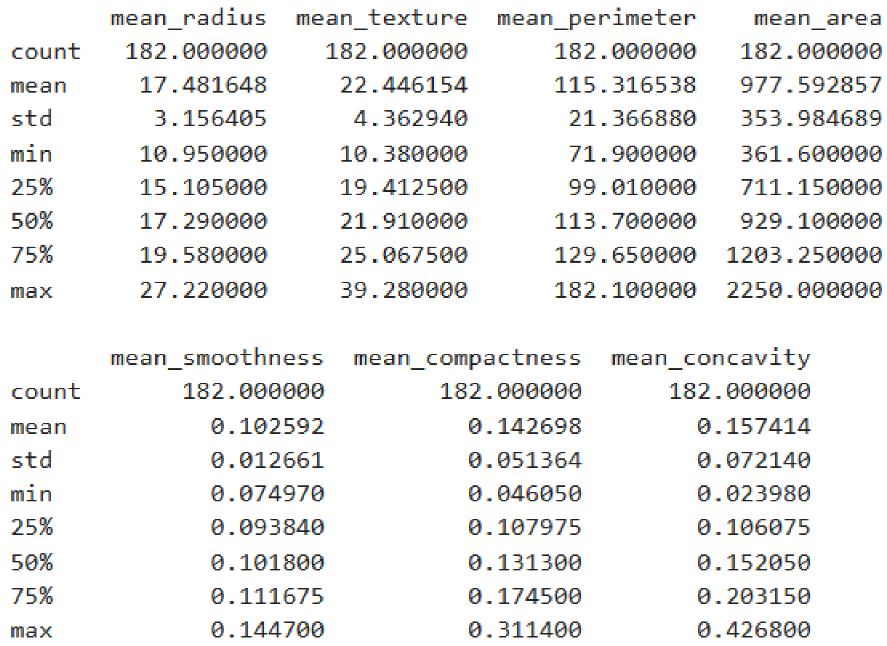
ASSIGNMENT 2 – MACHINE LEARNING_REPORT

SUMATH CHANDRA KILLE

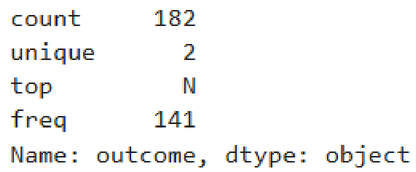
1A. These are the values for Summarization of statistics of these variables into count, mean, standard deviation, minimum, 25% percentile, 50% percentile, 75% percentile, and maximum. I have used describe to find these.



```
       mean_radius  mean_texture  mean_perimeter    mean_area
count   182.000000    182.000000      182.000000   182.000000
mean     17.481648     22.446154      115.316538   977.592857
std       3.156405      4.362940       21.366880   353.984689
min      10.950000     10.380000       71.900000   361.600000
25%      15.105000     19.412500       99.010000   711.150000
50%      17.290000     21.910000      113.700000   929.100000
75%      19.580000     25.067500      129.650000  1203.250000
max      27.220000     39.280000      182.100000  2250.000000

       mean_smoothness  mean_compactness  mean_concavity
count       182.000000        182.000000      182.000000
mean          0.102592          0.142698        0.157414
std           0.012661          0.051364        0.072140
min           0.074970          0.046050        0.023980
25%           0.093840          0.107975        0.106075
50%           0.101800          0.131300        0.152050
75%           0.111675          0.174500        0.203150
max           0.144700          0.311400        0.426800
```

.

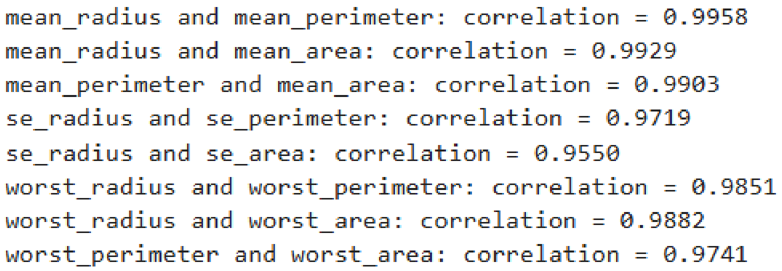
1B. As the target varibale is categorical variable we are summarizing stats of varibale into count, unique value, top value and frequency



```
count     182
unique      2
top         N
freq      141
Name: outcome, dtype: object
```

1C.Yes, we can encode the outcome varibale from categorical to numerical data type using a technique called label encodeing which is especiallly suitable for binay categories.

1D.

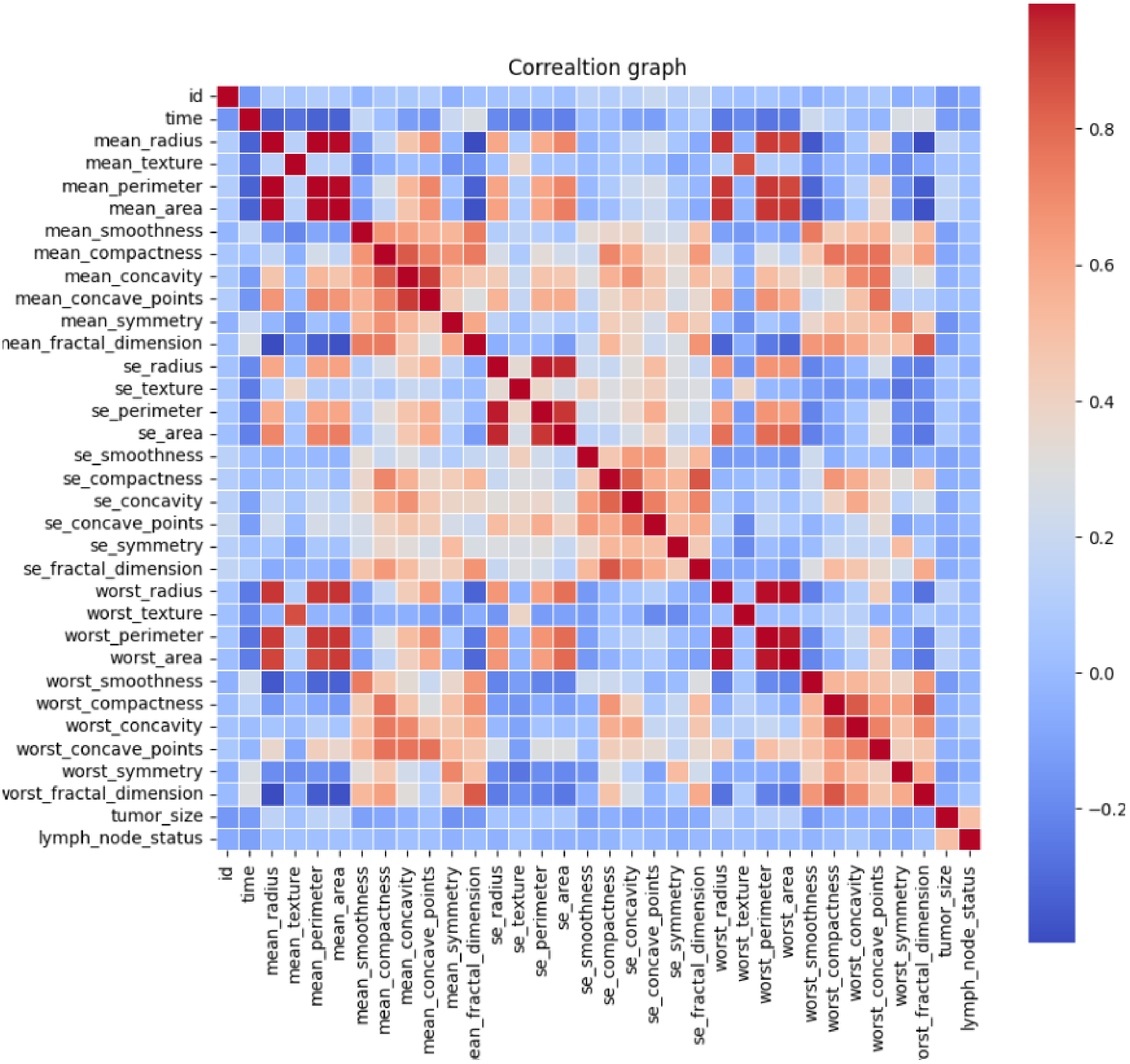


```
mean_radius and mean_perimeter: correlation = 0.9958
mean_radius and mean_area: correlation = 0.9929
mean_perimeter and mean_area: correlation = 0.9903
se_radius and se_perimeter: correlation = 0.9719
se_radius and se_area: correlation = 0.9550
worst_radius and worst_perimeter: correlation = 0.9851
worst_radius and worst_area: correlation = 0.9882
worst_perimeter and worst_area: correlation = 0.9741
```

