Contraceptive Prevalence Survey – Logistic-Regression

Name :

Apoorv Purohit

Padmini Subramanian

Preyal Deep Chhabra

Rathi Sadhasivan

Renuka Prasad GM

Sukrit Kalia

PGP-DSBA Offline

August - 23

Date: 22/12/2023

Contraceptive Prevalence Survey.: Logistic Regression – Business Report

Contents

List of Tables

List of Figures

Contraceptive Prevalence Survey.: Logistic Regression – Business Report

Executive Summary

The Republic of Indonesia ministry of health does a contraceptive prevalence survey on married women's. Wife and husband education, number of children born, standard of living index and media exposure variables are more and high chances of using contraceptive.

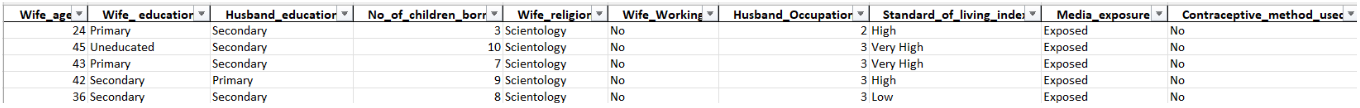
Introduction

Assignment is to deep understanding of dataset and perform exploratory data analysis. Explore datasets with logistic regression to validate, whether married women using contraceptive method/not based on their demographic and socio-economic characteristic's (depends on wife's education, working, religion, etc). The dataset consists of 10 columns having numerical and categorical data and 1473 rows. Analyse different features of categorical data present in dataset and how this data interrelationship with other categorical variables and which variables will help to predict contraceptive women/not using supervised logistic regression approach. Dataset will explore more on summary statistics, probabilities scores, null values, anomalies present in categorical variable, train and test the data under 70/30 combination, encode the data for logistic classification to find the accuracy of the model and data visualization across numerical and categorical subjects. Generate a confusion matrix to give more insight on accurate prediction of contraceptive. Plot ROC - AUC metrics to demonstrate the accuracy of the test and trained data on married women pregnant/not.

Data Description

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

Sample of the dataset:



Wife_age	Wife_ educatior	Husband_educatior	No_of_children_borr	Wife_religior	Wife_Working	Husband_Occupatior	Standard_of_living_inde:	Media_exposure	Contraceptive_method_usec
24	Primary	Secondary	3	Scientology	No	2	High	Exposed	No
45	Uneducated	Secondary	10	Scientology	No	3	Very High	Exposed	No
43	Primary	Secondary	7	Scientology	No	3	Very High	Exposed	No
42	Secondary	Primary	9	Scientology	No	3	High	Exposed	No
36	Secondary	Secondary	8	Scientology	No	3	Low	Exposed	No

Table 1. Dataset Sample

Data has 10 variables with more categorical variables in contraceptive data and which attributes influences more towards the classification prediction

Exploratory Data Analysis

Let’s check types of variables present in data frame

```
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   Wife_age                  1402 non-null   float64
 1   Wife_ education           1473 non-null   object 
 2   Husband_education         1473 non-null   object 
 3   No_of_children_born       1452 non-null   float64
 4   Wife_religion             1473 non-null   object 
 5   Wife_Working              1473 non-null   object 
 6   Husband_Occupation        1473 non-null   int64  
 7   Standard_of_living_index  1473 non-null   object 
 8   Media_exposure            1473 non-null   object 
 9   Contraceptive_method_used 1473 non-null   object 
```

Total of 1473 rows and 10 columns in the dataset. Out of 10 , 7 columns are of categorical type and rest 3 are of either integer or float data type.

Check for missing/null values in the dataset

```
Wife_age                     71
Wife_ education               0
Husband_education             0
No_of_children_born          21
Wife_religion                 0
Wife_Working                  0
Husband_Occupation            0
Standard_of_living_index      0
Media_exposure                0
Contraceptive_method_used     0
dtype: int64
```

From the above data observed that wife age and number of children born attributes having a missing value present in the model.

