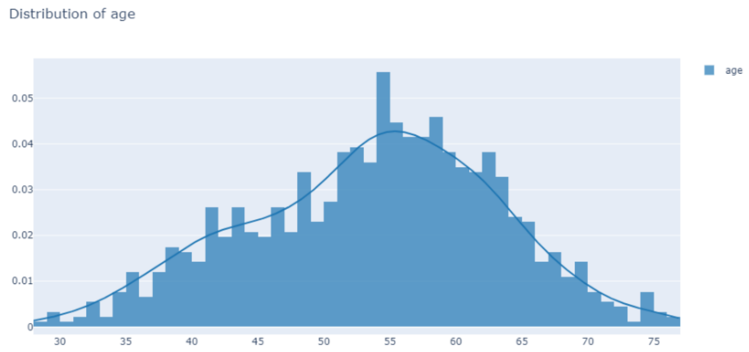
Introduction:

According to Long Beach and Cleveland Clinic Foundation, cardiovascular diseases are the number one cause of death globally, taking a toll of eighteen million lives each year, and accounts for thirty-one percent of all deaths worldwide. In a simple sense, four out of five deaths in cardiovascular diseases are due to heart attacks and strokes, and one-third of these deaths occur prematurely in people under seventy years of age. Heart failure is a common event caused by cardiovascular diseases and this dataset contains twelve features that can be used to predict a possible heart disease:

1. Age: age of the patient [years].
2. Sex: sex of the patient [M: Male, F: Female].
3. ChestPainType: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic].
4. RestingBP: resting blood pressure [mm Hg].
5. Cholesterol: serum cholesterol [mm/dl].
6. FastingBS: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise].
7. RestingECG: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria].
8. MaxHR: maximum heart rate achieved [Numeric value between 60 and 202].
9. ExerciseAngina: exercise-induced angina [Y: Yes, N: No].
10. Oldpeak: oldpeak = ST [Numeric value measured in depression].
11. ST_Slope: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping].
12. HeartDisease: output class [1: heart disease, 0: Normal].

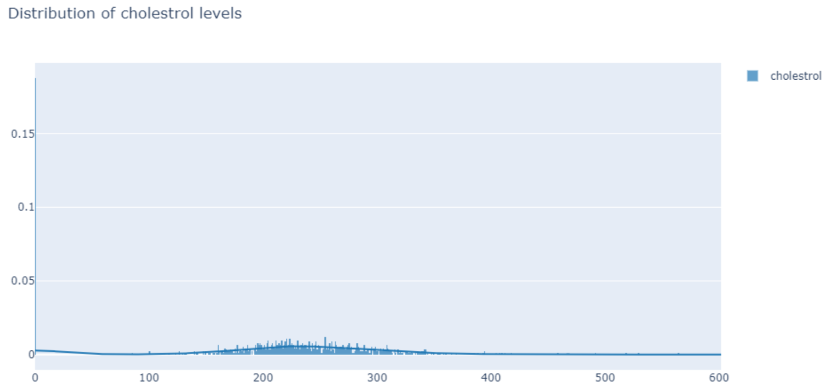
People with cardiovascular disease or who are at high cardiovascular risk is due to the presence of one or more risk factors such as hypertension, diabetes, hyperlipidaemia or already established disease. Therefore, machine learning can be a great asset to the early detection and mangement of people who have cardiovascular disease or who simply have high cardiovascular risk.

Description Statistics:

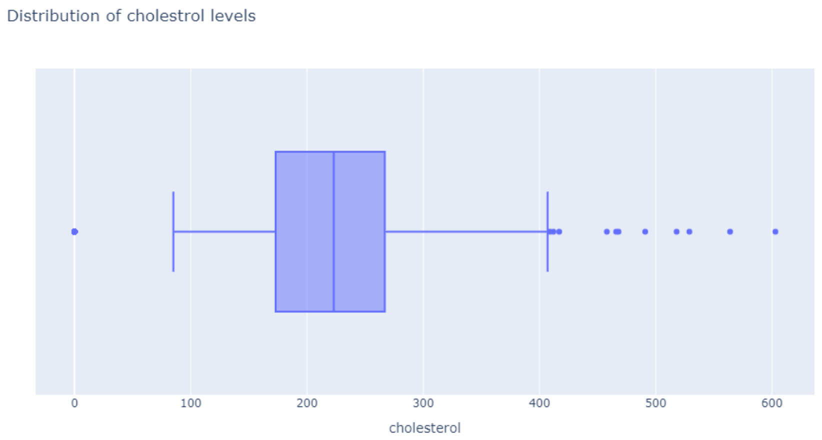


Mean age: 53.510893246187365

Median age: 54.0



The values are concentrated between 200-300 but we see the range is 438. The reason is because of outliers in the data.