Indeed, the dataset contains redundant features. Pairs that hold overlapping information, such as mean_radius and mean_perimeter (correlation = 0.9958) and worst_radius and worst_area (correlation = 0.9882), are highly correlated.

Eliminating a single feature from these pairs will simplify the model, decrease multicollinearity, and increase efficiency without affecting the analysis.

Observation: The mean_, se_, and worst_ groups' attributes are highly correlated, and the outcome variable is unbalanced.

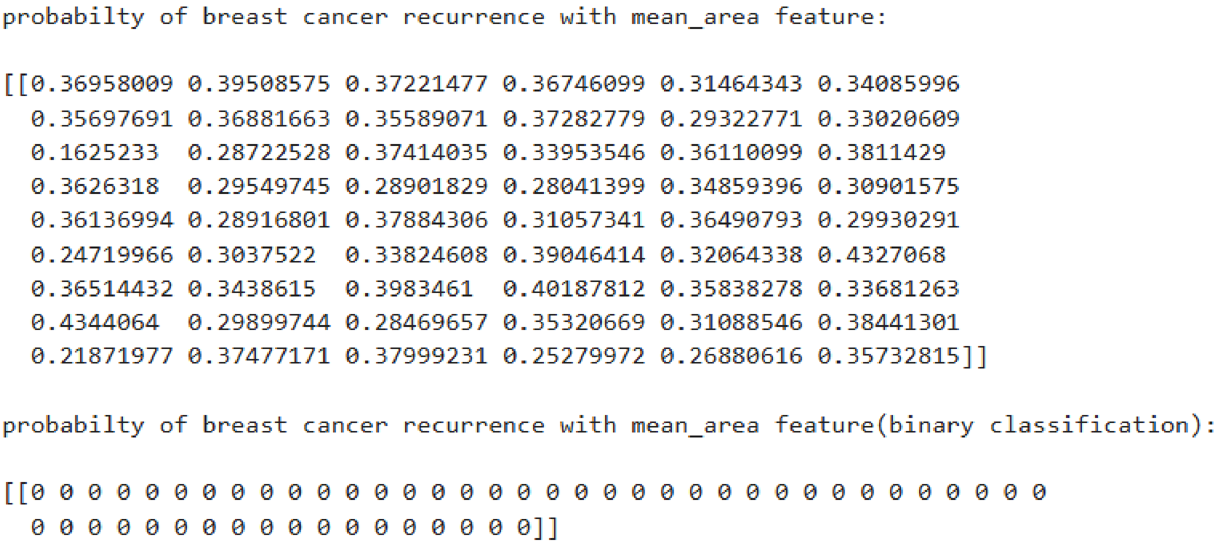


In the graph we can see the red colored sqares are highly correalted which means srong positive correaltion like mean_radius, mean_perimeter, mean_area.
Blue colored are strong negative corelated
white or light colored are no linear relationship among those feature pairs.
1E. The correaltion between mean perimeter and SE perimeter is 0.60668619726589

Q2. Logistic regressiion with one variable

A.) In this task, we investigated the connection between a tumor’s mean_area and the chances of breast cancer recurrence. We implemented logistic regression from the ground up, creating essential components such as the sigmoid, cost, and prediction functions. The dataset was divided into training and testing sets, and the model was trained using only the mean_area feature.

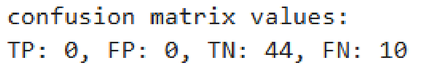
lr = 0.00001, iterations_count = 1000



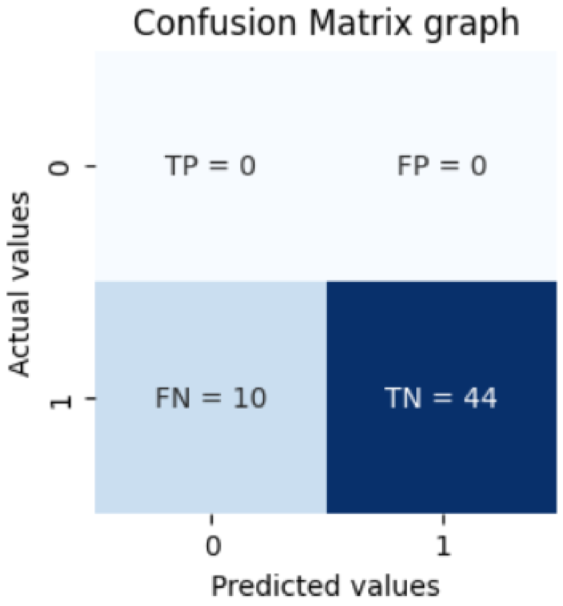
```
probabilty of breast cancer recurrence with mean_area feature:

[[0.36958009 0.39508575 0.37221477 0.36746099 0.31464343 0.34085996
  0.35697691 0.36881663 0.35589071 0.37282779 0.29322771 0.33020609
  0.1625233  0.28722528 0.37414035 0.33953546 0.36110099 0.3811429
  0.3626318  0.29549745 0.28901829 0.28041399 0.34859396 0.30901575
  0.36136994 0.28916801 0.37884306 0.31057341 0.36490793 0.29930291
  0.24719966 0.3037522  0.33824608 0.39046414 0.32064338 0.4327068
  0.36514432 0.3438615  0.3983461  0.40187812 0.35838278 0.33681263
  0.4344064  0.29899744 0.28469657 0.35320669 0.31088546 0.38441301
  0.21871977 0.37477171 0.37999231 0.25279972 0.26880616 0.35732815]]

probabilty of breast cancer recurrence with mean_area feature(binary classification):

[[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]]
```

