Tool and Techniques Name

1) Python Libraries

- Numpy
- Pandas
- Matplotlib
- Seaborn
- Sklearn.metrics
- Sklearn.preprocessing

2) Dataset Link

- https://www.kaggle.com/datasets/bidemiayinde/thyroid-sickness-determination

3) Preprocessing Techniques

- Label encoded the categorical values
- Replaced the question marks with np.nan values.
- Filled these nan values with mean of the particular column.
- Applied Random Undersampler for undersampling.

4) Feature Extraction/ Selection Technique

- Applied CHI2 technique for Feature Selection.
- Since, all these features were not present in our base paper that's why I have dropped them and left with features like sex, goitre, tumor, pregnant, etc.

6) Classification Technique

- We have 2 classes in our base paper which are whether the person is suffering from the thyroid disease or not.
- Classes are Yes or No

7) Data Split Ratio

- We splitted the data into training and testing in the ratio of 70:30.

8) Base Model

- K Neighbors Classifier
- Random Forest Classifier
- Ann Classifier

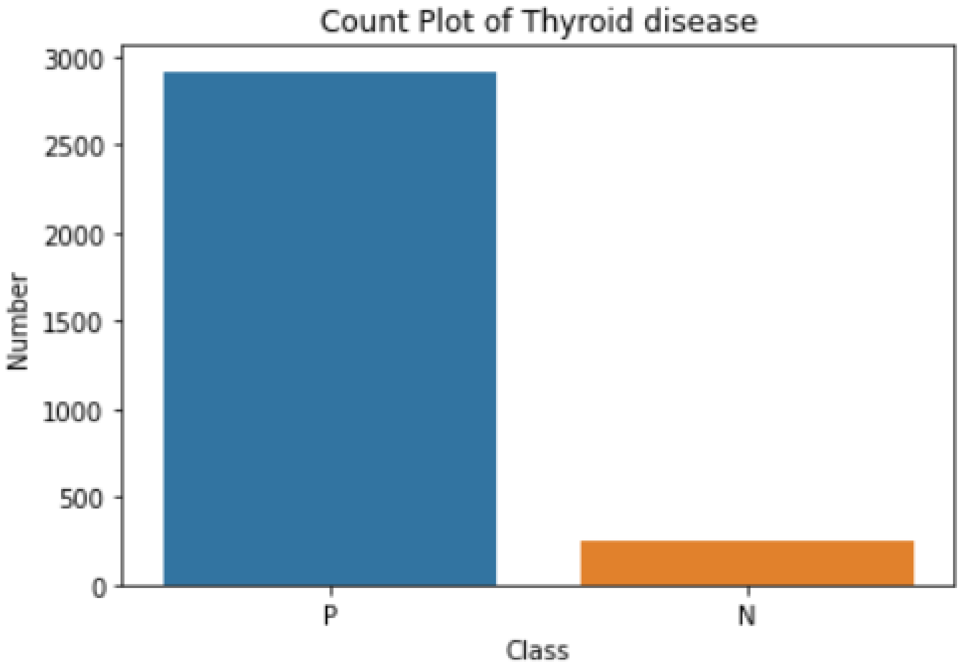
if any another tool and technique used so pls include it and remove it to above

points

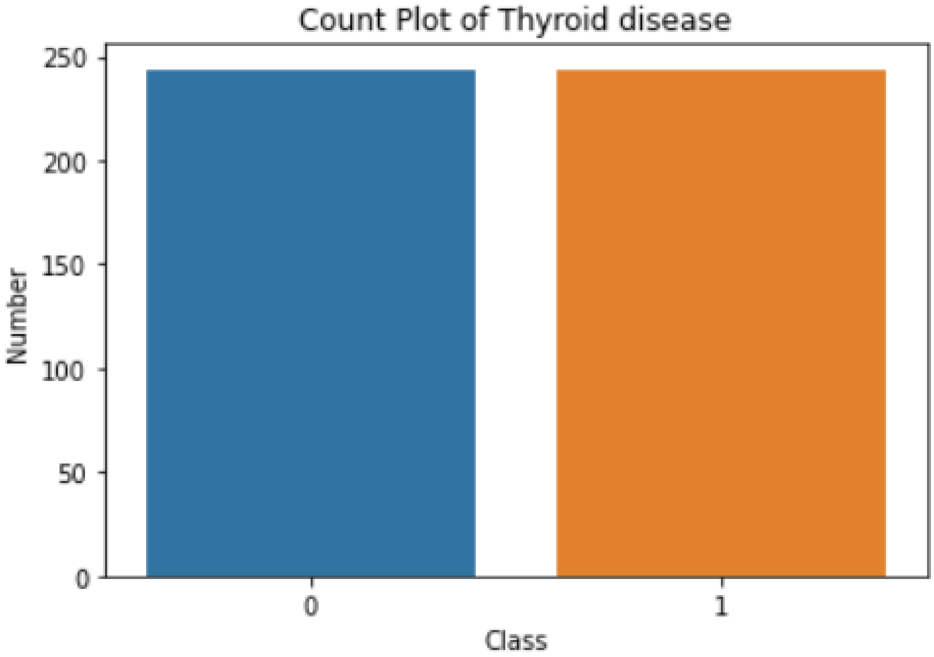
Screenshots of base and proposed results