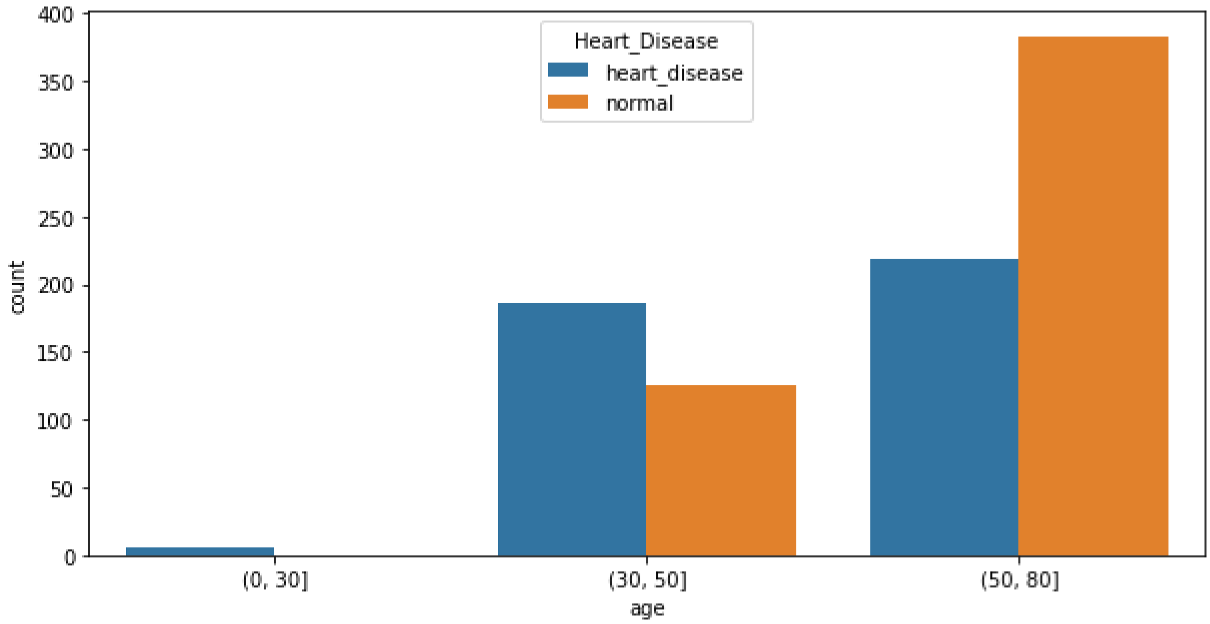


The numbers outside the range of 115.75 and 369.75 will be considered as outliers The box plot verify our calculation. All the number greater than 369.75 are shown as outliers.

