PROJECT - PREDICTIVE MODELLING

Team # 3 Assignment Report

Santhosh , Siddique, Kiran, Karan, Piyush, Vivek

Contents

Problem 2: Logistic Regression, LDA and CART

You are a statistician at the Republic of Indonesia Ministry of Health, and you are provided with a data of 1473 females collected from a Contraceptive Prevalence Survey. The samples are married women who were either not pregnant or do not know if they were at the time of the survey.

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

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.**

Basic Information about the dataset:

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1473 entries, 0 to 1472
Data columns (total 10 columns):
 #   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
dtypes: float64(2), int64(1), object(7)
memory usage: 115.2+ KB
```

- There are 1473 records with 10 columns present in the dataset.
- There is a mixture of numeric and categorical data types.
    - 3 numeric features
    - 7 categorical features.
- There are few rows with the null entries.

Checking for the count of null values:

Let us make Husband occupation as Categorical variable from an integer. Also validate the null checks. We are getting null for Wife age and number of children born.

```
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
```

After imputing the null value with modes, output is as shown below.

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

Checking for duplicate records:

```
In [11]: .data.duplicated().sum()
Out[11]: 87
```

- Two columns contain missing records.
- There are 87 records which are duplicated.

Total 87 records are duplicated, we are dropping those duplicates, as the same information is recorded in other records.

After treating missing and duplicate values, the dataset contains:

- 1386 rows
- 10 columns

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1386 entries, 0 to 1472
Data columns (total 10 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Wife_age                   1386 non-null   float64
 1   Wife_ education            1386 non-null   object
 2   Husband_education          1386 non-null   object
 3   No_of_children_born        1386 non-null   float64
 4   Wife_religion              1386 non-null   object
 5   Wife_Working               1386 non-null   object
 6   Husband_Occupation         1386 non-null   category
 7   Standard_of_living_index   1386 non-null   object
 8   Media_exposure             1386 non-null   object
 9   Contraceptive_method_used  1386 non-null   object
dtypes: category(1), float64(2), object(7)
memory usage: 109.8+ KB
```

Statistical Summary or 5 point summary of the dataset.

	count	unique	top	freq	mean	std	min	25%	50%	75%	max
Wife_age	1386.0	NaN	NaN	NaN	32.227273	8.254237	16.0	25.0	31.0	38.0	49.0
Wife_ education	1386	4	Tertiary	511	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Husband_education	1386	4	Tertiary	822	NaN	NaN	NaN	NaN	NaN	NaN	NaN
No_of_children_born	1386.0	NaN	NaN	NaN	3.277056	2.390657	0.0	1.25	3.0	5.0	16.0
Wife_religion	1386	2	Scientology	1179	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Wife_Working	1386	2	No	1036	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Husband_Occupation	1386.0	4.0	3.0	566.0	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Standard_of_living_index	1386	4	Very High	614	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Media_exposure	1386	2	Exposed	1277	NaN	NaN	NaN	NaN	NaN	NaN	NaN
Contraceptive_method_used	1386	2	Yes	772	NaN	NaN	NaN	NaN	NaN	NaN	NaN

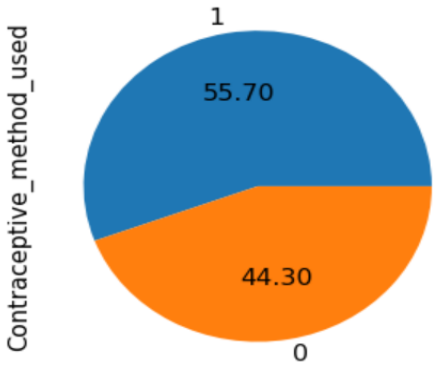
Findings:

1. wife_education, Husband_education and Standard_of_living_index has 4 unique labels.
2. Wife_religion, Wife_working and Media_exposure has 2 unique labels.
3. There is no major difference between Mean and Median.
4. There seems to be outliers in No_of_children_born.