The NaN values in No_of_children_born could be 0/not. It should be discrete value and not an appropriate way to fill with mean/median. Will perform the following treatments

- No_of_children_born = 2 mode value
- Create a function based on mean wife age and fill accordingly, eg. <30 ~ 1, 30-35 ~ 2, etc.
- Drop 21 null values, which are small compared to total records of 1473

Check for duplicate value treatment

No duplicate value present in the model

Q1: 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.

Description of the variables and measurements for logistic regression analysis of the determinants of contraceptive method utilization among women in Indonesia. The attributes are influencing factor for contraceptive/not like religion, number of children's, working, education, husband education etc. Exploratory data analysis explained in the above slides

Outliers: Represented in boxplot data visualisation

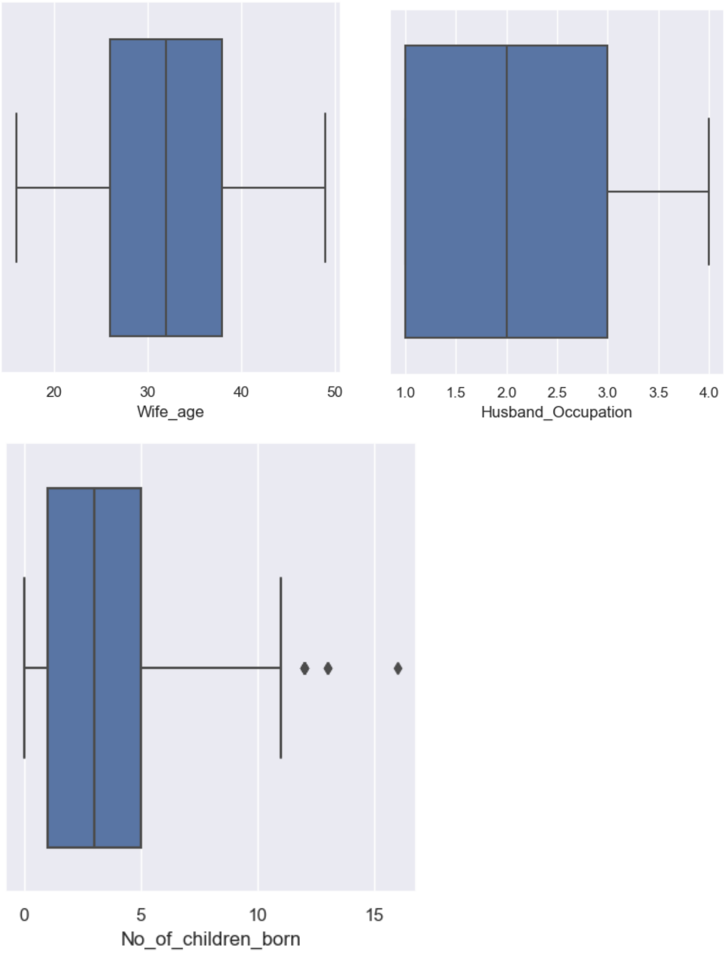


Fig 1. Outliers

Ignore 3 data of outliers in No_of_children_born and also has finite unique values.