1) Dataset visualization screenshots

- Count plot of thyroid disease before undersampler

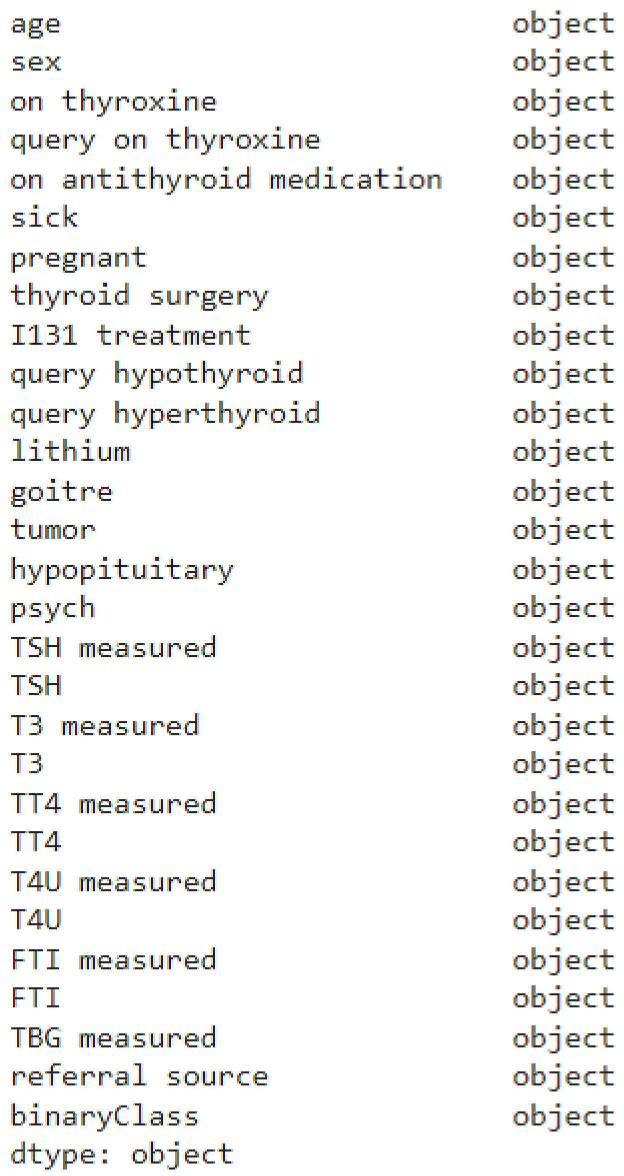


- Count plot after Undersampler



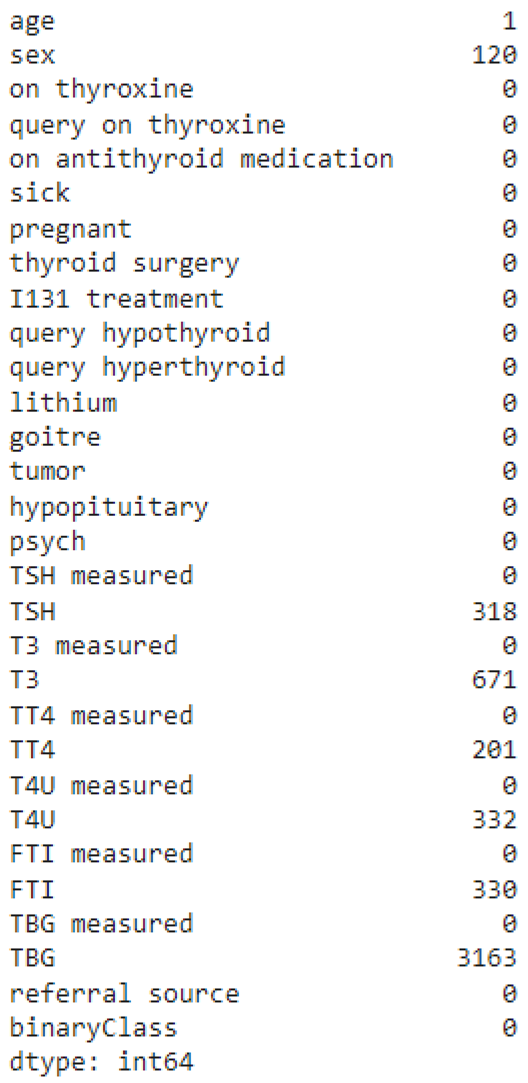
2) Preprocessing results

- Datatypes of features



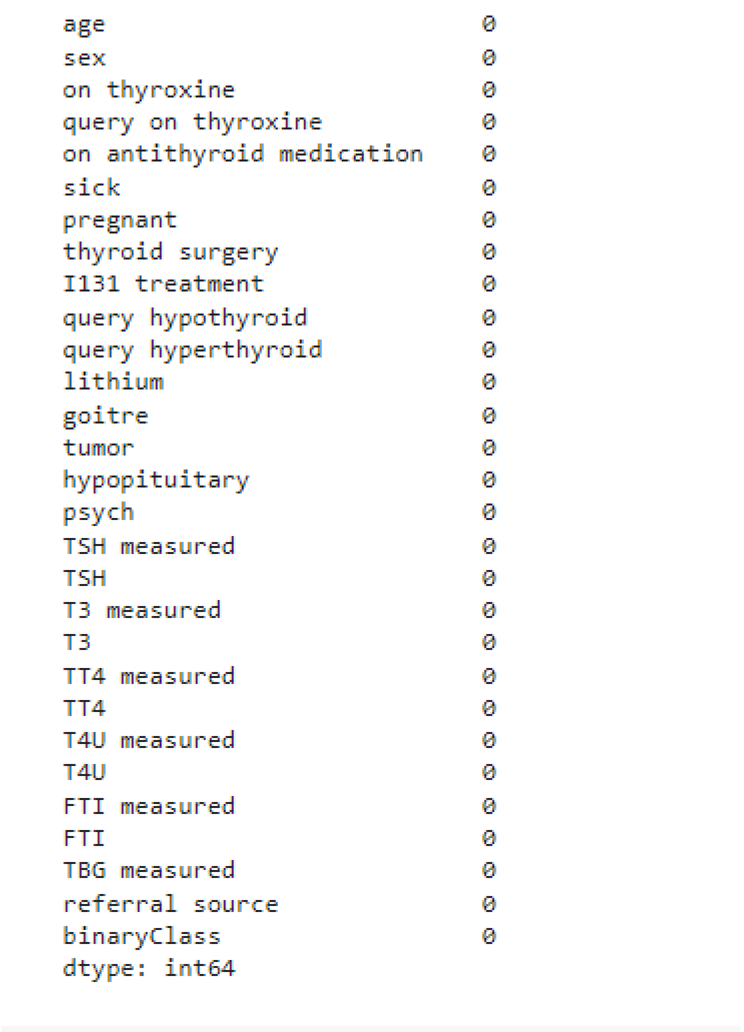
```
age                          object
sex                          object
on thyroxine                 object
query on thyroxine           object
on antithyroid medication    object
sick                         object
pregnant                     object
thyroid surgery              object
I131 treatment               object
query hypothyroid            object
query hyperthyroid           object
lithium                      object
goitre                       object
tumor                        object
hypopituitary                object
psych                        object
TSH measured                 object
TSH                          object
T3 measured                  object
T3                           object
TT4 measured                 object
TT4                          object
T4U measured                 object
T4U                          object
FTI measured                 object
FTI                          object
TBG measured                 object
referral source              object
binaryClass                  object
dtype: object
```