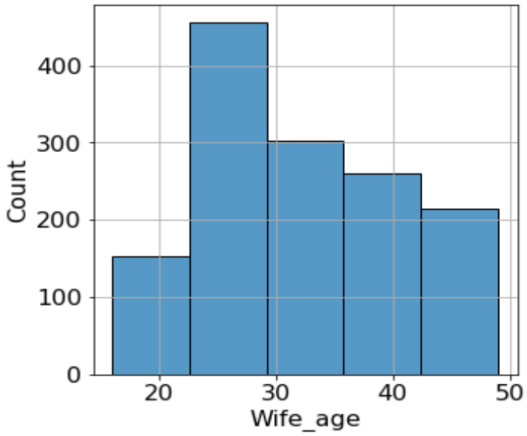
Univariate Analysis

1. **Contraceptive method used.**

The dataset is with respect to target/dependent fetaures seems to be balanced (There is no much difference).



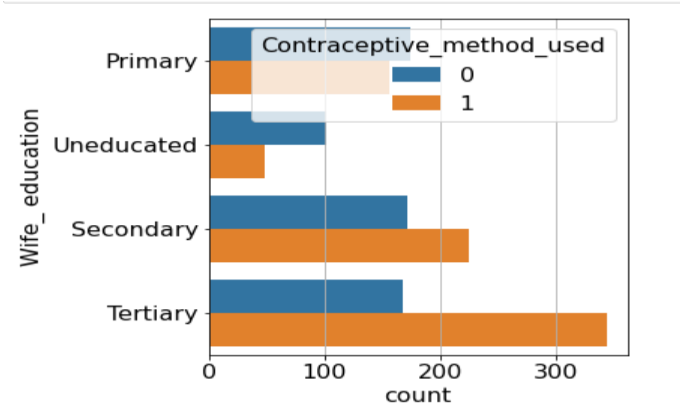
2. **Wife_Age distribution**



When we describe Wife age feature and here are few observations

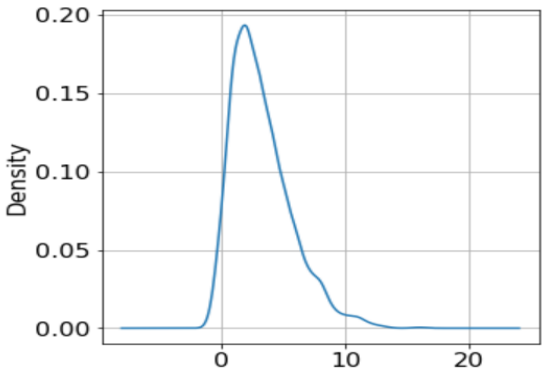
- A dataset consists of Women's Age ranges from 16 to 49.
- Average Women's Age is 32.5
- Age is not normally distributed; it is bit right skewed with fatter tail

3. Count Plot for Wife Education



- 90% of the females are having at least a minimum education, only 10% are illiterates.

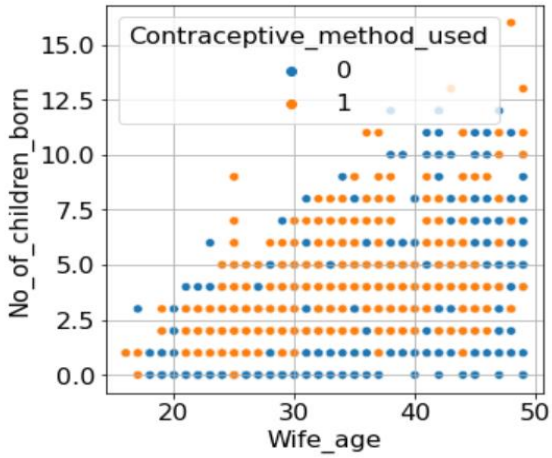
4. Density plot for number of children born



1. No_of_Childeren_Born is right skewed.
    - Minimum is 0
    - Maximum is 16
    - Average is 3
2. There seems to be Outliers in this feature.