B. To evaluate performance of logistic regression model we ues confusion matrix which has TP,TN,FP, FN which are use to model eavaluaton metrices like accuarcy, precision, recall and f1 score

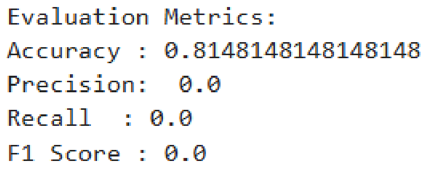
->The model is based towards predicting the non-recurrecne and fails to identify recurrence cases(breast cnacer), which is critical.



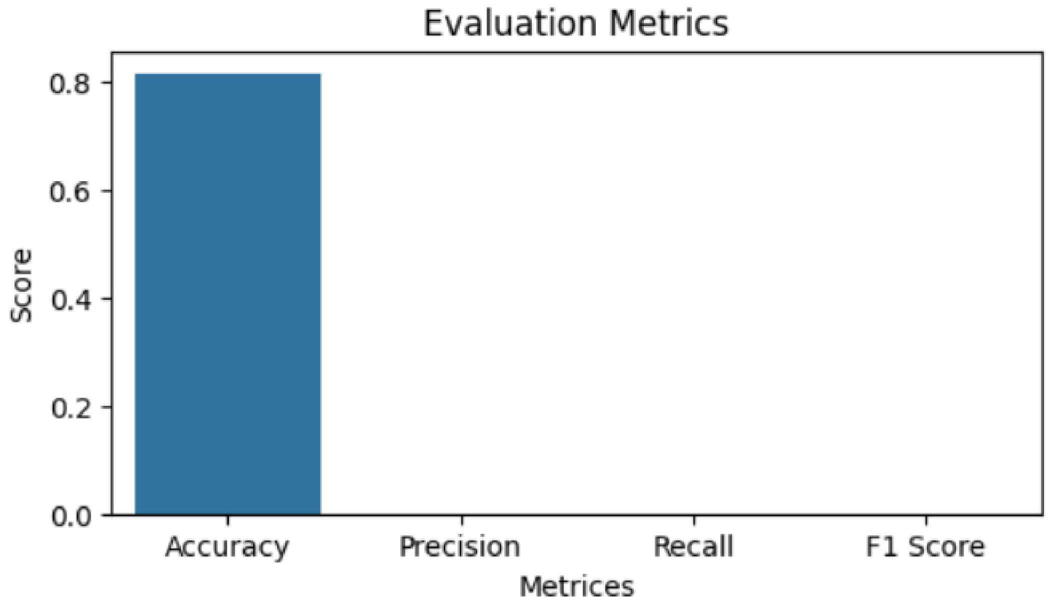
```
confusion matrix values:
TP: 0, FP: 0, TN: 44, FN: 10
```



->From evaluation metrices we can say that despite high accuracy the modle performs poorly on recurrence classes means mean_area is insuficeant and we need to use other fetaures for better performance.



```
Evaluation Metrics:
Accuracy : 0.8148148148148148
Precision:  0.0
Recall   : 0.0
F1 Score : 0.0
```

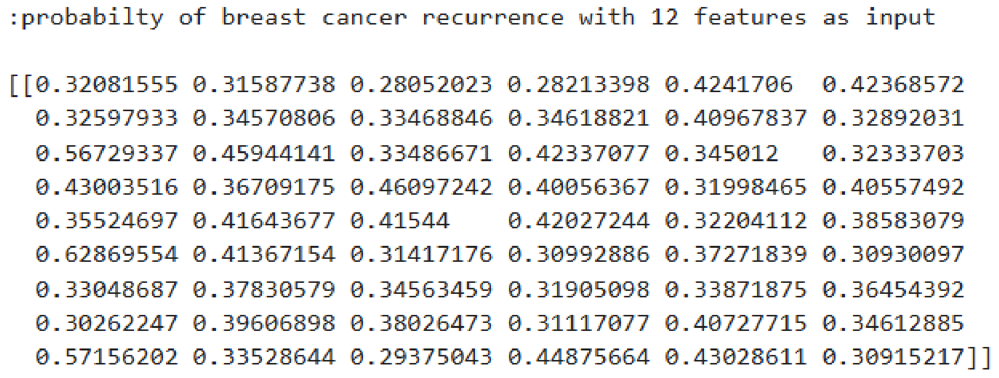


Q3.Logistic regression with multiple variables.

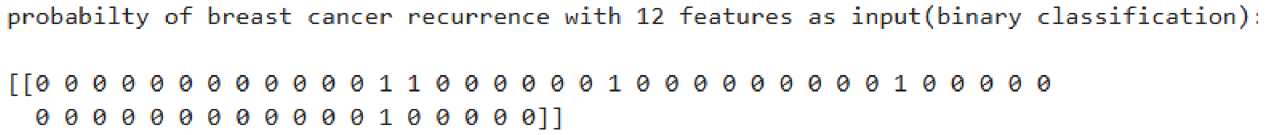
3A. In this task, we aim to build a Logistic Regression model to predict the likelihood of breast cancer recurrence based on a selected set of 12 features from the dataset

Features: mean_radius, mean_texture, mean_perimeter, mean_area, mean_smoothness, mean_compactness, mean_concavity, mean_concave_points, mean_fractal_dimension, se_perimeter, se_texture, se_area

-> Below are the probabilities of breast cancer recurrence using 12 variables as a input.