- Checking the null values

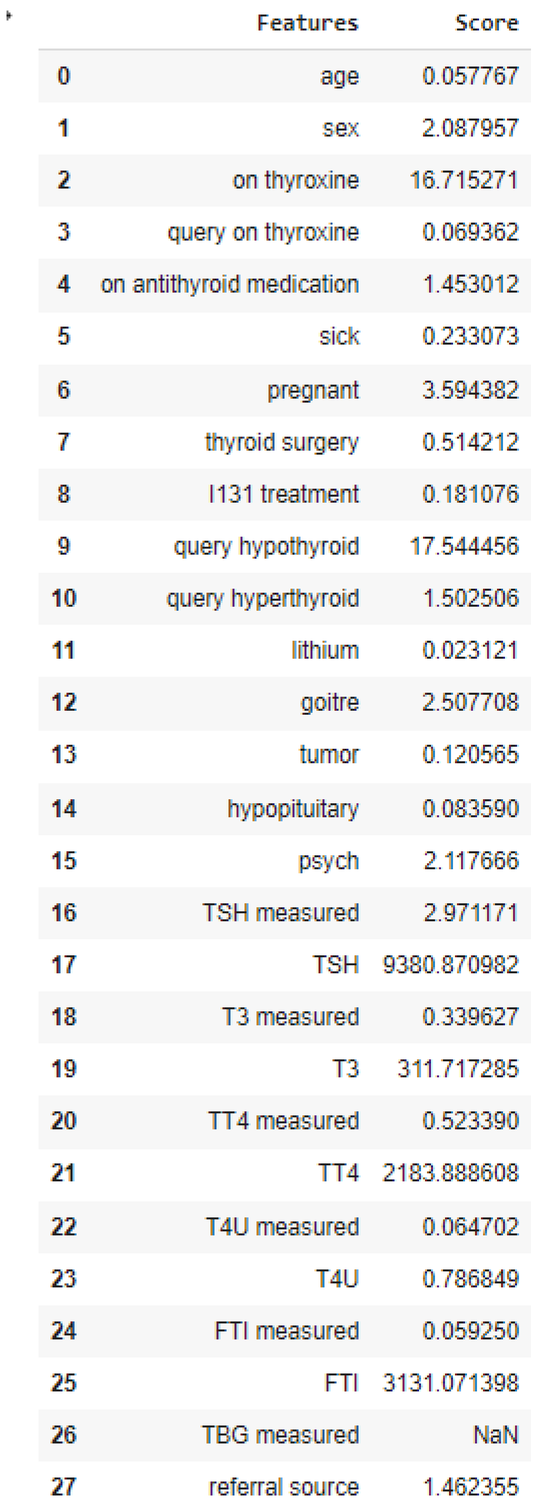


```
age                               1
sex                             120
on thyroxine                      0
query on thyroxine                0
on antithyroid medication         0
sick                              0
pregnant                          0
thyroid surgery                   0
I131 treatment                    0
query hypothyroid                 0
query hyperthyroid                0
lithium                           0
goitre                            0
tumor                             0
hypopituitary                     0
psych                             0
TSH measured                      0
TSH                             318
T3 measured                       0
T3                              671
TT4 measured                      0
TT4                             201
T4U measured                      0
T4U                             332
FTI measured                      0
FTI                             330
TBG measured                      0
TBG                            3163
referral source                   0
binaryClass                       0
dtype: int64
```

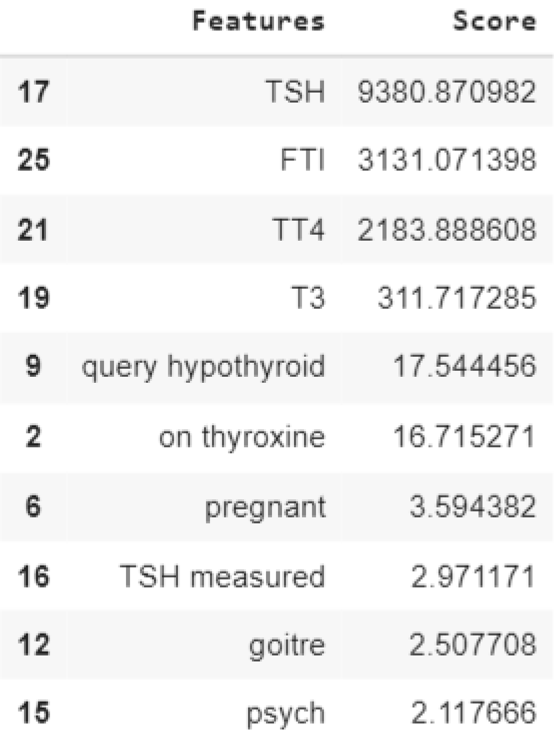
- **After Removing the null values from the dataset**



```
age                          0
sex                          0
on thyroxine                 0
query on thyroxine           0
on antithyroid medication    0
sick                         0
pregnant                     0
thyroid surgery              0
I131 treatment               0
query hypothyroid            0
query hyperthyroid           0
lithium                      0
goitre                       0
tumor                        0
hypopituitary                0
psych                        0
TSH measured                 0
TSH                          0
T3 measured                  0
T3                           0
TT4 measured                 0
TT4                          0
T4U measured                 0
T4U                          0
FTI measured                 0
FTI                          0
TBG measured                 0
referral source              0
binaryClass                  0
dtype: int64
```