Bivariate Analysis

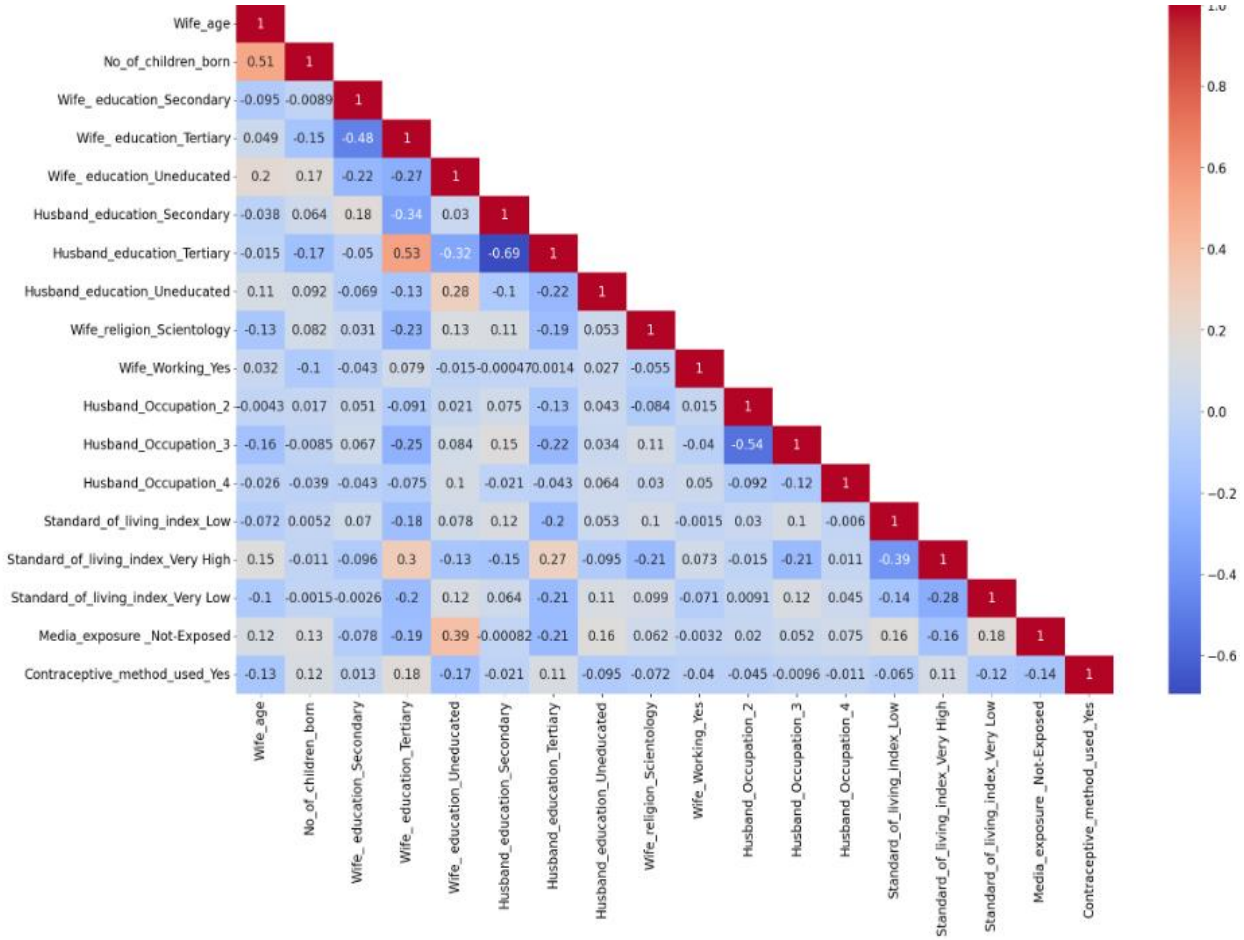
- **No of Children Born vs Wife_Age**



In the above scatterplot we can see there is a positive co-relation between Wife_age" and "No_of_children_born", As the Age increases number of children is also increasing.

Multivariate Analysis

Heat Map



In the above heatmap, we see the relationship between the features.