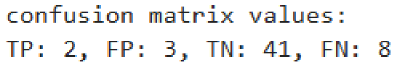
```
:probabilty of breast cancer recurrence with 12 features as input

[[0.32081555 0.31587738 0.28052023 0.28213398 0.4241706  0.42368572
  0.32597933 0.34570806 0.33468846 0.34618821 0.40967837 0.32892031
  0.56729337 0.45944141 0.33486671 0.42337077 0.345012   0.32333703
  0.43003516 0.36709175 0.46097242 0.40056367 0.31998465 0.40557492
  0.35524697 0.41643677 0.41544    0.42027244 0.32204112 0.38583079
  0.62869554 0.41367154 0.31417176 0.30992886 0.37271839 0.30930097
  0.33048687 0.37830579 0.34563459 0.31905098 0.33871875 0.36454392
  0.30262247 0.39606898 0.38026473 0.31117077 0.40727715 0.34612885
  0.57156202 0.33528644 0.29375043 0.44875664 0.43028611 0.30915217]]
```



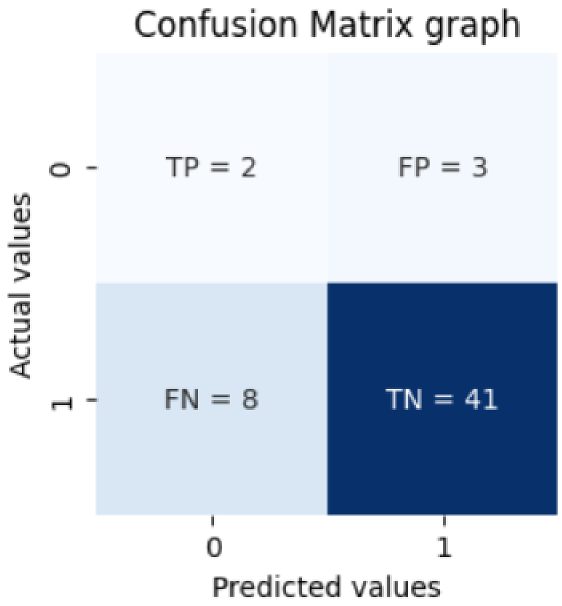
```
probabilty of breast cancer recurrence with 12 features as input(binary classification):

[[0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
  0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0]]
```

->Below is the confusion matrix with values of TP FP TN FN and with graph

-> Although the model does a respectable job of detecting non-recurrence situations, it still overlooks a sizable portion of recurrence cases. This indicates poor performance with room for development, particularly in terms of identifying real positives.

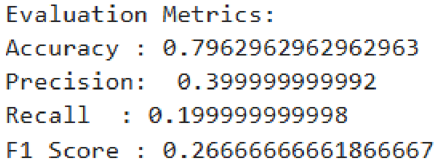


```
confusion matrix values:
TP: 2, FP: 3, TN: 41, FN: 8
```

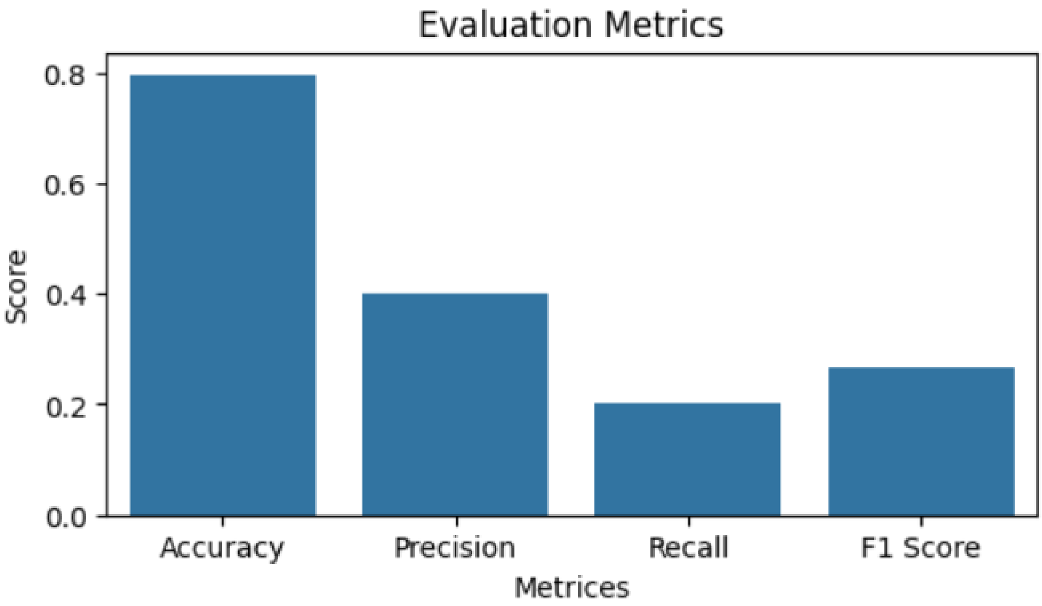


->Below are the Evaluation metrice values and graph for 12 fetaure model

-> Although the model shows good overall accuracy, it has difficulty identifying recurrence cases, as seen in its low recall. This means it may miss high-risk patients and isn't yet reliable for medical decision-making without further tuning or adding more informative features

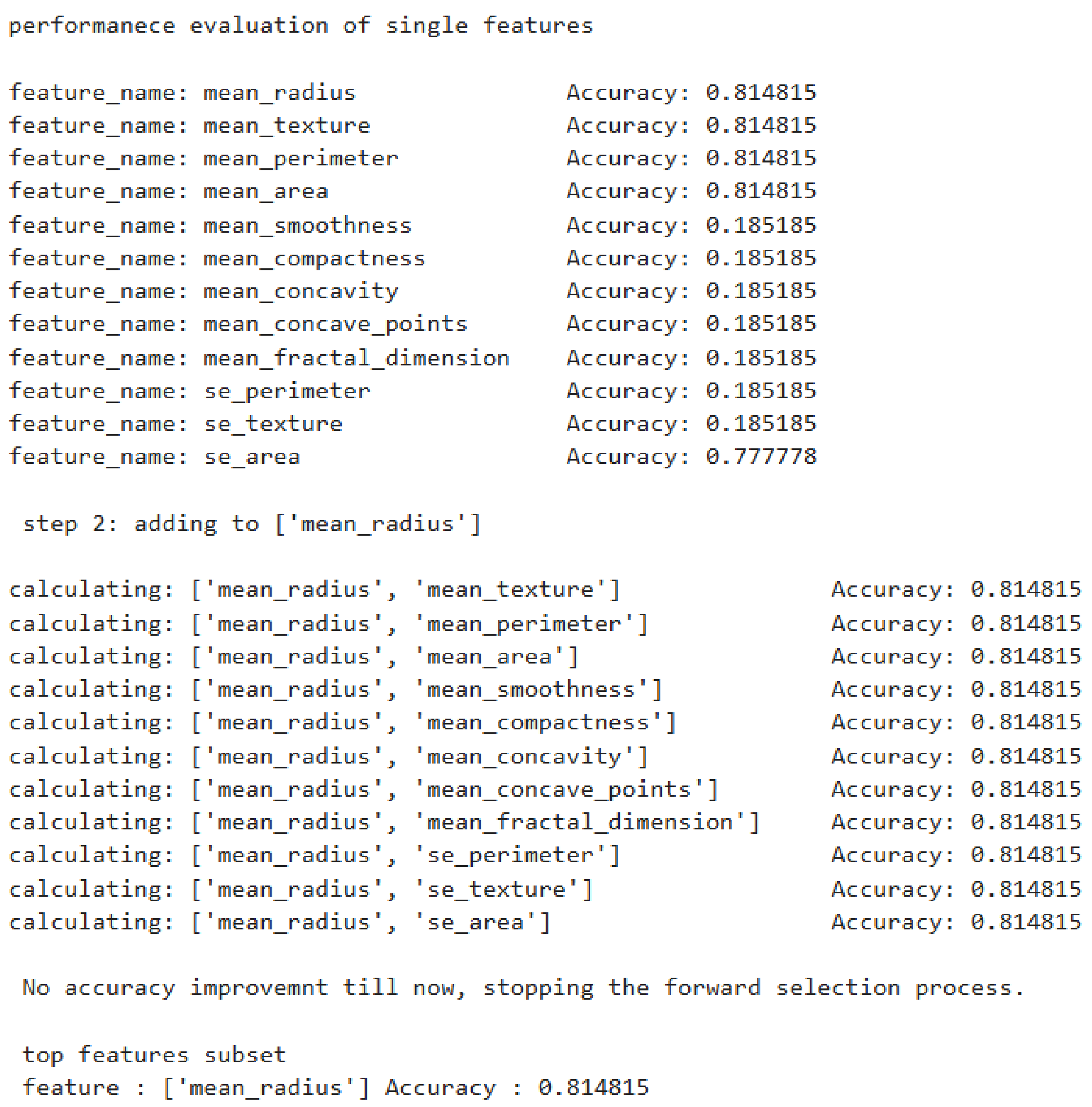


```
Evaluation Metrics:
Accuracy : 0.7962962962962963
Precision:  0.399999999992
Recall  : 0.199999999998
F1 Score : 0.26666666661866667
```