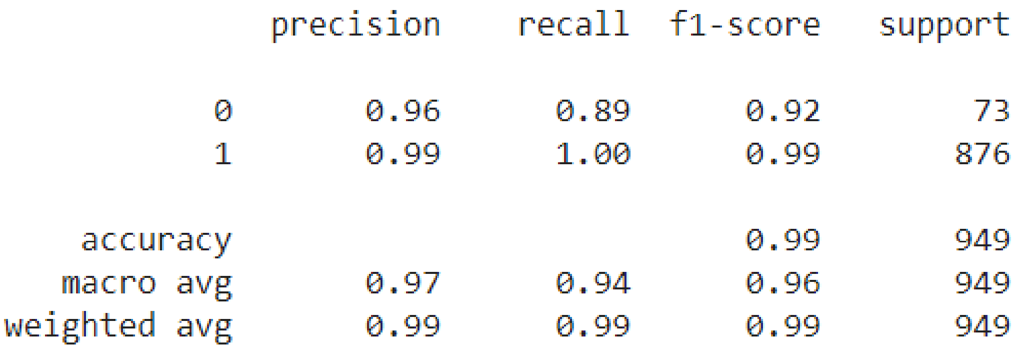
	Features	Score
0	age	0.057767
1	sex	2.087957
2	on thyroxine	16.715271
3	query on thyroxine	0.069362
4	on antithyroid medication	1.453012
5	sick	0.233073
6	pregnant	3.594382
7	thyroid surgery	0.514212
8	I131 treatment	0.181076
9	query hypothyroid	17.544456
10	query hyperthyroid	1.502506
11	lithium	0.023121
12	goitre	2.507708
13	tumor	0.120565
14	hypopituitary	0.083590
15	psych	2.117666
16	TSH measured	2.971171
17	TSH	9380.870982
18	T3 measured	0.339627
19	T3	311.717285
20	TT4 measured	0.523390
21	TT4	2183.888608
22	T4U measured	0.064702
23	T4U	0.786849
24	FTI measured	0.059250
25	FTI	3131.071398
26	TBG measured	NaN
27	referral source	1.462355



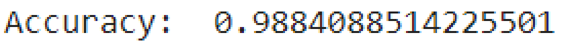
	Features	Score
17	TSH	9380.870982
25	FTI	3131.071398
21	TT4	2183.888608
19	T3	311.717285
9	query hypothyroid	17.544456
2	on thyroxine	16.715271
6	pregnant	3.594382
16	TSH measured	2.971171
12	goitre	2.507708
15	psych	2.117666

5) Base model results---like confusion matrix, ROC curve, classification report etc.

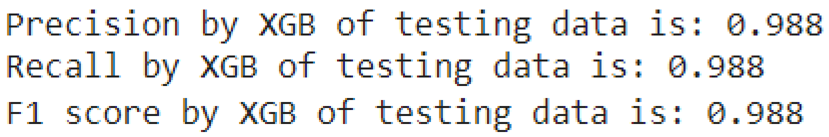
XGB Classifier



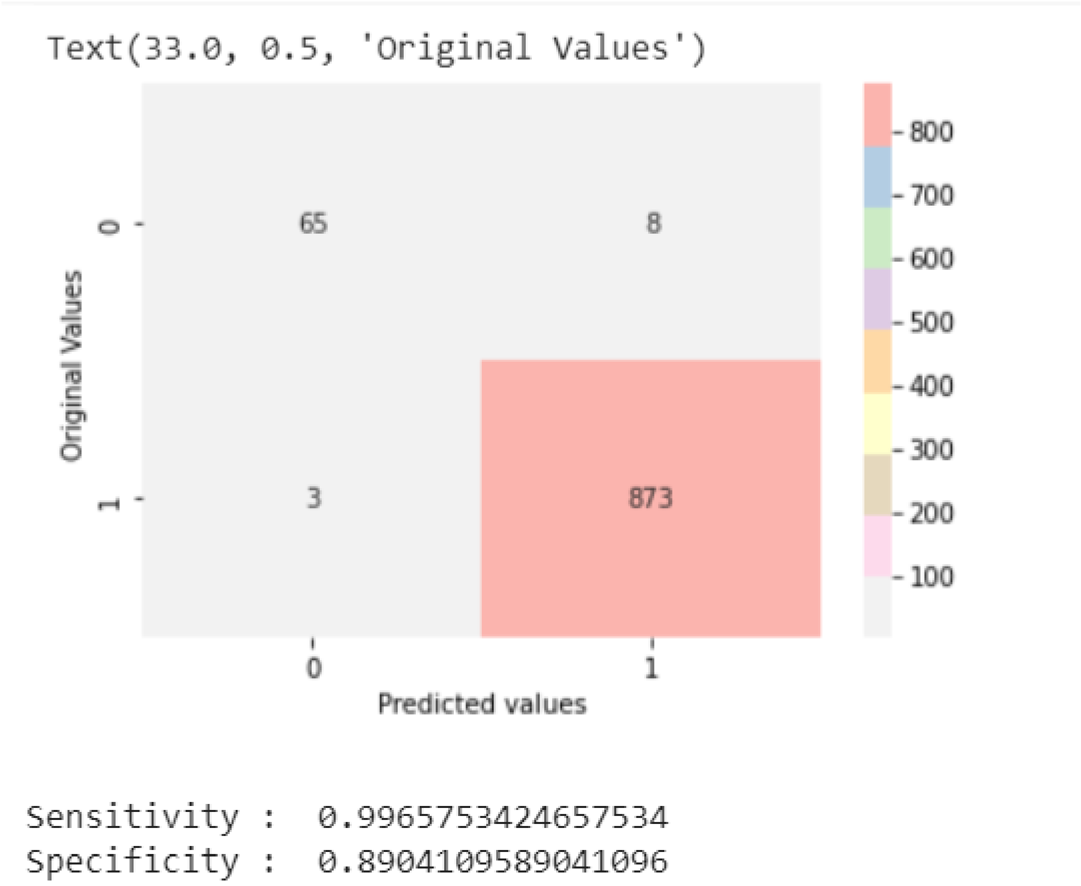
```
              precision    recall  f1-score   support

           0       0.96      0.89      0.92        73
           1       0.99      1.00      0.99       876

    accuracy                           0.99       949
   macro avg       0.97      0.94      0.96       949
weighted avg       0.99      0.99      0.99       949
```



```
Accuracy:  0.9884088514225501
```