Uni-Variate Analysis: Represented in count and heat plot data visualisation

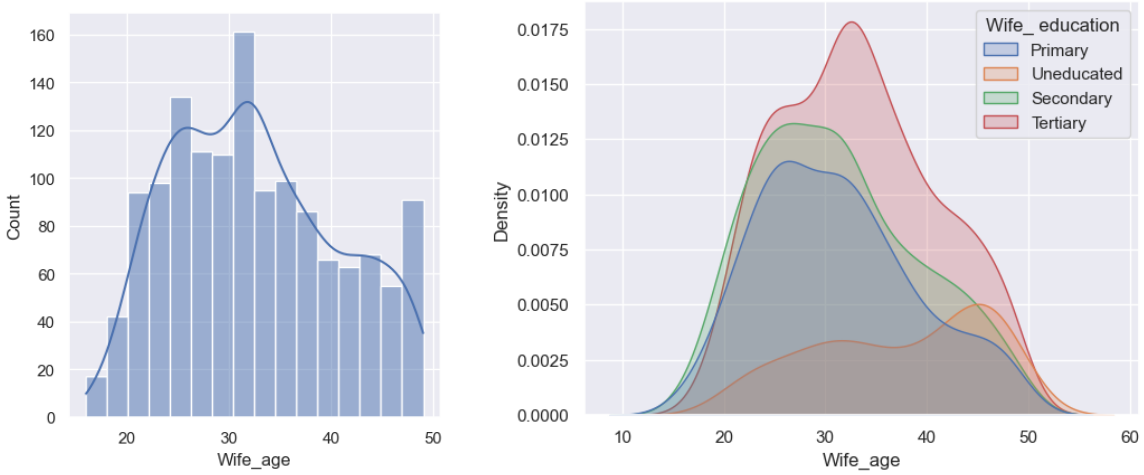


Fig 2. Wife age with level of education

Wife age is not normally distributed as observed in boxplot and overlapping of data observed in histogram plot means that some other attributes are influencing it.

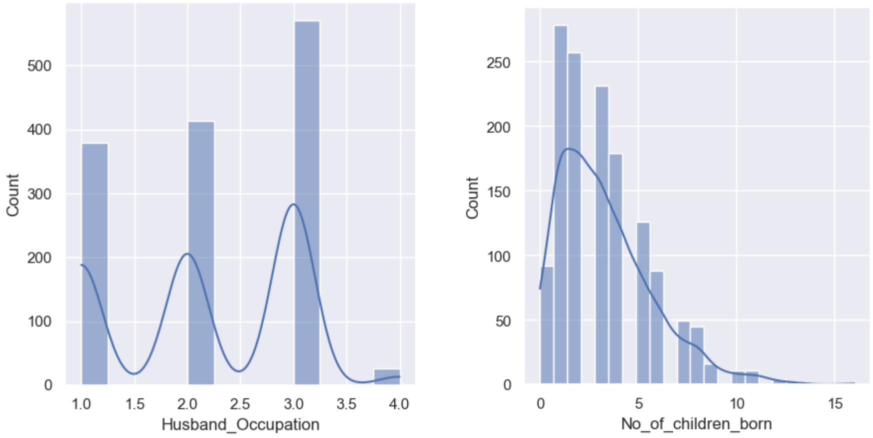


Fig 3. Husband education and No of Children Born

Husband occupation is a discrete categorical and ignore the outliers present in No. of children born