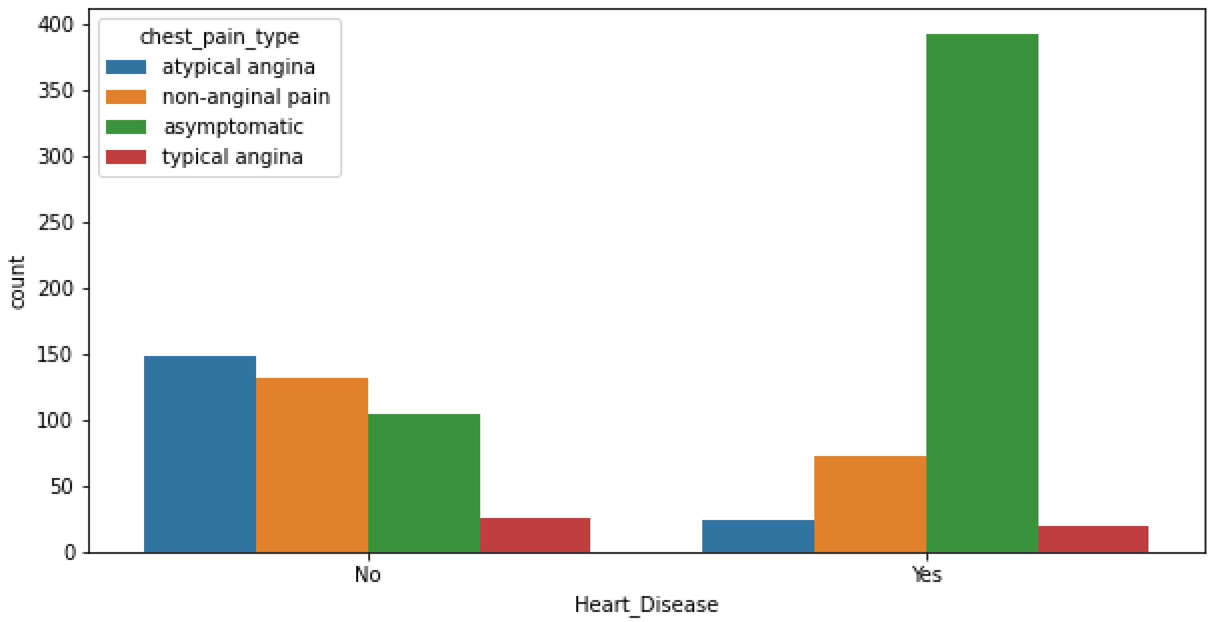
Data Cleaning Problems:

1. Correcting the data types.
2. Removing the unnecessary columns or values.
3. Handling missing values.

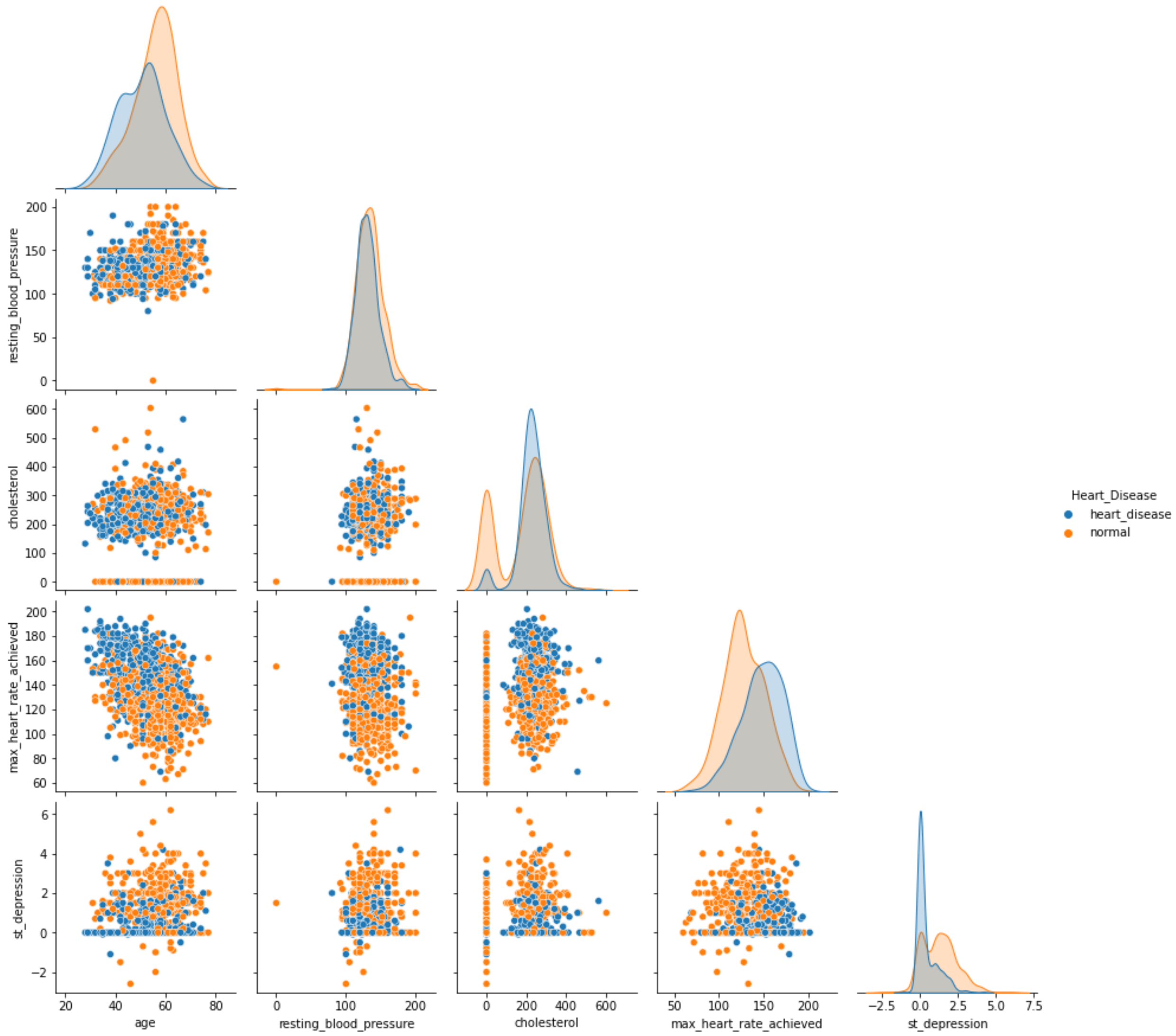
EDA:



Adults age 50 and older are more likely than younger people to have heart disease.



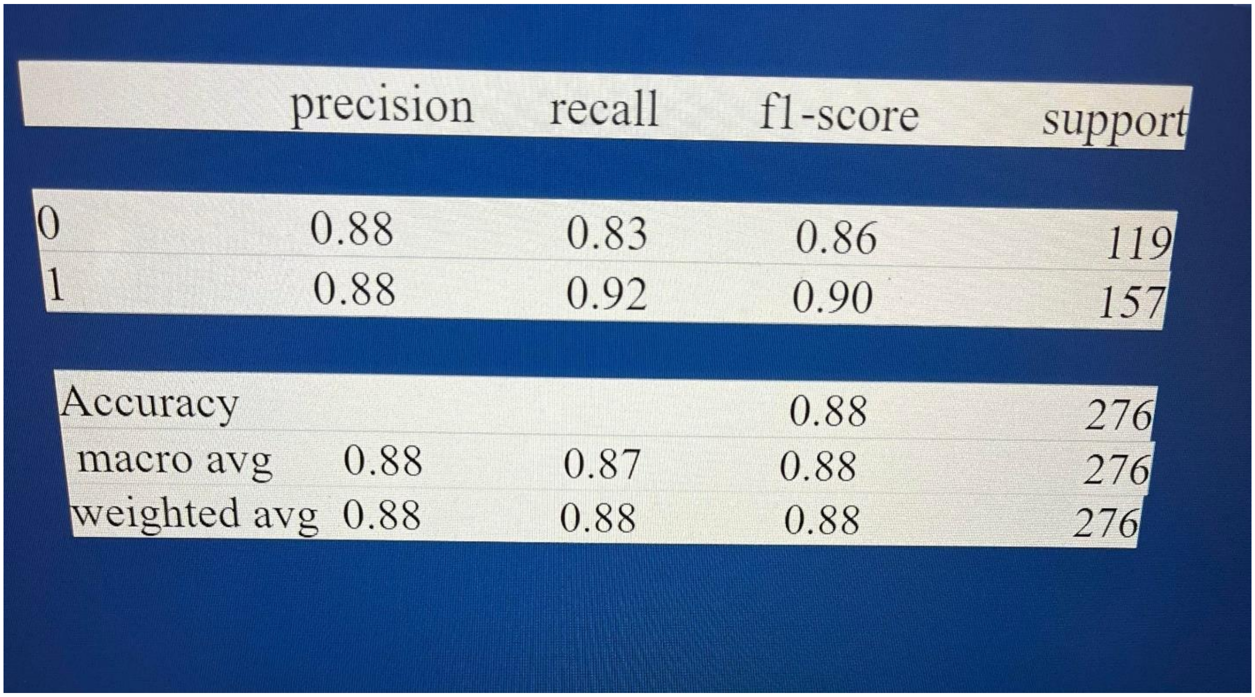
Asymptomatic is the most common type of chest pain among patients with heart disease.



A pair plot between heart disease with age, sex, chestPainType, restingBP, cholesterol, fastingBS, restingECG, maxHR ExerciseAngina, Old_Peak, and ST_Slope.