- No of children born and wife age is positively co-related (51)
- Husband education tertiary and wife education tertiary is also positively co-related (53)
- Husband education tertiary and husband education secondary are negatively co-related (-69)
- Husband occupation 2 and husband occupation 3 is negatively co-related (-54)

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.**

Encoding the data using get_dummies() function.

Before encoding: Initially we are having 10 columns.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1386 entries, 0 to 1472
Data columns (total 10 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   Wife_age                   1386 non-null   float64
 1   Wife_ education            1386 non-null   object
 2   Husband_education          1386 non-null   object
 3   No_of_children_born        1386 non-null   float64
 4   Wife_religion              1386 non-null   object
 5   Wife_Working               1386 non-null   object
 6   Husband_Occupation         1386 non-null   category
 7   Standard_of_living_index   1386 non-null   object
 8   Media_exposure             1386 non-null   object
 9   Contraceptive_method_used  1386 non-null   object
dtypes: category(1), float64(2), object(7)
memory usage: 109.8+ KB
```

After encoding, there are 18 columns.

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 1386 entries, 0 to 1472
Data columns (total 18 columns):
 #   Column                              Non-Null Count  Dtype
---  ------                              --------------  -----
 0   Wife_age                            1386 non-null   float64
 1   No_of_children_born                 1386 non-null   float64
 2   Wife_ education_Secondary           1386 non-null   uint8
 3   Wife_ education_Tertiary            1386 non-null   uint8
 4   Wife_ education_Uneducated          1386 non-null   uint8
 5   Husband_education_Secondary         1386 non-null   uint8
 6   Husband_education_Tertiary          1386 non-null   uint8
 7   Husband_education_Uneducated        1386 non-null   uint8
 8   Wife_religion_Scientology           1386 non-null   uint8
 9   Wife_Working_Yes                    1386 non-null   uint8
 10  Husband_Occupation_2                1386 non-null   uint8
 11  Husband_Occupation_3                1386 non-null   uint8
 12  Husband_Occupation_4                1386 non-null   uint8
 13  Standard_of_living_index_Low        1386 non-null   uint8
 14  Standard_of_living_index_Very High  1386 non-null   uint8
 15  Standard_of_living_index_Very Low   1386 non-null   uint8
 16  Media_exposure _Not-Exposed         1386 non-null   uint8
 17  Contraceptive_method_used_Yes       1386 non-null   uint8
dtypes: float64(2), uint8(16)
memory usage: 54.1 KB
```

Logistics Regression.

- In descriptive approach we did for 10 different models.
- Here we are using backward approach.
- VIF Values for all the columns seems to be not significant (There is no much multicollinearity), so we can drop features based on p values.
- P-Value for some features is very high. Those features are not statistically important. We can drop the feature with the higher P value.

Model 1 : With all the features.

- VIF values for all the features are less than 5, there seems to be no multi-colinearity.
- Husband_education_Tertiary is having highest p value 88%, SO we are dropping the feat ure in the next model.

Model_1 Performance is 0.1040

Model 2 : We are dropping Husband_education_Tertiary.

- Husband_occupation_3 is having highest p value 70%, SO we are dropping the feature in the next model.

Model_2 Performance is 0.1067

Model 3 : We are dropping Husband_occupation_3.

- Husband_education_Uneducated is having highest p value 50%, SO we are dropping the f eature in the next model.

Model_3 Performance is 0.1077

Model 4: We are dropping Husband_education_Uneducated

- Wife_Working_Yes is having highest p value 50%, SO we are dropping the feature in the next model.

Model_4 Performance is 0.1085

Model 5: We are dropping Wife_Working_Yes

- **Standard_of_living_index_Low** is having highest p value 30%, SO we are dropping the feature in the next model.

Model_5 Performance is 0.1091

Model 6: We are dropping Standard_of_living_index_Low

- Husband_Occupation_4 is having highest p value 30%, SO we are dropping the feature in the next model.

Model_6 Performance is 0.1096

Model 7: We are dropping Husband_Occupation_4

- Husband_education_Secondary is having highest p value 21%, SO we are dropping the fe ature in the next model.