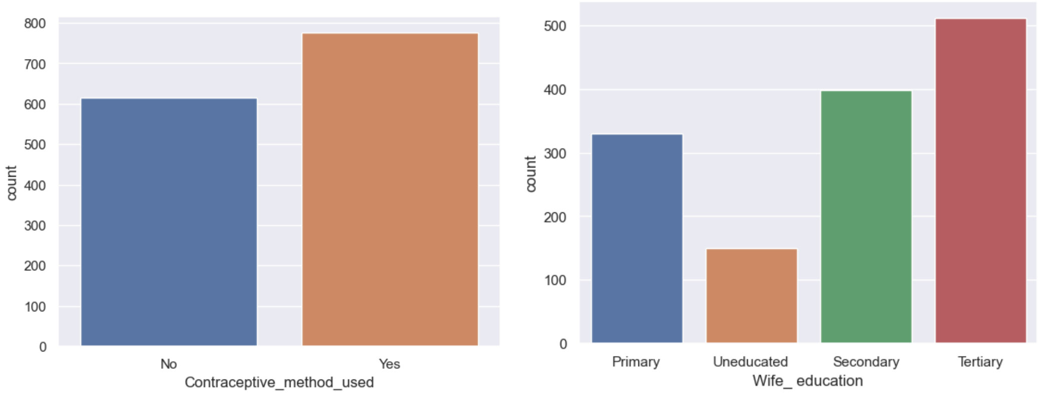


Fig 4. Contraceptive usage with respect to level of education

From the above data, women used 56% contraceptive and 44% non-contraceptive

The dataset has more educated women

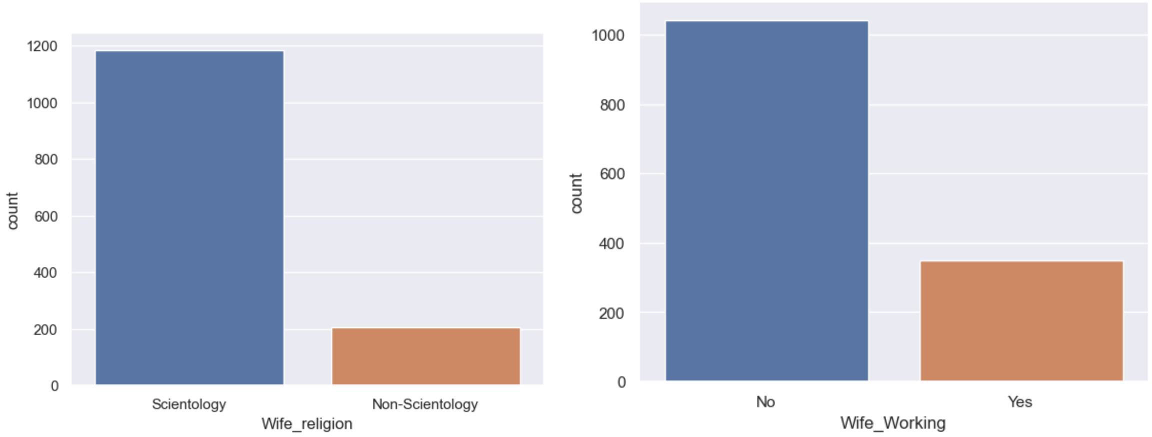


Fig 5. Wife religion and working

The ratio of Scientology to Non-Scientology is very high.

The working ratio is high.

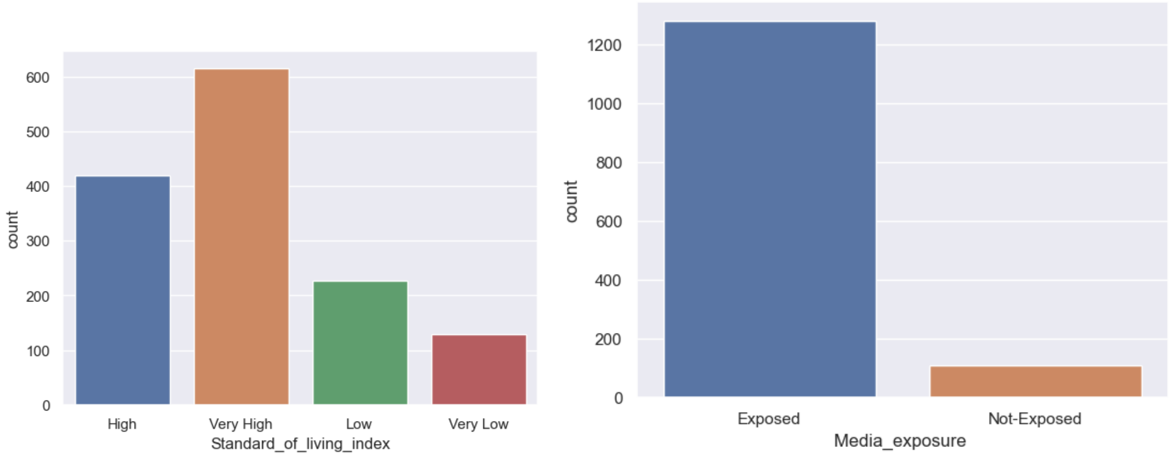


Fig 6. Standard of living index and media exposure

High chances of women exposed to media and majority of women has high standard of living index