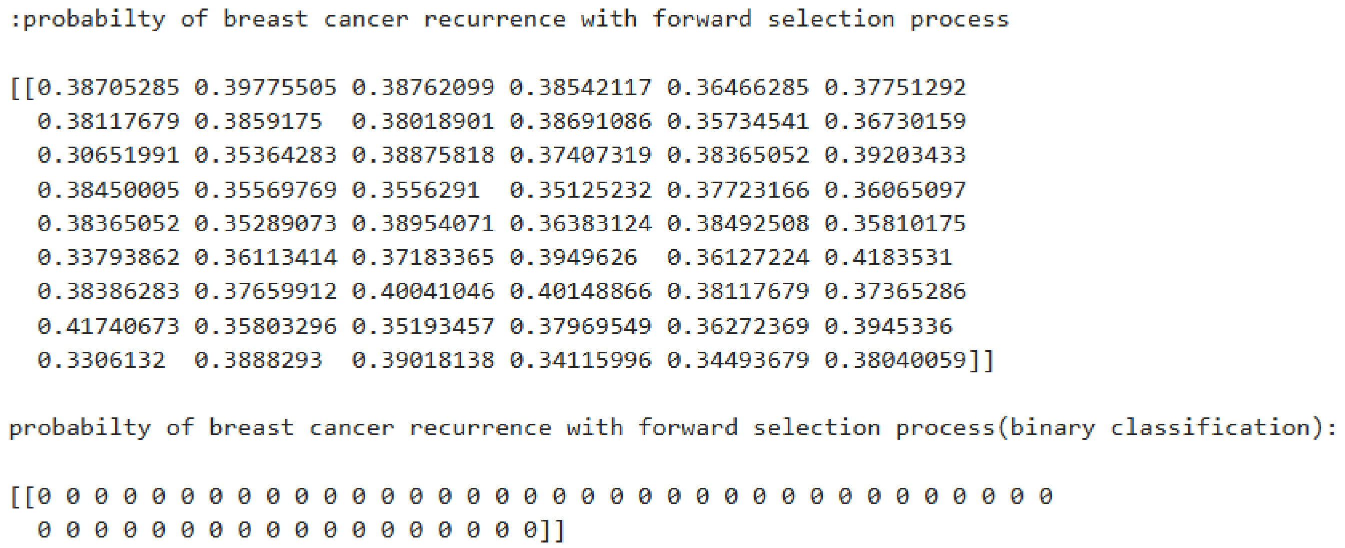
3B. FORWARD SELECTION

The feature subset ['mean_radius'] achieved the highest accuracy of 0.8148. Since adding additional features didn't enhance the performance, it indicates that mean_radius alone holds significant predictive value, and the rest may offer little extra benefit. It's quite notable that a single feature can be this effective, emphasizing the importance of tumor size in forecasting recurrence.



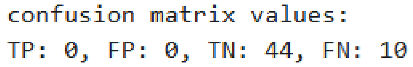
```
performanece evaluation of single features

feature_name: mean_radius                Accuracy: 0.814815
feature_name: mean_texture               Accuracy: 0.814815
feature_name: mean_perimeter             Accuracy: 0.814815
feature_name: mean_area                  Accuracy: 0.814815
feature_name: mean_smoothness            Accuracy: 0.185185
feature_name: mean_compactness           Accuracy: 0.185185
feature_name: mean_concavity             Accuracy: 0.185185
feature_name: mean_concave_points        Accuracy: 0.185185
feature_name: mean_fractal_dimension     Accuracy: 0.185185
feature_name: se_perimeter               Accuracy: 0.185185
feature_name: se_texture                 Accuracy: 0.185185
feature_name: se_area                    Accuracy: 0.777778

 step 2: adding to ['mean_radius']

calculating: ['mean_radius', 'mean_texture']                Accuracy: 0.814815
calculating: ['mean_radius', 'mean_perimeter']              Accuracy: 0.814815
calculating: ['mean_radius', 'mean_area']                   Accuracy: 0.814815
calculating: ['mean_radius', 'mean_smoothness']             Accuracy: 0.814815
calculating: ['mean_radius', 'mean_compactness']            Accuracy: 0.814815
calculating: ['mean_radius', 'mean_concavity']              Accuracy: 0.814815
calculating: ['mean_radius', 'mean_concave_points']         Accuracy: 0.814815
calculating: ['mean_radius', 'mean_fractal_dimension']      Accuracy: 0.814815
calculating: ['mean_radius', 'se_perimeter']                Accuracy: 0.814815
calculating: ['mean_radius', 'se_texture']                  Accuracy: 0.814815
calculating: ['mean_radius', 'se_area']                     Accuracy: 0.814815

 No accuracy improvemnt till now, stopping the forward selection process.

 top features subset
 feature : ['mean_radius'] Accuracy : 0.814815
```

->Below is the ss of probability if braest cancer using the Mean_radius from forward selection.