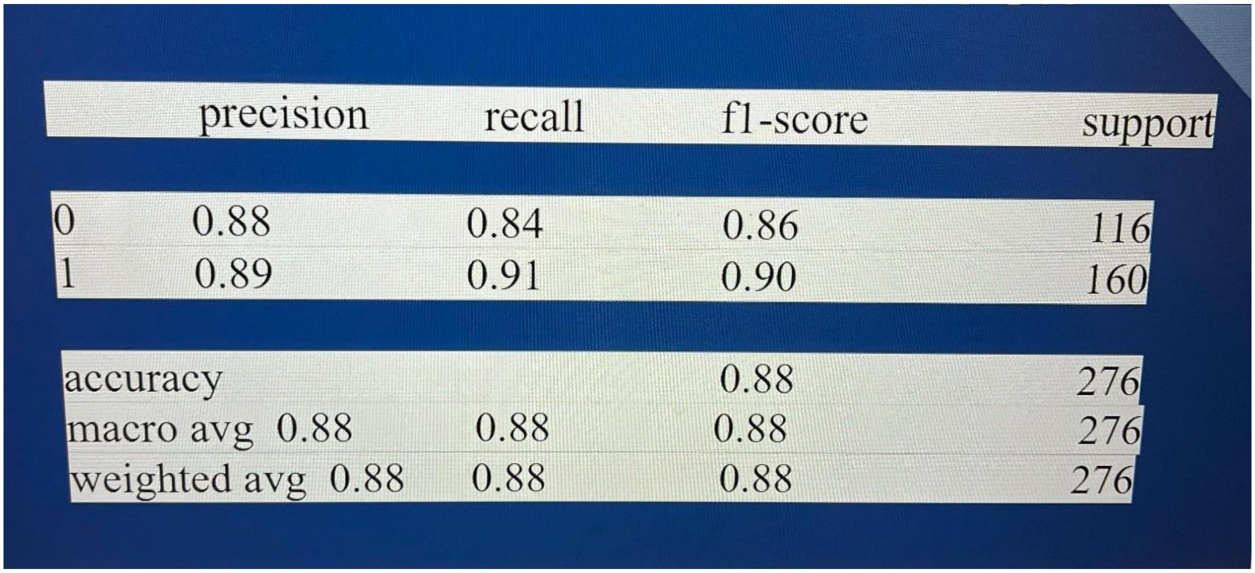
Machine Learning:

Logistic Regression Result



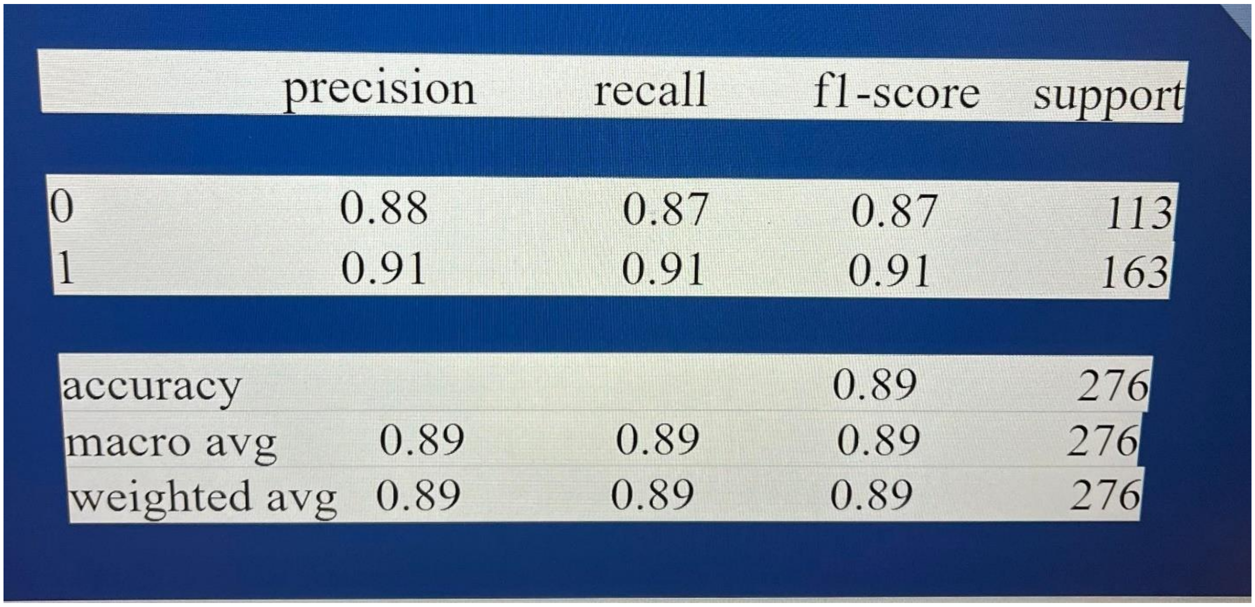
	precision	recall	f1-score	support
0	0.88	0.83	0.86	119
1	0.88	0.92	0.90	157
Accuracy			0.88	276
macro avg	0.88	0.87	0.88	276
weighted avg	0.88	0.88	0.88	276