Bi-variate and multi-variate Analysis

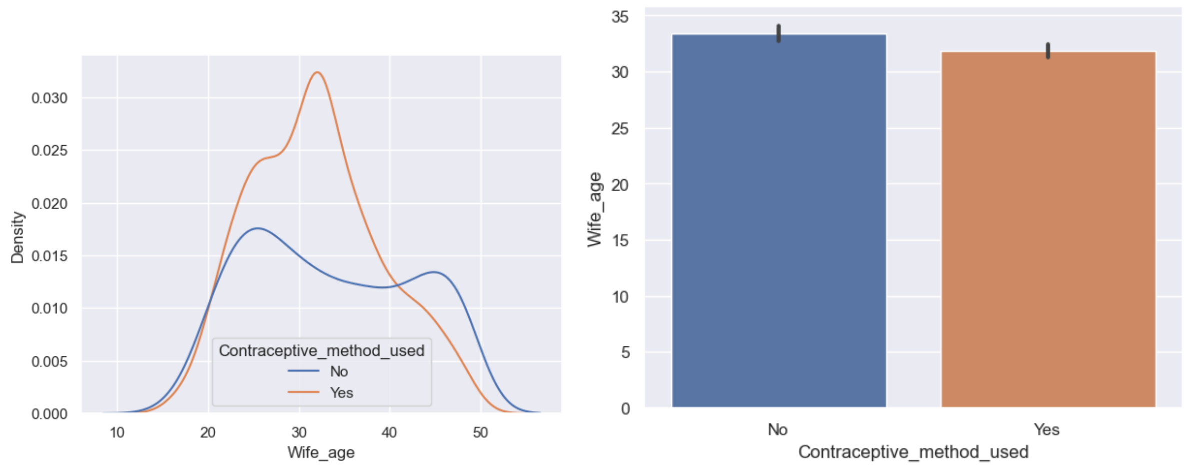


Fig 7. Wife age with contraceptive usage

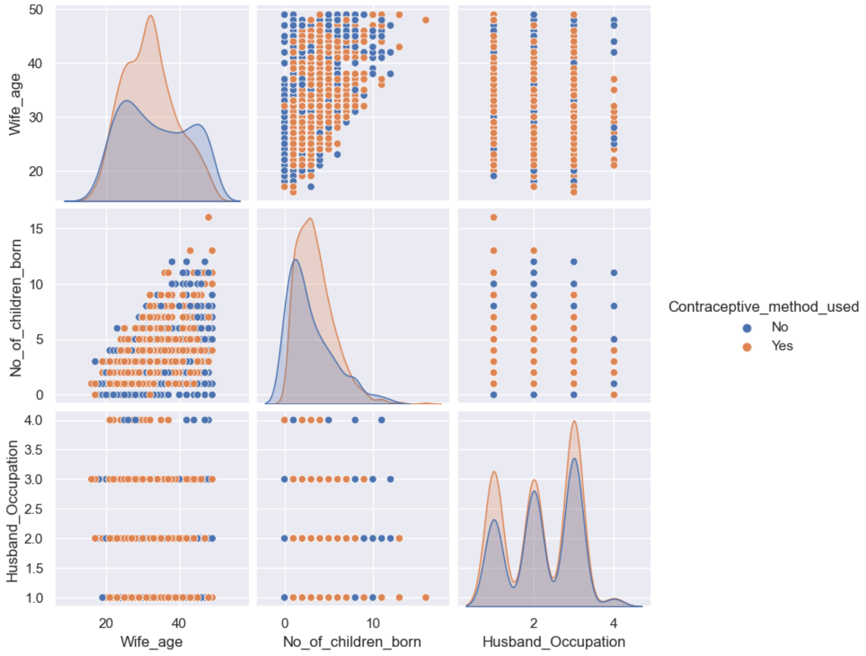
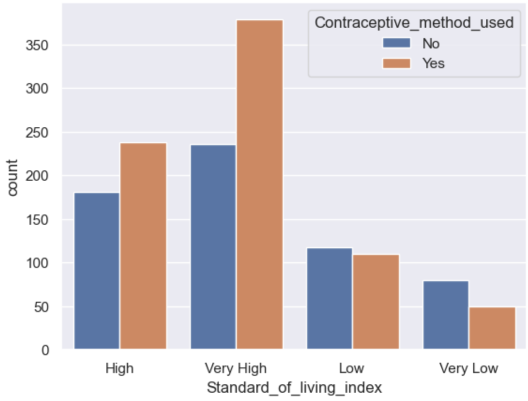
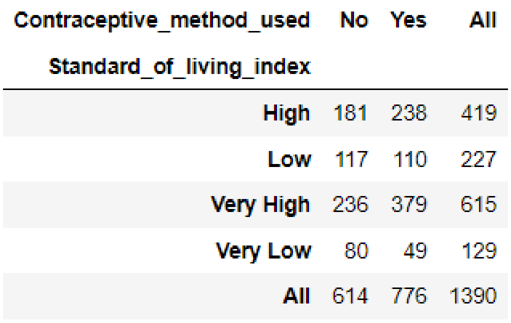