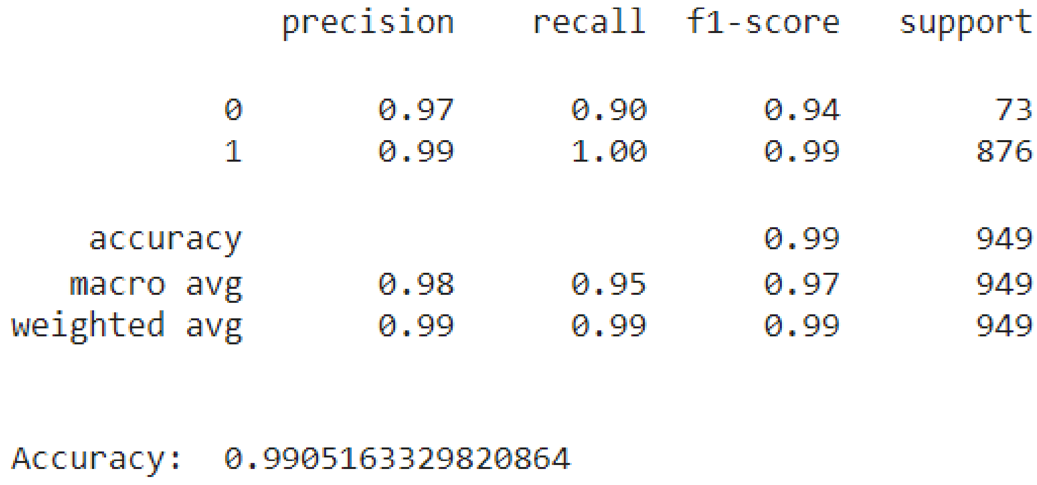
```
Precision by XGB of testing data is: 0.988
Recall by XGB of testing data is: 0.988
F1 score by XGB of testing data is: 0.988
```



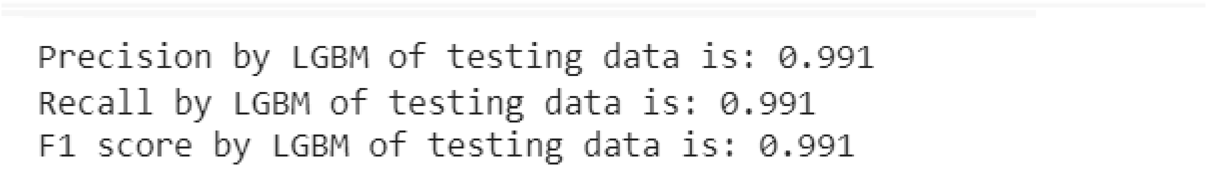
```
Text(33.0, 0.5, 'Original Values')
```

```
Sensitivity :  0.9965753424657534
Specificity :  0.8904109589041096
```

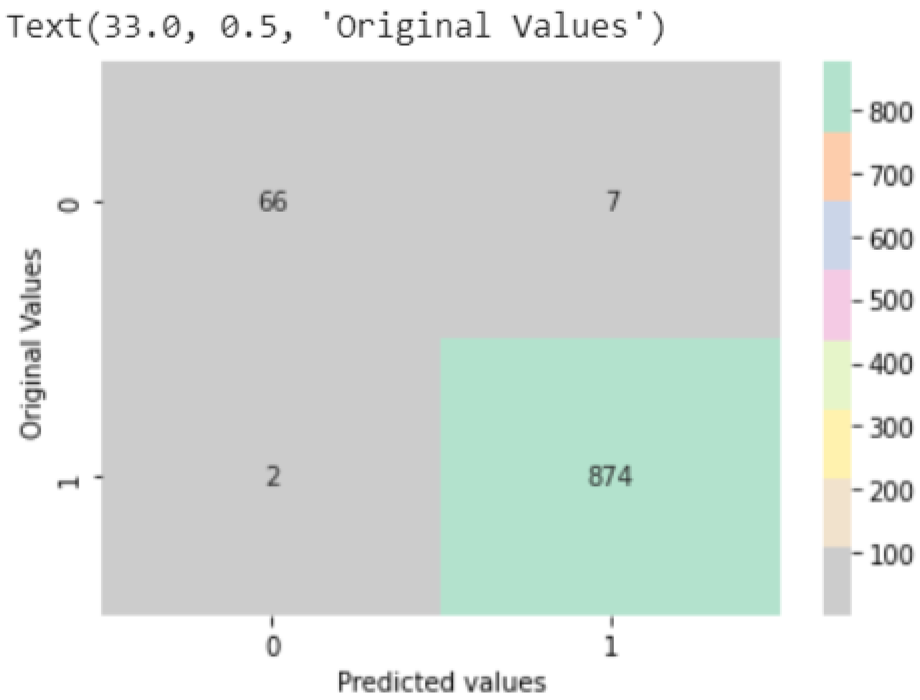
LGBM Classifier



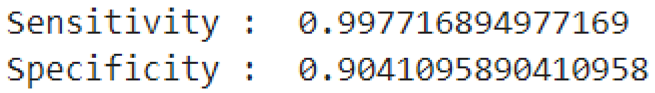
```
              precision    recall  f1-score   support

           0       0.97      0.90      0.94        73
           1       0.99      1.00      0.99       876

    accuracy                           0.99       949
   macro avg       0.98      0.95      0.97       949
weighted avg       0.99      0.99      0.99       949


Accuracy:  0.9905163329820864
```



```
Precision by LGBM of testing data is: 0.991
Recall by LGBM of testing data is: 0.991
F1 score by LGBM of testing data is: 0.991
```