Model_7 Performance is 0.1100

Model 8: We are dropping Husband_education_Secondary is

- **Wife_education_Uneducated** is having highest p value 15%, SO we are dropping the fea ture in the next model.

Model_8 Performance is 0.1103

Model 9: We are dropping Wife_education_Uneducated

- Husband_Occupation_2 is having highest p value 30%, SO we are dropping the feature in the next model.

Model_9 Performance is 0.1132

Model 10: We are dropping Husband_Occupation_2

- In this model all the features are statistically significant(Which is less than P value 0.05), so we are not dropping any features further.

Model_10 Performance is 0.1098

Findings from descriptive approach and performance of each Logistic Regression model:

	model_name	model_perf
1	Model_1	0.105719
2	Model_2	0.106759
3	Model_3	0.107735
4	Model_4	0.108549
5	model_5	0.109168
6	model_6	0.109653
7	model_7	0.110085
8	model_8	0.110326
9	model_9	0.110277
10	model_10	0.109893

- Dataset has no multicollinearity between the features.(VIF criteria <5)
- Model 8, Model 9 & Model 10 seems to be the best models with the statistically significant features.
- Model_8 is having higher Pseudo_R_Square, but some features are having p-value more than 0.05.
- Model_10 is having features which are statistically significant, but slight decrease in the Pseudo_R_Sqaure.

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.**

Performance of Predictions on Train and Test sets using Accuracy:

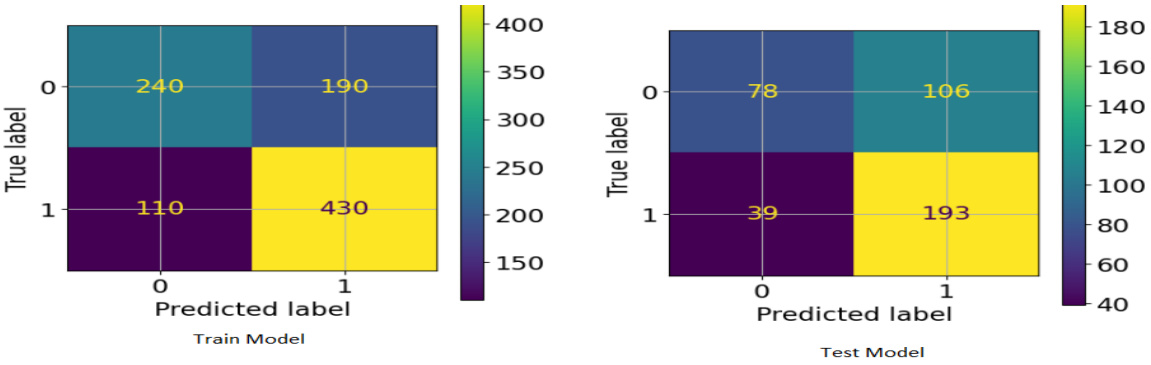
	Classification Technique	Train accuracy	Test accuracy
0	Logistic Regression 1	0.690722	0.651442
1	Logistic Regression 8	0.685567	0.661058
2	Logistic Regression 9	0.678351	0.665865
3	Logistic Regression 10	0.675258	0.661058

In Predictive Approach we are doing for 4 models.

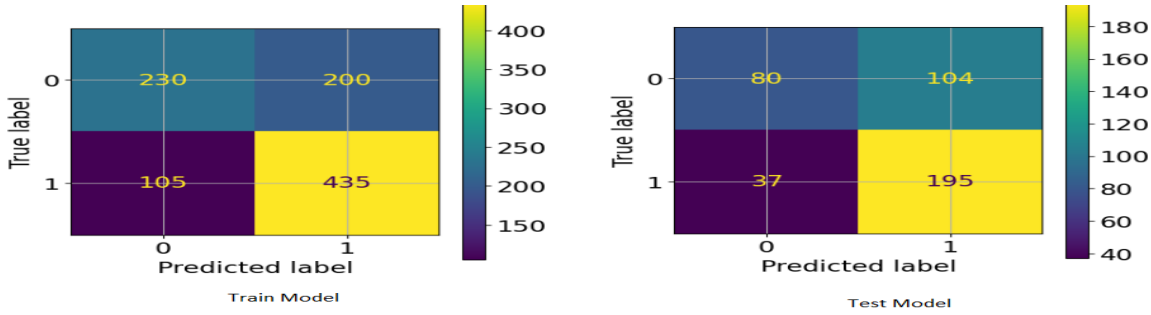
- We are doing Logistic Regression for Model_1, Model_8 and Model_10.
- Model _1 is performing better in train and not performing well in test model. And also, we are adding all the features, here some features are not statistically significant.
- Model_8 is best in terms of descriptive approach and also it is with 2 not significant feature.
- Model_10 is best in terms of Predictive approach, and in this model it contains only significant features. It is having the higher accuracy.