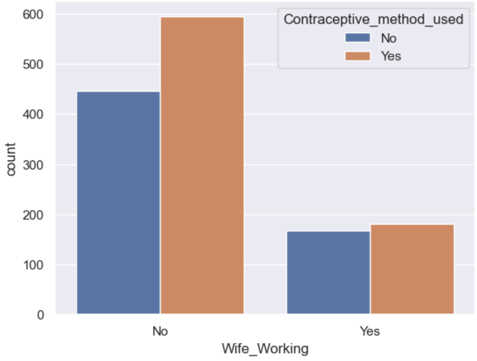
Fig 8. Pair plot of Wife age-No of Children Born-Husband education

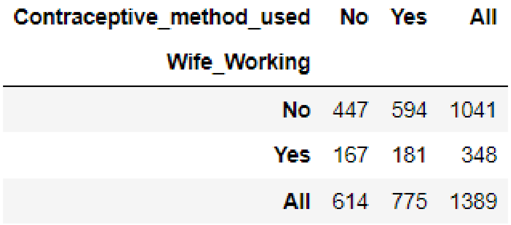




Contraceptive_method_used	No	Yes	All
Standard_of_living_index			
High	181	238	419
Low	117	110	227
Very High	236	379	615
Very Low	80	49	129
All	614	776	1390

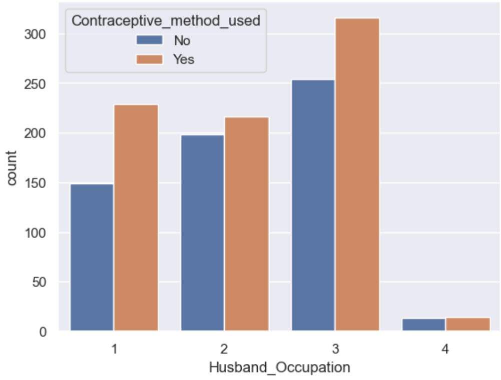
Fig 9. Contraceptive usage based on standard of living index

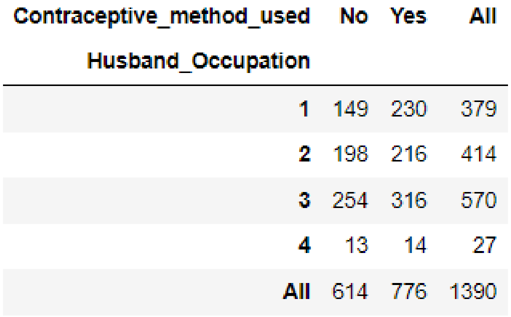




Contraceptive_method_used	No	Yes	All
Wife_Working			
No	447	594	1041
Yes	167	181	348
All	614	775	1389

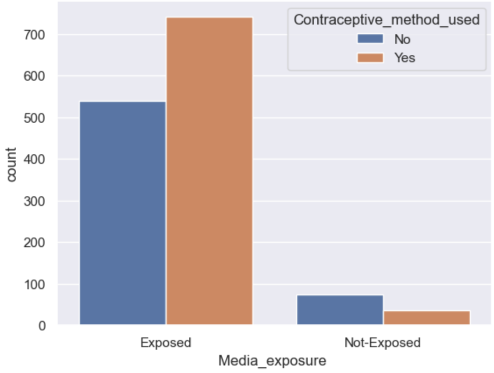
Fig 10. Contraceptive usage-based on wife working