Random Forest Classification Result



	precision	recall	f1-score	support
0	0.88	0.84	0.86	116
1	0.89	0.91	0.90	160
accuracy			0.88	276
macro avg	0.88	0.88	0.88	276
weighted avg	0.88	0.88	0.88	276

SVC Result

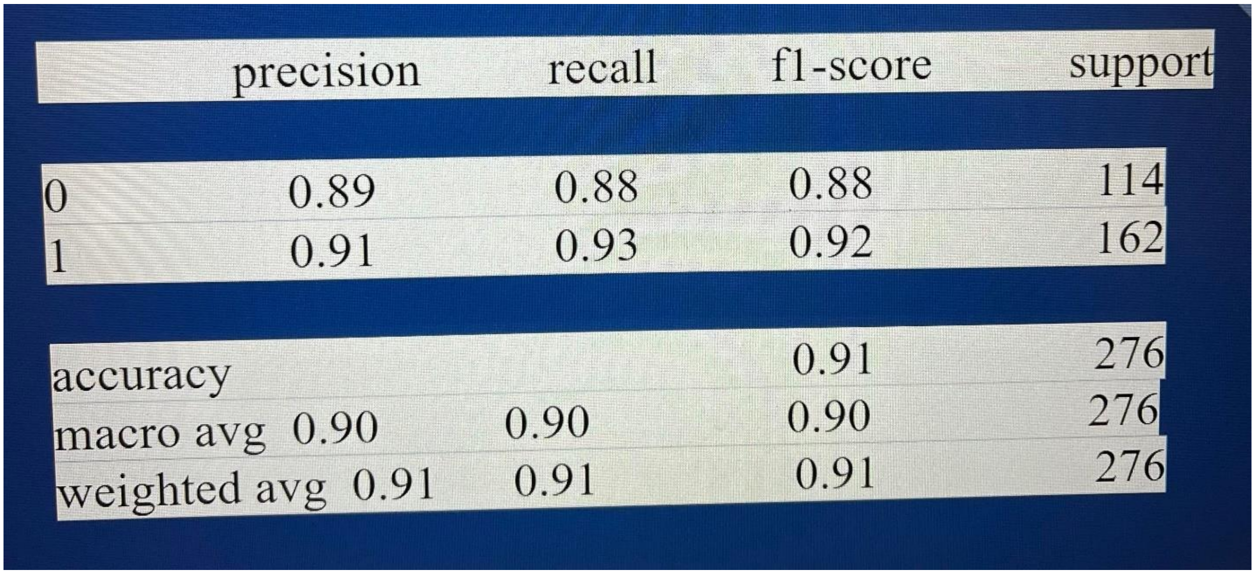


```
              precision    recall  f1-score   support

0                  0.88      0.87      0.87       113
1                  0.91      0.91      0.91       163

accuracy                               0.89       276
macro avg          0.89      0.89      0.89       276
weighted avg       0.89      0.89      0.89       276
```