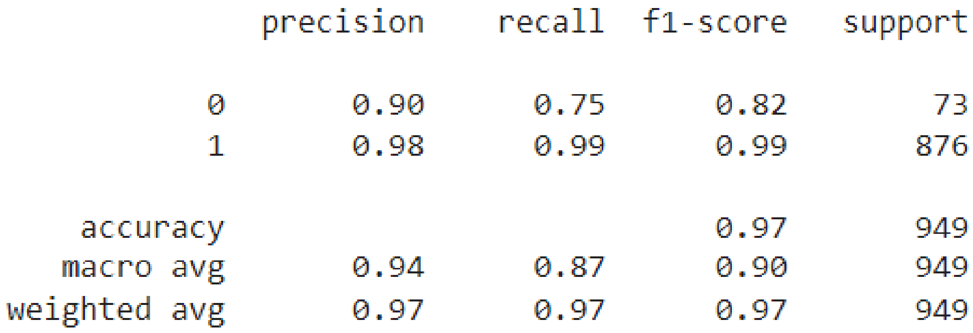


```
Text(33.0, 0.5, 'Original Values')
```

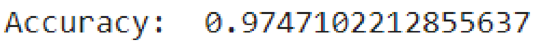


```
Sensitivity :  0.997716894977169
Specificity :  0.9041095890410958
```

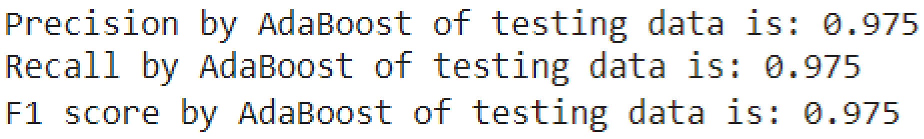
AdaBoost Classifier



```
              precision    recall  f1-score   support

           0       0.90      0.75      0.82        73
           1       0.98      0.99      0.99       876

    accuracy                           0.97       949
   macro avg       0.94      0.87      0.90       949
weighted avg       0.97      0.97      0.97       949
```

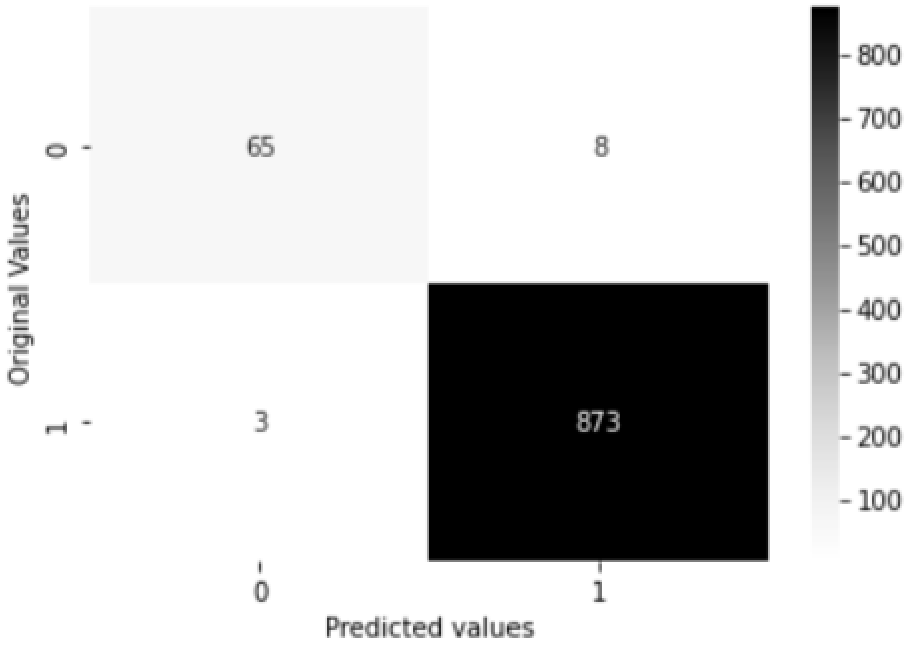


```
Accuracy:  0.9747102212855637
```



```
Precision by AdaBoost of testing data is: 0.975
Recall by AdaBoost of testing data is: 0.975
F1 score by AdaBoost of testing data is: 0.975
```

```
Text(33.0, 0.5, 'Original Values')
```

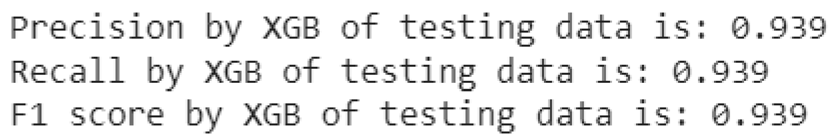


```
Sensitivity :  0.9965753424657534
Specificity :  0.8904109589041096
```