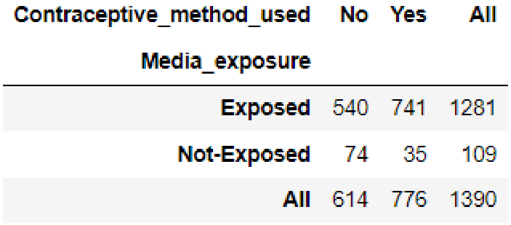




Contraceptive_method_used	No	Yes	All
Husband_Occupation			
1	149	230	379
2	198	216	414
3	254	316	570
4	13	14	27
All	614	776	1390

Fig 11. Contraceptive usage-based on Husband occupation



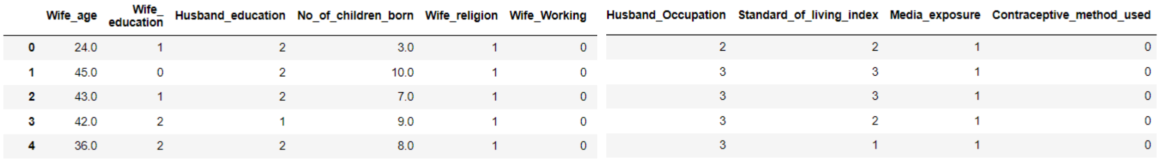


Contraceptive_method_used	No	Yes	All
Media_exposure			
Exposed	540	741	1281
Not-Exposed	74	35	109
All	614	776	1390

Fig 12. Contraceptive usage-based on Media Exposure

Q2]: 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

Encoding the Data for Logistic Regression: String values represented in discrete values as shown in the below tabular column and easy to predict the probabilities of attributes.



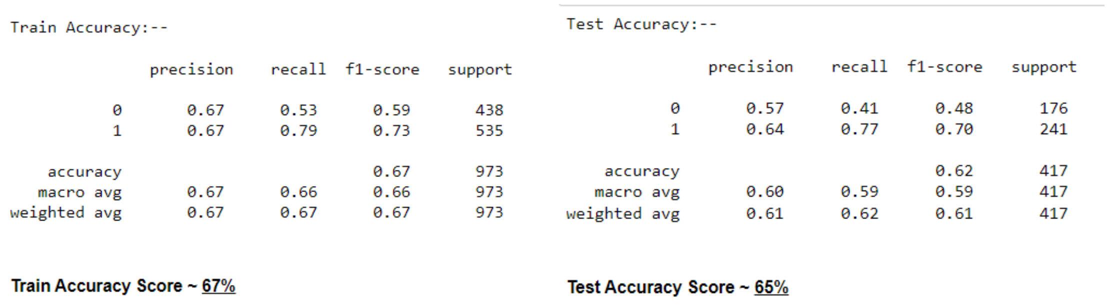
	Wife_age	Wife_ education	Husband_education	No_of_children_born	Wife_religion	Wife_Working	Husband_Occupation	Standard_of_living_index	Media_exposure	Contraceptive_method_used
0	24.0	1	2	3.0	1	0	2	2	1	0
1	45.0	0	2	10.0	1	0	3	3	1	0
2	43.0	1	2	7.0	1	0	3	3	1	0
3	42.0	2	1	9.0	1	0	3	2	1	0
4	36.0	2	2	8.0	1	0	3	1	1	0

Table 2. Encode string to discrete values

Using Sklearn libraries to split the data into train and test and to find the accuracy of the train and test results with ratio of 70:30 respectively. Apply logistic regression to predict the precision and recall scores. Key indicator for classification model

Q3] : Performance Metrics: Check the performance of Predictions on Train and Test sets using Accuracy, Confusion Matrix, Plot ROC curve and get ROC_AUC score

Data Visualization: Train and Test Accuracy scores



```
Train Accuracy:--

              precision    recall  f1-score   support

           0       0.67      0.53      0.59       438
           1       0.67      0.79      0.73       535

    accuracy                           0.67       973
   macro avg       0.67      0.66      0.66       973
weighted avg       0.67      0.67      0.67       973
```

