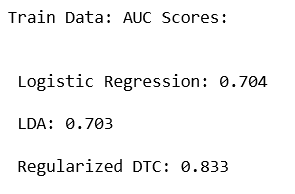
**Observations:**

* From the above, it can be inferred that CART model outperforms the other two





**Fig.2.13. ROC Curves across models for train and test data**



**Fig.2.14. AUC scores comparison**

**Observations:**

* From the graphs and scores above, it can be inferred that of the three models, CART gives better fit in both train and test sets.

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

**Summary:**

* For the given dataset, the performance of LDA and logistic regression were similar. However, CART gave better results
* This can be an indication of a non-linear relation between the predictors and target
* Based on the equations derived from LDA and logistic regression, the features that create maximum separability are Wife\_education, Media\_exposure and Number\_of\_children\_born
* Hence, in order to improve contraceptive usage these actors could be targeted