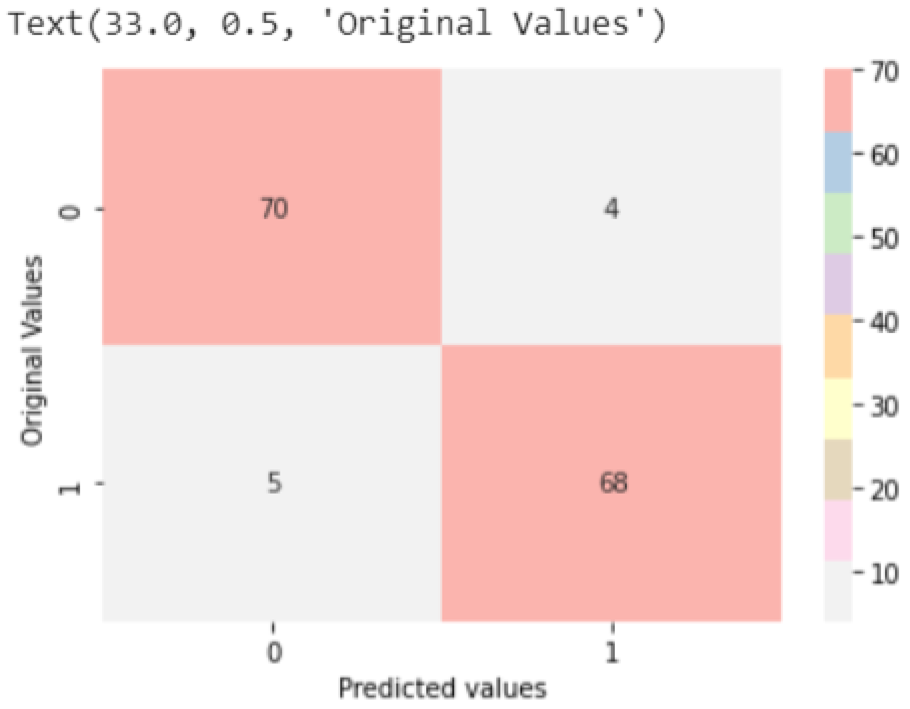
After UnderSampling

XGB Classifier

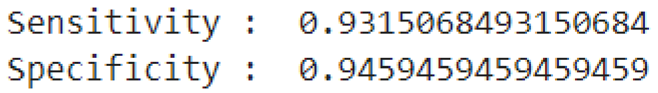
```
              precision    recall  f1-score   support

           0       0.93      0.95      0.94        74
           1       0.94      0.93      0.94        73

    accuracy                           0.94       147
   macro avg       0.94      0.94      0.94       147
weighted avg       0.94      0.94      0.94       147


 Accuracy:  0.9387755102040817
```



```
Precision by XGB of testing data is: 0.939
Recall by XGB of testing data is: 0.939
F1 score by XGB of testing data is: 0.939
```

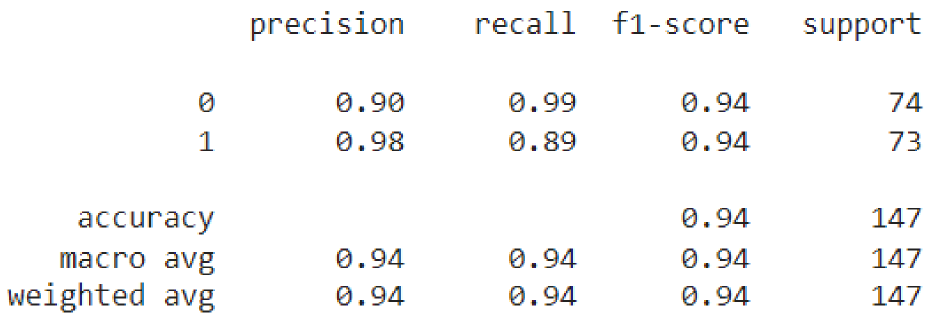


```
Text(33.0, 0.5, 'Original Values')
```

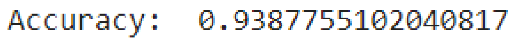


```
Sensitivity :  0.9315068493150684
Specificity :  0.9459459459459459
```

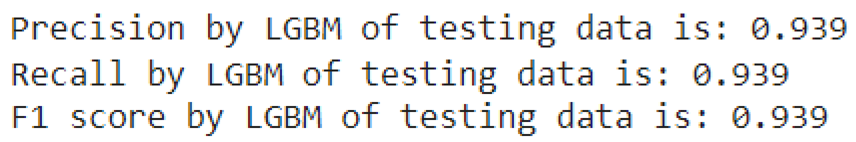
LGBM Classifier



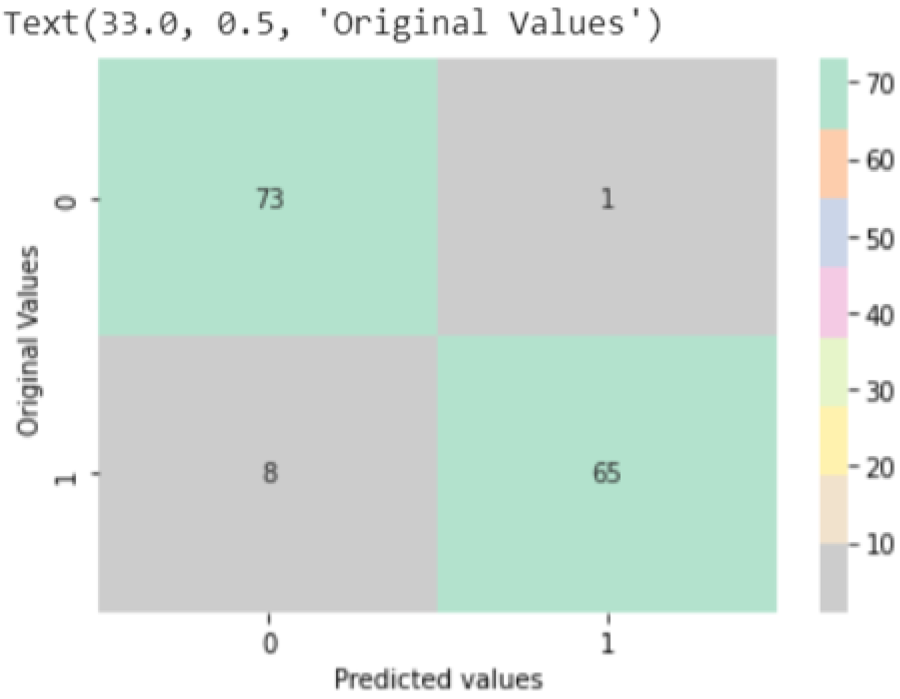
```
              precision    recall  f1-score   support

           0       0.90      0.99      0.94        74
           1       0.98      0.89      0.94        73

    accuracy                           0.94       147
   macro avg       0.94      0.94      0.94       147
weighted avg       0.94      0.94      0.94       147
```



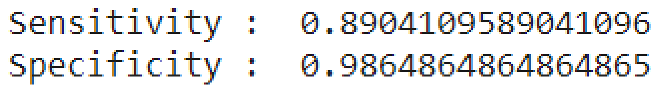
```
Accuracy:  0.9387755102040817
```



```
Precision by LGBM of testing data is: 0.939
Recall by LGBM of testing data is: 0.939
F1 score by LGBM of testing data is: 0.939
```