Catboost Result

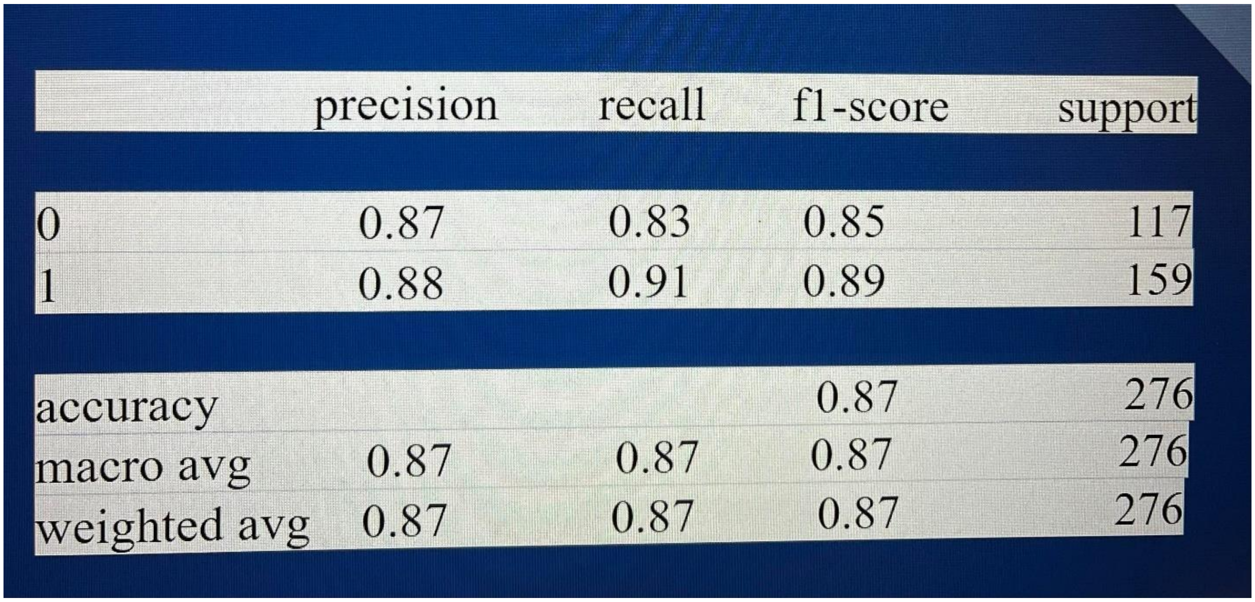


```
              precision    recall  f1-score   support

0                  0.89      0.88      0.88       114
1                  0.91      0.93      0.92       162

accuracy                               0.91       276
macro avg          0.90      0.90      0.90       276
weighted avg       0.91      0.91      0.91       276
```

LGBM Result

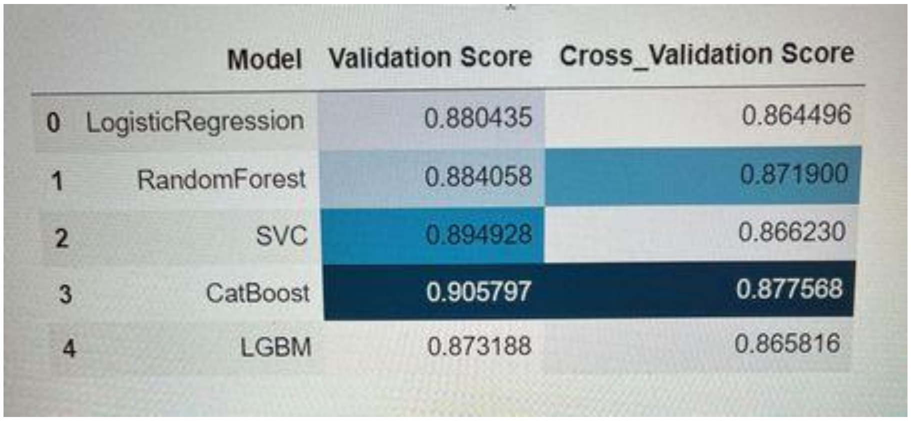


```
              precision    recall  f1-score   support

0                0.87       0.83     0.85        117
1                0.88       0.91     0.89        159

accuracy                             0.87        276
macro avg        0.87       0.87     0.87        276
weighted avg     0.87       0.87     0.87        276
```

Best Result



	Model	Validation Score	Cross_Validation Score
0	LogisticRegression	0.880435	0.864496
1	RandomForest	0.884058	0.871900
2	SVC	0.894928	0.866230
3	CatBoost	0.905797	0.877568
4	LGBM	0.873188	0.865816

As we can see, all of our models obtained decent results within the validation and cross validation score. We can see that the Light GBM model performed the worse than the Logistic Regression, SVC, and Random Forest within the validation score. It seems CatBoost performed the best within the validation score Within the cross-validation score we can see that Logistic Regression performed the worst. While CatBoost seems to performed the best within the cross-validation score as well