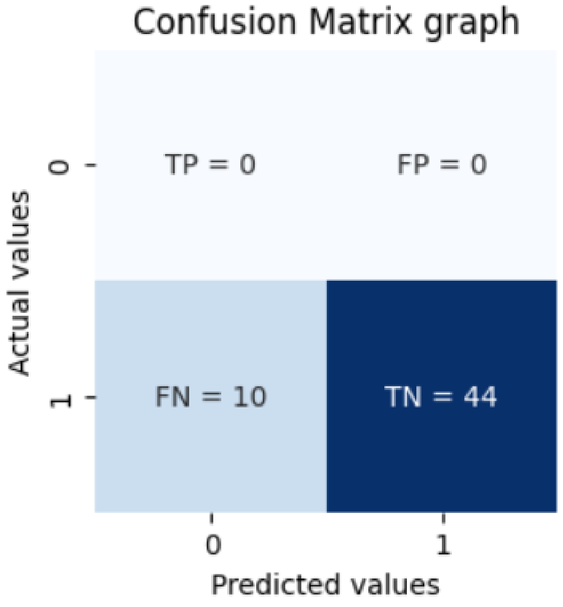


```
:probabilty of breast cancer recurrence with forward selection process

[[0.38705285 0.39775505 0.38762099 0.38542117 0.36466285 0.37751292
  0.38117679 0.3859175  0.38018901 0.38691086 0.35734541 0.36730159
  0.30651991 0.35364283 0.38875818 0.37407319 0.38365052 0.39203433
  0.38450005 0.35569769 0.3556291  0.35125232 0.37723166 0.36065097
  0.38365052 0.35289073 0.38954071 0.36383124 0.38492508 0.35810175
  0.33793862 0.36113414 0.37183365 0.3949626  0.36127224 0.4183531
  0.38386283 0.37659912 0.40041046 0.40148866 0.38117679 0.37365286
  0.41740673 0.35803296 0.35193457 0.37969549 0.36272369 0.3945336
  0.3306132  0.3888293  0.39018138 0.34115996 0.34493679 0.38040059]]

probabilty of breast cancer recurrence with forward selection process(binary classification):

[[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]]
```

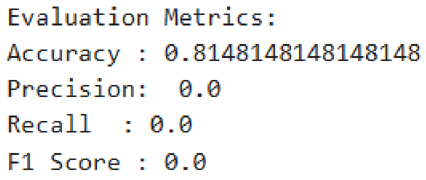
->The model fails to identify recurrence(breast cnacer)



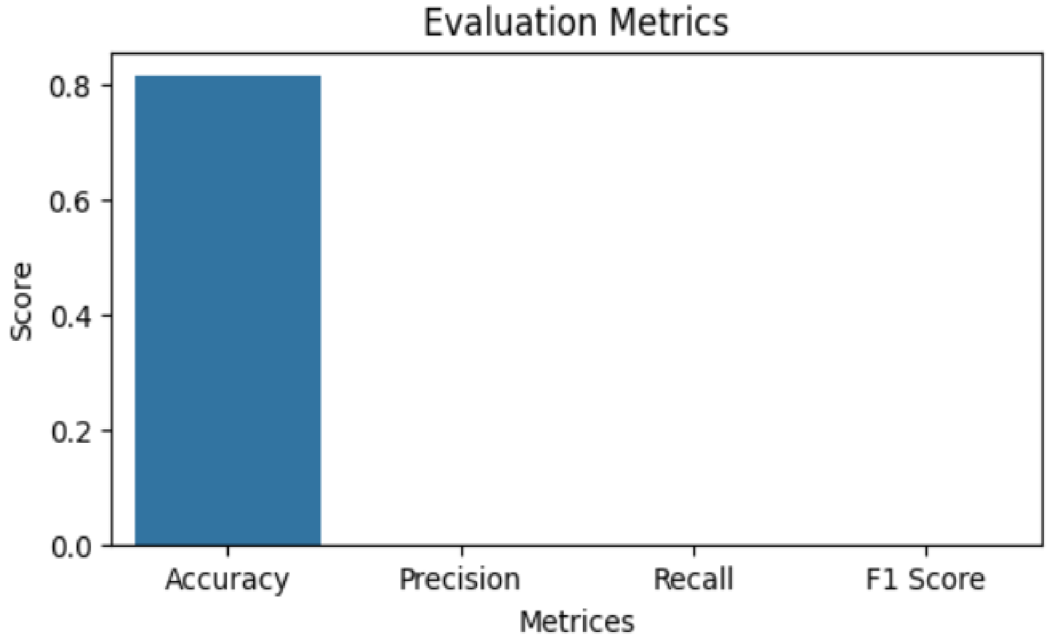
```
confusion matrix values:
TP: 0, FP: 0, TN: 44, FN: 10
```



-> Even though the model reached a high accuracy of 81.48%, it didn't identify any actual recurrence cases. As a result, precision, recall, and F1 score were all zero, showing that the model performs poorly on the minority class despite its overall accuracy looking good at first glance.



```
Evaluation Metrics:
Accuracy : 0.8148148148148148
Precision:  0.0
Recall   : 0.0
F1 Score : 0.0
```

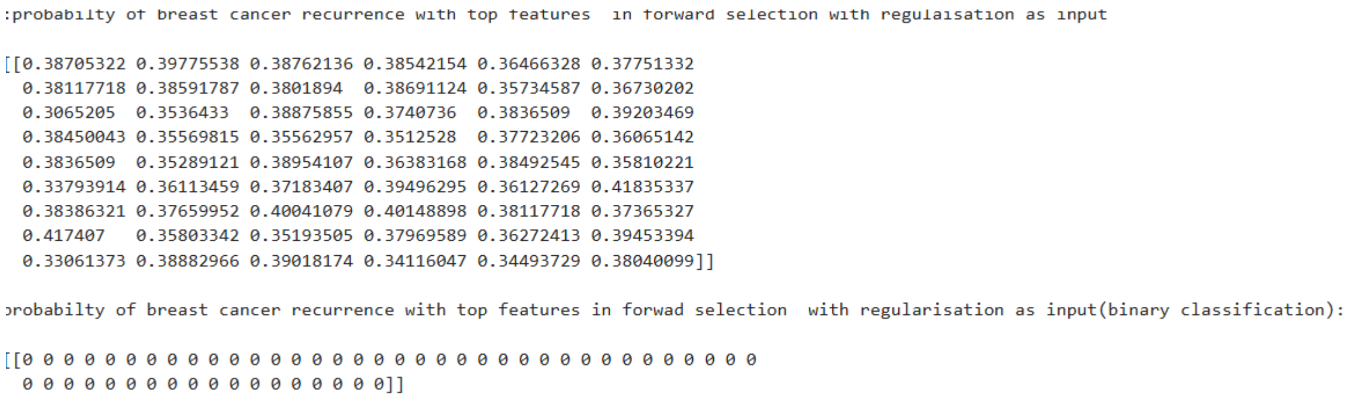


3C: The forward selection feature is having the more accuracy than the 12 feature model. By comparing this we can tell that forward feature model has more performance than the other model.

