|  |
| --- |
| **Gender Prediction+ Breast Cancer+ Delhi Accident** |
| Classification |
| Rabi |

**Gender Prediction**

Step-1:

Dataset overview, identifying explanatory and response variables

Step-2

Finding the features with maximum missing values, separating the categorical features

Step-3

EDA is done, label encoding is performed.

Step-4

Train and test, Model fitting- Logistic, Decision Tree, XGboost

Step-5

Test Accuracy is 35% using the Logistic Regression.

Step-6

Threshold probabilities=0.38

Step-7

Hyper parameter tuning

C=1.623776739188721

Test Accuracy is =35%

**Breast Cancer**

Step-1

Dataset Overview and separate explanatory and response variables

Step-2

Check for balance and imbalance of the dataset, EDA for response variable, perform the Up or down sampling in order to make the dataset balanced.

Step-3

Finding the missing values, impute the missing values based on the respective distribution of each of the features or remove the row based on the percentage of missing values.

Step-4

Counts the different features of the dataset, Find the correlation among the numerical features among themselves.

Step-5

Find correlation and multicollinearity , label encoding, splitting dataset,

Step-6

* The threshold probability should be 0.50.
* C=78.47599703514607, penalty='l1', solver='liblinear'

Step-7

* Accuracy Score on Test set:- 0.9532163742690059
* F1 Score:- 0.9354838709677421
* Average Precision Score:- 0.8830317863324109
* Log Loss:- 1.6158819201245438
* Precision Score:- 0.8923076923076924
* Recall Score:- 0.9830508474576272
* ROC-AUC Score:- 0.9602754237288136

Delhi Accident

Step-1

Read the dataset, Find response variable.

Step-2

Label encoding and hot encoding can be used

Step- 3

Feed into the model –Logistic Regression

Step-4

Hyper parameters tuning: NO effect is observed

Step-5

Accuracy is same as it was earlier

But still

* The threshold probability should be 0.75.
* C=0.615848211066026