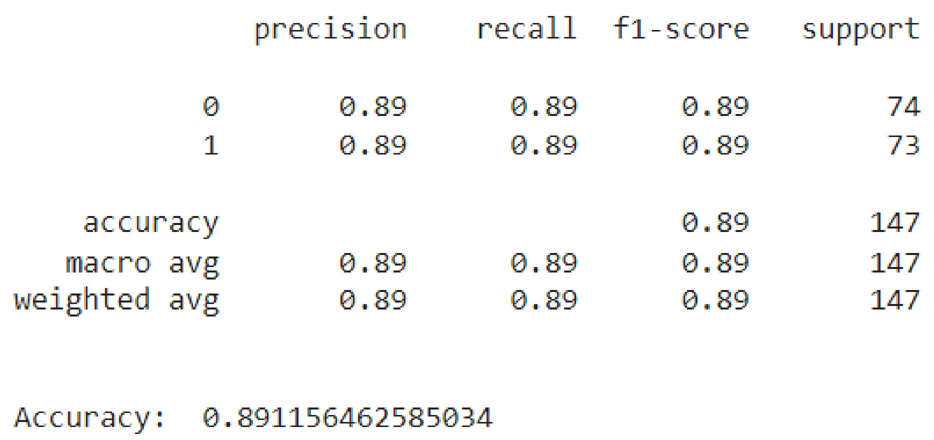


```
Text(33.0, 0.5, 'Original Values')
```

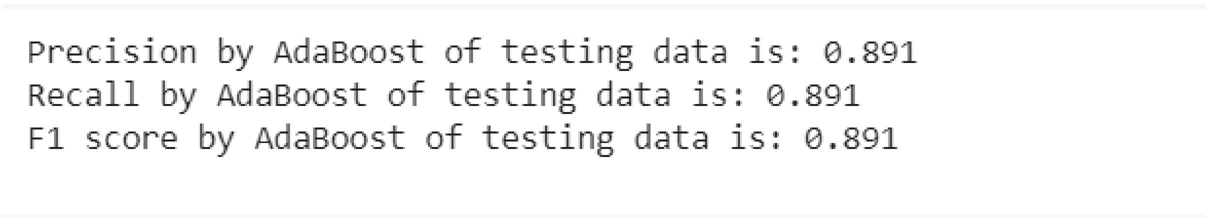


```
Sensitivity :  0.8904109589041096
Specificity :  0.9864864864864865
```

Adaboost Classifier

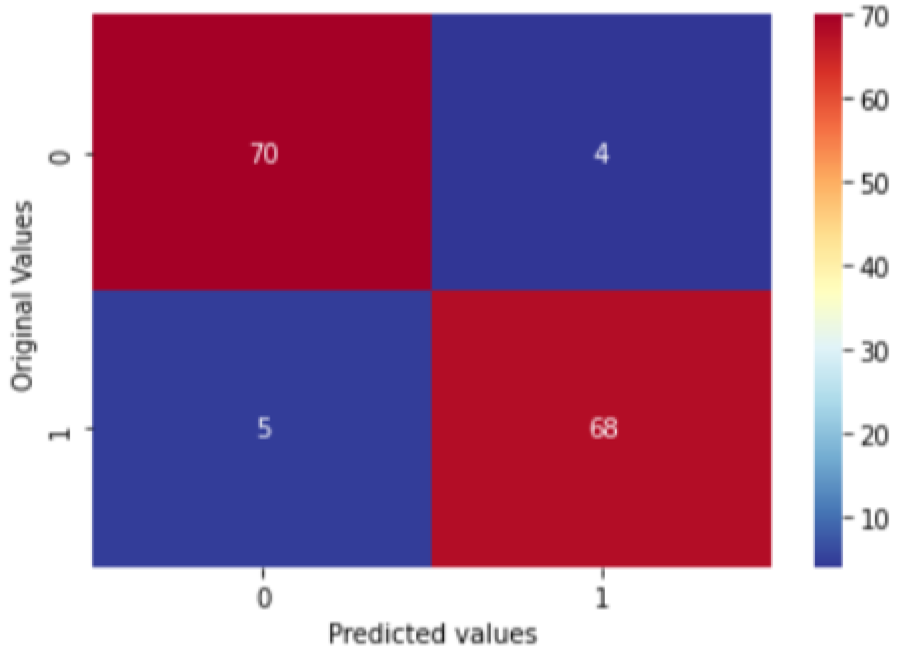


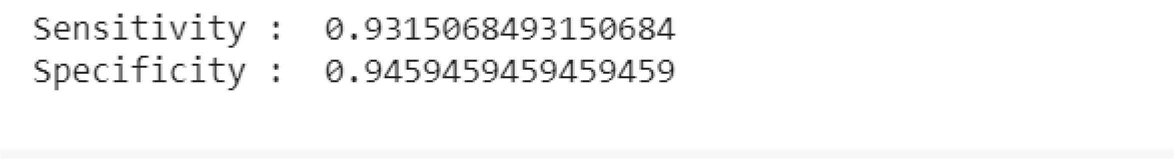
```
              precision    recall  f1-score   support

           0       0.89      0.89      0.89        74
           1       0.89      0.89      0.89        73

    accuracy                           0.89       147
   macro avg       0.89      0.89      0.89       147
weighted avg       0.89      0.89      0.89       147


Accuracy:  0.891156462585034
```



```
Precision by AdaBoost of testing data is: 0.891
Recall by AdaBoost of testing data is: 0.891
F1 score by AdaBoost of testing data is: 0.891
```

```
Text(33.0, 0.5, 'Original Values')
```





```
Sensitivity :  0.9315068493150684
Specificity :  0.9459459459459459
```

Propose Result table (before undersampling)

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
XGBC	**98%**	**0.99**	**0.99**	**0.99**
ADBC	**97%**	**0.97**	**0.97**	**0.97**
LGBM	**99%**	**0.99**	**0.99**	**0.99**

Propose Result table (After undersampling)

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**
XGBC	**93%**	**0.94**	**0.94**	**0.94**
ADBC	**89%**	**0.89**	**0.89**	**0.89**
LGBM	**93%**	**0.94**	**0.94**	**0.94**

Base and proposed results comparison table

	Accuracy	**Precision**	**Recall**	**F1 score**
Base	**97.26%**	**0.97**	**0.97**	**0.97**
Propose	**99%**	**0.99**	**0.99**	**0.99**