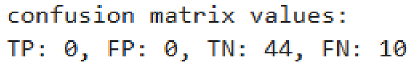
4A.

1.Regularisation:

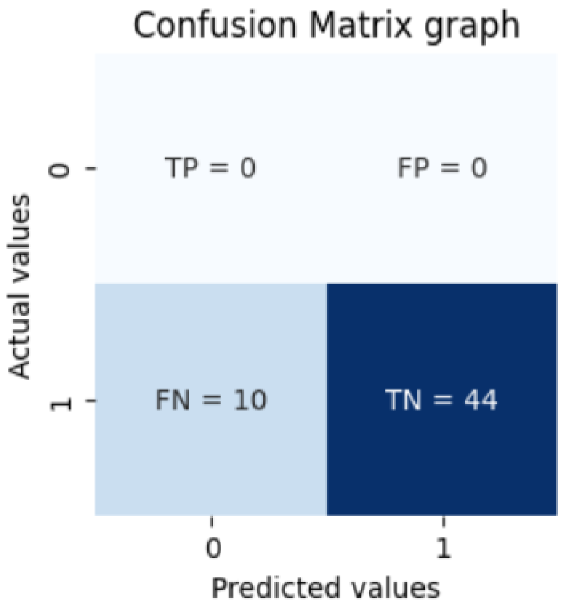
Below is the regularization applied for forward selection model and accuracy mentioned.

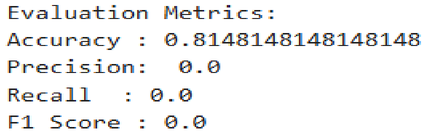


```
:probability of breast cancer recurrence with top features  in forward selection with regulaisation as input

[[0.38705322 0.39775538 0.38762136 0.38542154 0.36466328 0.37751332
  0.38117718 0.38591787 0.3801894  0.38691124 0.35734587 0.36730202
  0.3065205  0.3536433  0.38875855 0.3740736  0.3836509  0.39203469
  0.38450043 0.35569815 0.35562957 0.3512528  0.37723206 0.36065142
  0.3836509  0.35289121 0.38954107 0.36383168 0.38492545 0.35810221
  0.33793914 0.36113459 0.37183407 0.39496295 0.36127269 0.41835337
  0.38386321 0.37659952 0.40041079 0.40148898 0.38117718 0.37365327
  0.417407   0.35803342 0.35193505 0.37969589 0.36272413 0.39453394
  0.33061373 0.38882966 0.39018174 0.34116047 0.34493729 0.38040099]]

probabilty of breast cancer recurrence with top features in forwad selection  with regularisation as input(binary classification):

[[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]]
```

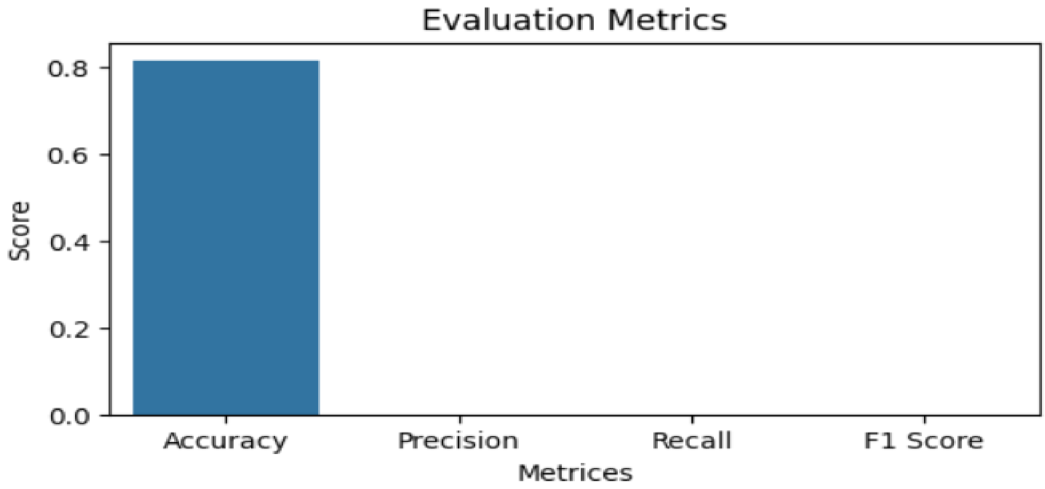


```
confusion matrix values:
TP: 0, FP: 0, TN: 44, FN: 10
```



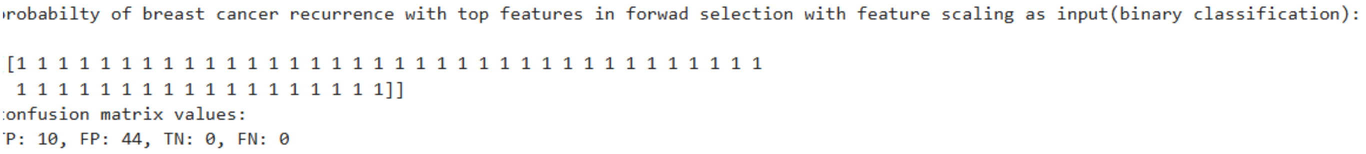


```
Evaluation Metrics:
Accuracy : 0.8148148148148148
Precision:  0.0
Recall   : 0.0
F1 Score : 0.0
```

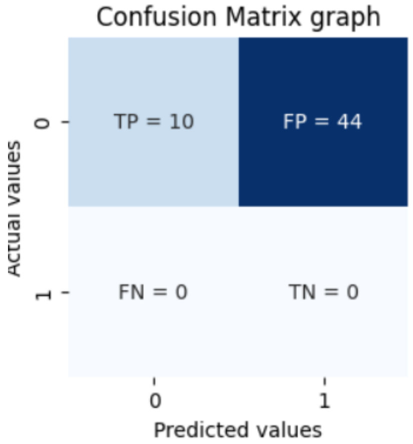


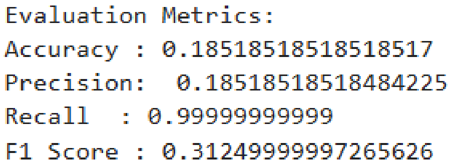
->does regularization improve the performance?
 No, there isn’ t any difference in model performance.