Train Accuracy Score ~ 67%

```
Test Accuracy:--

              precision    recall  f1-score   support

           0       0.57      0.41      0.48       176
           1       0.64      0.77      0.70       241

    accuracy                           0.62       417
   macro avg       0.60      0.59      0.59       417
weighted avg       0.61      0.62      0.61       417
```

Test Accuracy Score ~ 65%

Table 3. Train and test accuracy

Data Visualisation: Confusion Matrix

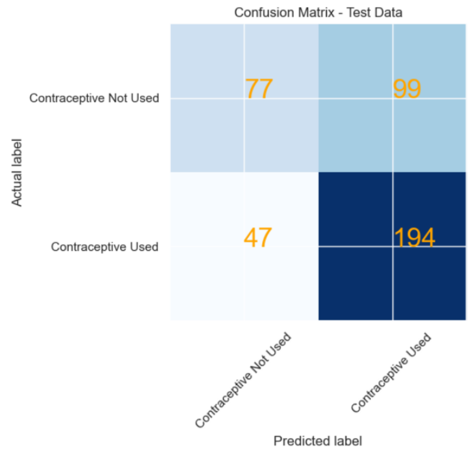


Fig 13. Confusion Matrix

Matrix gives insight on performance of classification model. How well model is predicted for contraceptive used / not used as shown in the above tabular column.

Data Visualisation: ROC - AUC Metrics

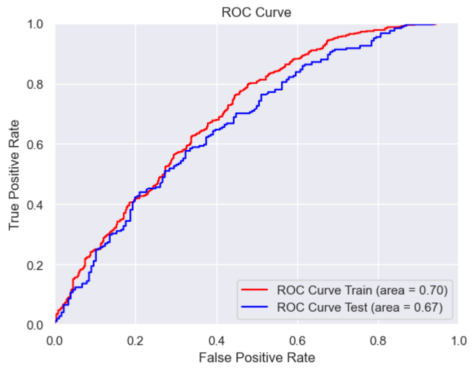


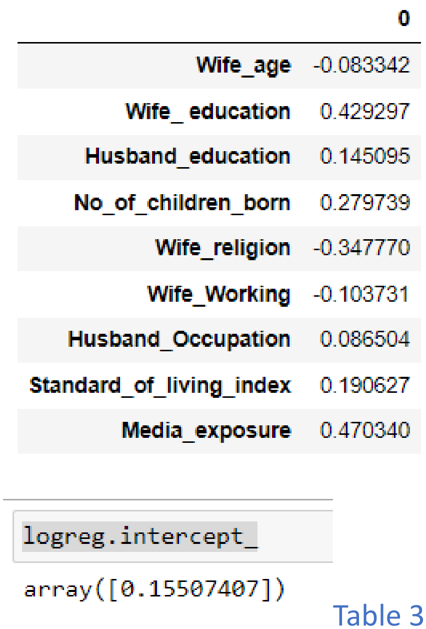
Fig 14. ROC – AUC features

Train ROC curve = 0.70

Test ROC curve = 0.67

Using ROC-AUC method increases 3% of prediction than compare to logistic split method approach.

Coefficient features and intercept after Logistic Regression



	0
Wife_age	-0.083342
Wife_ education	0.429297
Husband_education	0.145095
No_of_children_born	0.279739
Wife_religion	-0.347770
Wife_Working	-0.103731
Husband_Occupation	0.086504
Standard_of_living_index	0.190627
Media_exposure	0.470340

```
logreg.intercept_
```

```
array([0.15507407])
```

Table 3. Co-efficient features

Conclusion & Recommendation

- Wife Age (-0.083342):

  older wives are associated with lower odds of using contraceptives.

- Wife Education (0.429297), Husband Education (0.145095):

  The education level of both the wife and husband positively influences contraceptive use.

- Number of Children Born (0.279739):

  Couples with more children are more likely to use contraceptives.

- Husband Occupation (0.086504):

  The nature of the husband's occupation may influence family planning decisions.

- Standard of Living Index (0.190627):

  Couples with a higher standard of living are more likely to use contraceptives.

- Media Exposure (0.470340):

  Media exposure may play a role in influencing family planning decisions.

THE END