Confusion Matrix:

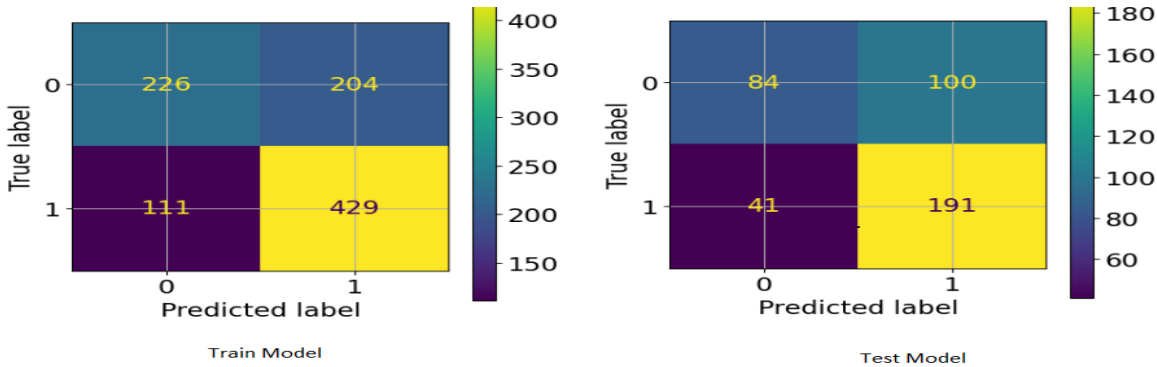
Model_1



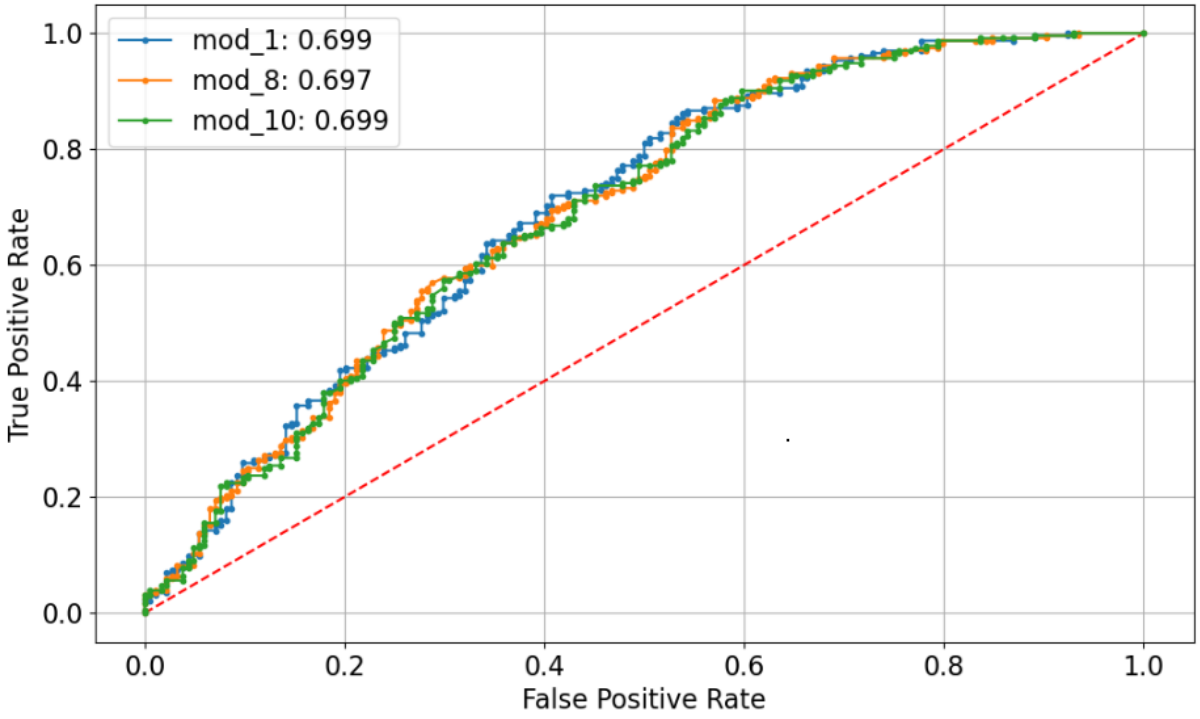
Model_8



Model_10



Plot ROC curve and get ROC_AUC score for each model:



From all the 3 models, model_1 and model_10 is having same AUC_ROC Score is same, but in model_1 we are considering all significant and insignificant features. In Model_10 we are only focused on significant features.

Final Model: Compare Both the models and write inference which model is best/optimized.

- In Logistic Regression and Linear Discriminant Analysis, Model_10 is giving similar accuracy. So, we are choosing Model_10 is the best model.
- In the Decision Tree, we are getting best accuracy when compared to all the models of Logistic Regression and Linear Discriminant Analysis.

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

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

Performance of each model with respect to train and test accuracy.

	Classification Technique	Train accuracy	Test accuracy
0	Logistic Regression 1	0.690722	0.651442
1	Logistic Regression 8	0.685567	0.661058
2	Logistic Regression 9	0.678351	0.665865
3	Logistic Regression 10	0.675258	0.661058
4	LDA 1	0.694845	0.651442
5	LDA 8	0.687629	0.665865
6	LDA 9	0.683505	0.665865
7	LDA 10	0.674227	0.661058
8	DT 1	0.722680	0.677885
9	DT 8	0.721649	0.677885
10	DT 9	0.721649	0.677885
11	DT 10	0.721649	0.677885

Inferences:

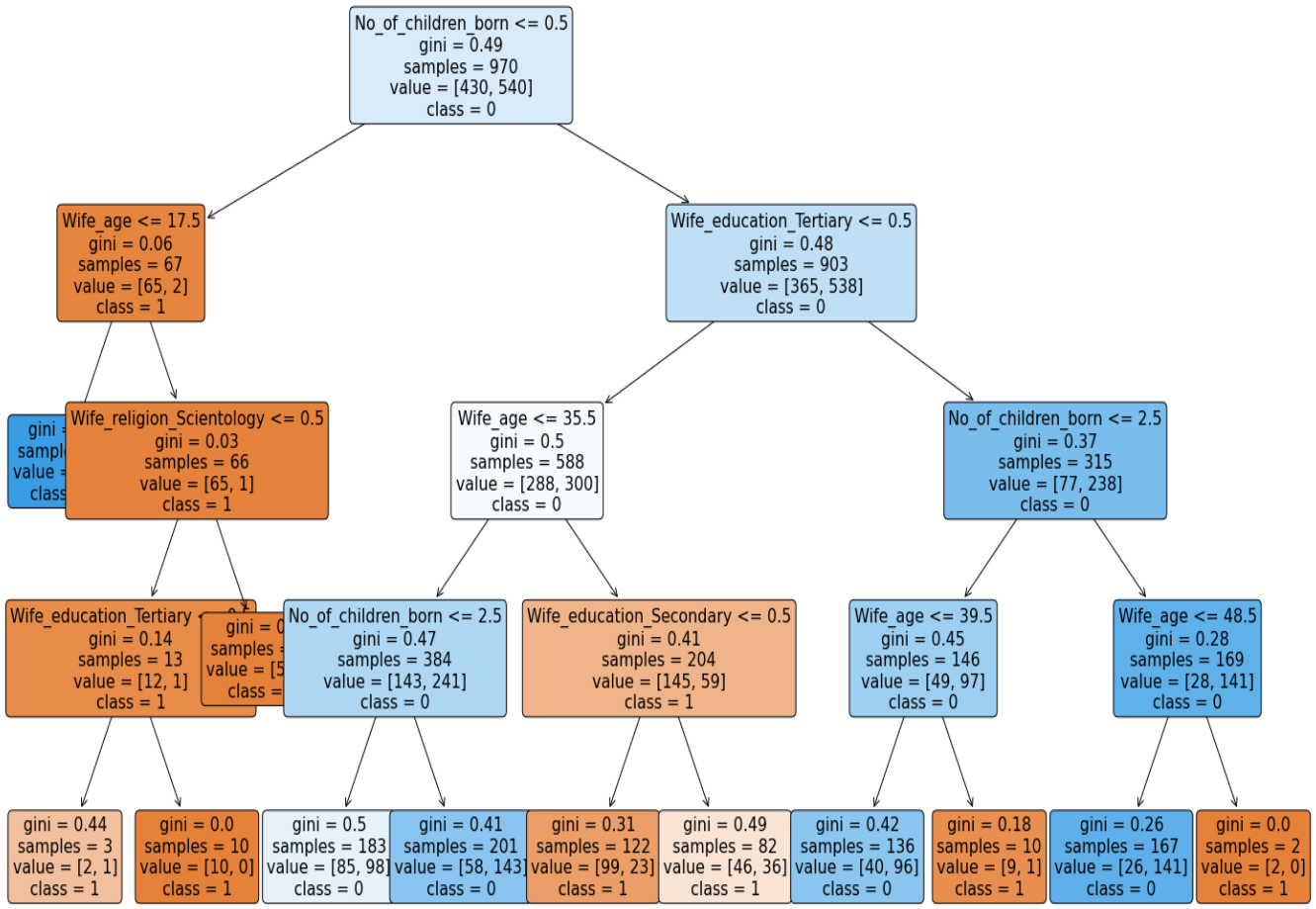
Logistic Regression:

- Model 1 with all the features in 3 different algorithm is performing better, but it contains non-significant features which are having p_value higher than 0.05.
- Model 8 is having some fewer features with statistically significant, but with 2 non-significant features.
- Performance wise in Logistic Regression, Model_8 is having higher accuracy.
- The accuracy for model_9 is dropping slightly, but with only 1 non-significant feature.
- Model_10 is having all the features which are statistically significant with p value lesser than 0.05.

Linear Discriminant Analysis:

- LDA models are performing similar to logistic regression models.

Decision Tree:



- In Decision tree also Model 10 is performing better when compared to other decision tree models

Recommendations:

Decision Tree is performing well in train and test model when compared to logistic regression and Linear Discriminant Analysis.

Steps performed

1. Read the dataset with Python Libraries

2. Checking of null and we are dropping it, as the count of null values are very less. As data points in some features are not normally distributed

3. Checked the duplicate values and there are duplicate values, we are dropping those duplicated values to avoid inconsistency in the dataset.

4. All features are Categorical features and the only feature Wife_Age is numerical.

5. The target feature ‘Contraceptive Method Used’ is categorial, hence model can be built with the Logistic Regression, Linear Discriminant Analysis and Decision Tree.

6. EDA process of the data like checking of outliers in the data set.

7. No of children born feature has outliers and we are capping it to improve the performance and build effective model.

8. First check the multicollinearity among the independent variables with the help of VIF (Variance Inflation Factor'), drop the feature which have highest VIF value and VIF value which have more than five.

9. Drop the features which have P value more than 0.05 which means the feature which are not significant to build the model.

10. Finally 10 different models were built and the model which performs best on test side like highest performance value and accuracy of the model will be considered from the business perspective.

11. All models will be submitted to the client and client will take decision for adopting of the model which best suited for his business out of the 10 models built.

END OF REPORT