4A.2.Feature scaling
Forward selection model with feature scaling using mean and standard deviation

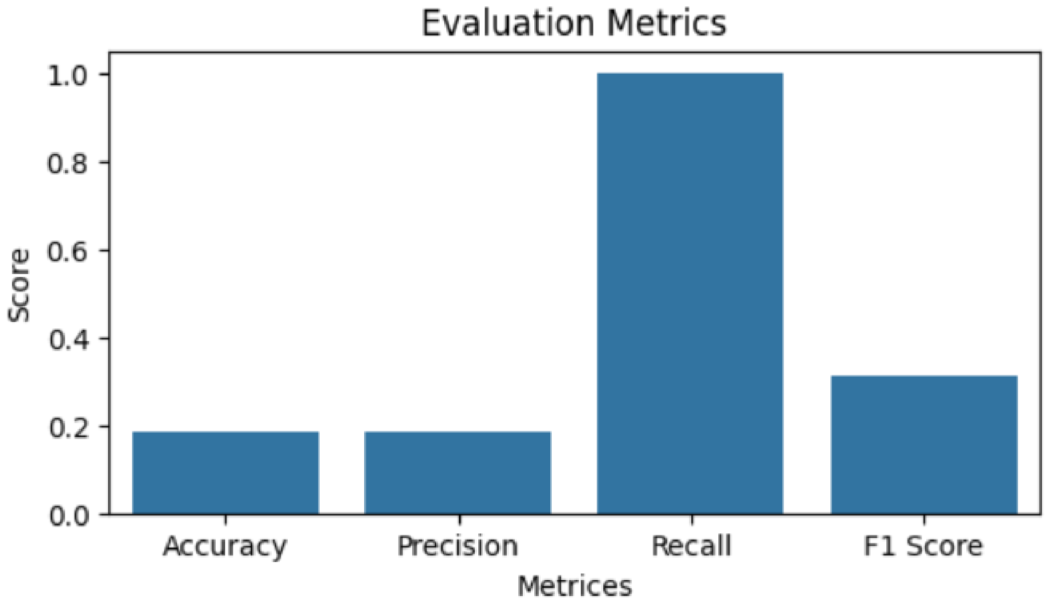


```
robabilty of breast cancer recurrence with top features in forwad selection with feature scaling as input(binary classification):

[1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1]]
onfusion matrix values:
P: 10, FP: 44, TN: 0, FN: 0
```





```
Evaluation Metrics:
Accuracy : 0.18518518518518517
Precision:  0.18518518518484225
Recall   : 0.99999999999
F1 Score : 0.31249999997265626
```

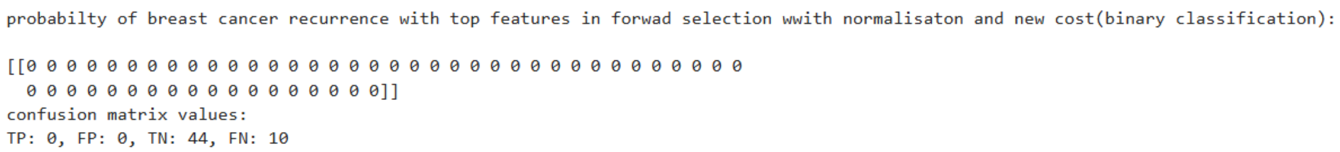


Despite a decline in overall accuracy, feature scaling increased recall from 0 to 1.0, making the model a far more useful tool for identifying cancer recurrence, which is the task's primary goal. In other words, feature scaling did indeed increase performance where it really counts.

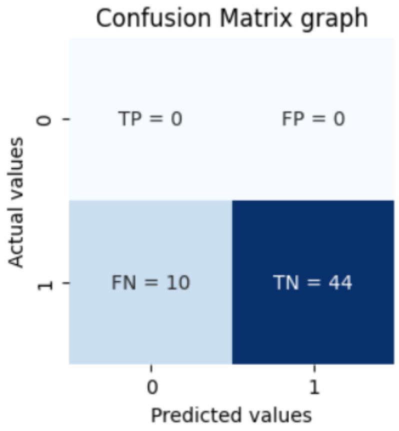
->I wll choose normalisation model among these 2 models based on accuracy.

4B. Appling the new cost function to normalisation

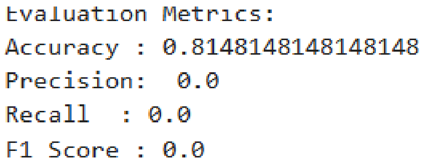
->Below are the confusion matrix for this model with new cost function.



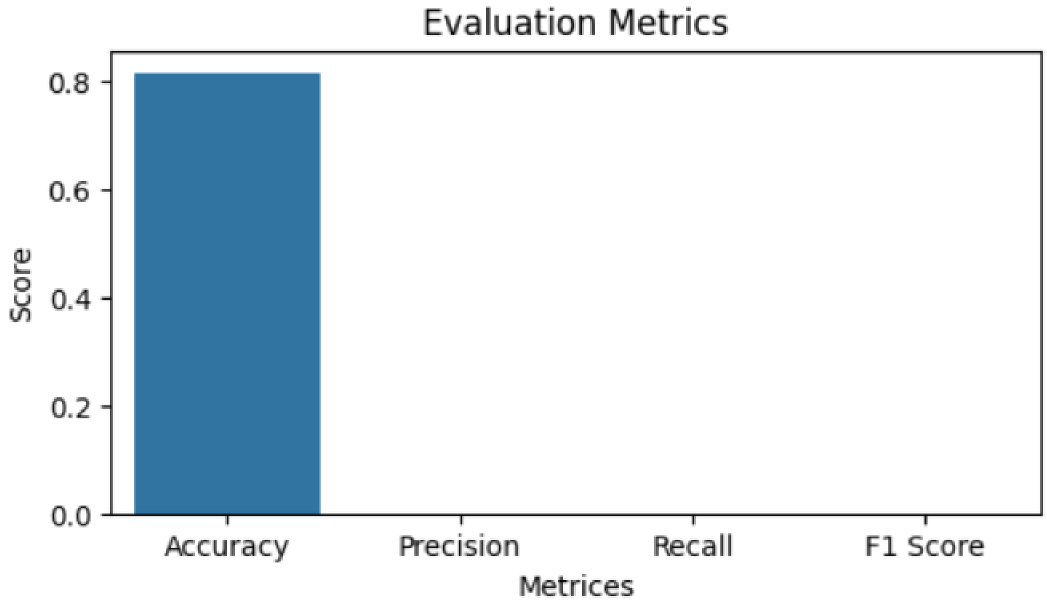
```
probabilty of breast cancer recurrence with top features in forwad selection wwith normalisaton and new cost(binary classification):

[[0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
  0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]]
confusion matrix values:
TP: 0, FP: 0, TN: 44, FN: 10
```



->Below is the Evaluation matrix of this model.



```
Evaluation Metrics:
Accuracy : 0.8148148148148148
Precision:  0.0
Recall   : 0.0
F1 Score : 0.0
```



->Compare the performance of both the models Do they give the same solution with a difference in cost function?

Yes, they gave the same solution with the difference in cost function.