Assignment 2 Report

1. Exploratory Data Analysis **(40 points)**

a) Consider the following numeric variables in the dataset: mean_radius, mean_texture, mean_perimeter, mean_area, mean_smoothness, mean_compactness, mean_concavity and mean_concave_points. Summarize the statistics of these variables into count, mean, standard deviation, minimum, 25% percentile, 50% percentile, 75% percentile, and maximum.

count	**mean**	**std**	**min**	**25%**	**50%**	**75%**	**max**	
mean_radius	198	17.060303	3.974541	0.00000	14.89000	17.195000	19.550000	27.2200
mean_texture	198	21.868131	5.280374	0.00000	19.24750	21.660000	24.517500	39.2800
mean_perimeter	198	114.856566	21.383402	71.90000	98.16000	113.700000	129.650000	182.1000
mean_area	198	970.040909	352.149215	361.60000	702.52500	929.100000	1193.500000	2250.0000
mean_smoothness	198	0.102681	0.012522	0.07497	0.09390	0.101900	0.110975	0.1447
mean_compactness	198	0.142648	0.049898	0.04605	0.11020	0.131750	0.172200	0.3114

count	mean	std	min	25%	50%	75%	max	
mean_concavity	198	0.156243	0.070572	0.02398	0.10685	0.151350	0.200500	0.4268
mean_concave_points	198	0.086776	0.033877	0.02031	0.06367	0.086075	0.103925	0.2012

b) Consider the categorical variable “outcome” in the dataset. Summarize the statistics of variable into count, unique value, top value, and frequency of top value.

Count: 198

Unique: ['N' 'R']

Top value: N

Frequency top value: 151

c) Is there a way to encode outcome variable from categorical to numerical data type? If so, how would you do that?

Yes, we can do label encoding. This type has been choosen instead of others like one hot encoding or frequency encoding or binary because there are only 2 unique values(N and R). The label encoding will map 0 and 1 to the unique values if N is given 0 then R will be 1.

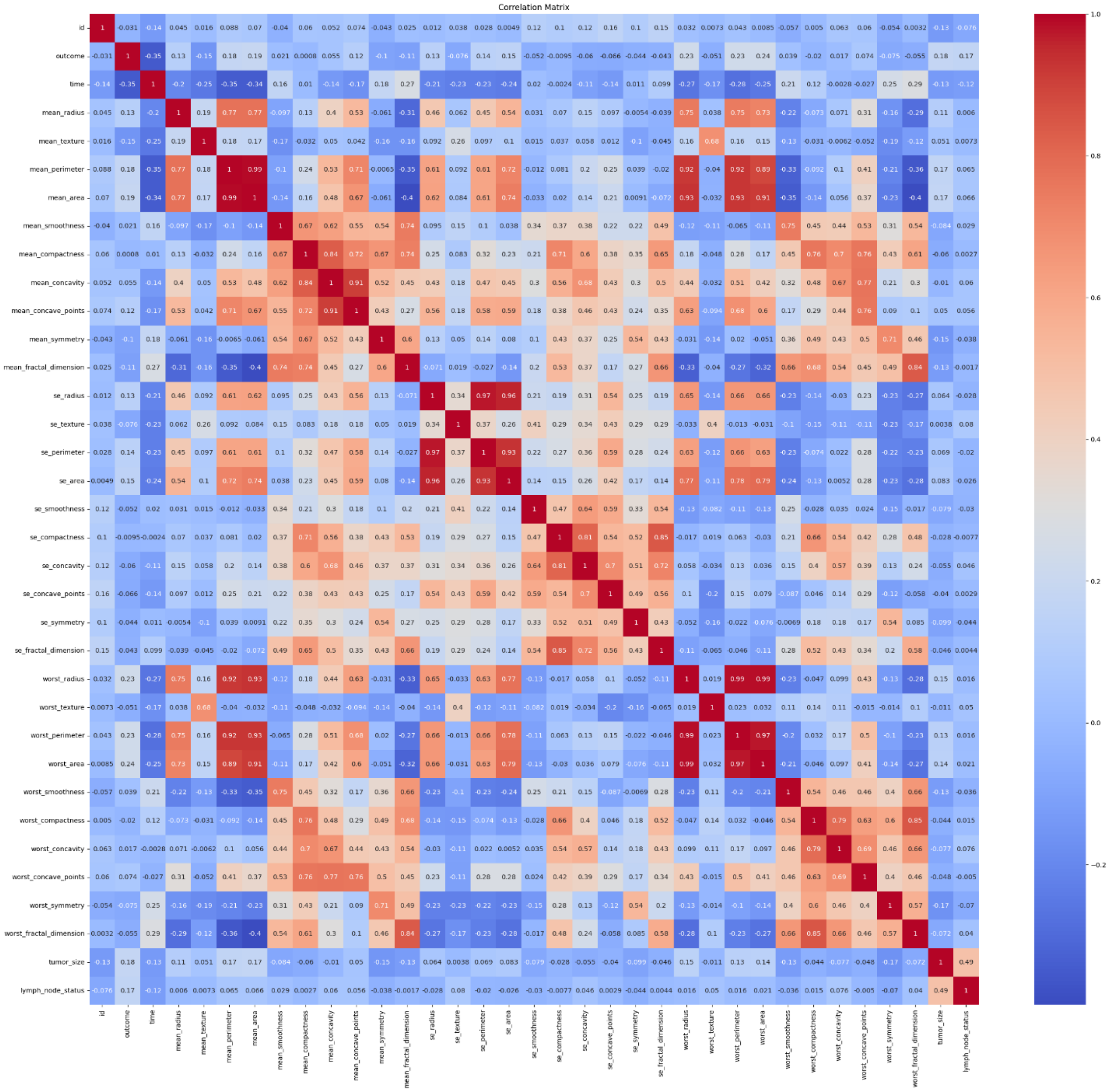
d) Do you think there are any redundant features present in the dataset? If so, explain how removing it won't impact the analysis. Also, based on the experiments so far, were there any interesting observations with respect to the variables?

There are some size related features that are highly correlated. They are:

1. mean_area - mean_perimeter

2. worst_area - worst_perimeter
3. se_perimeter - se_area
4. worst_radius - worst_area and worst_perimeter

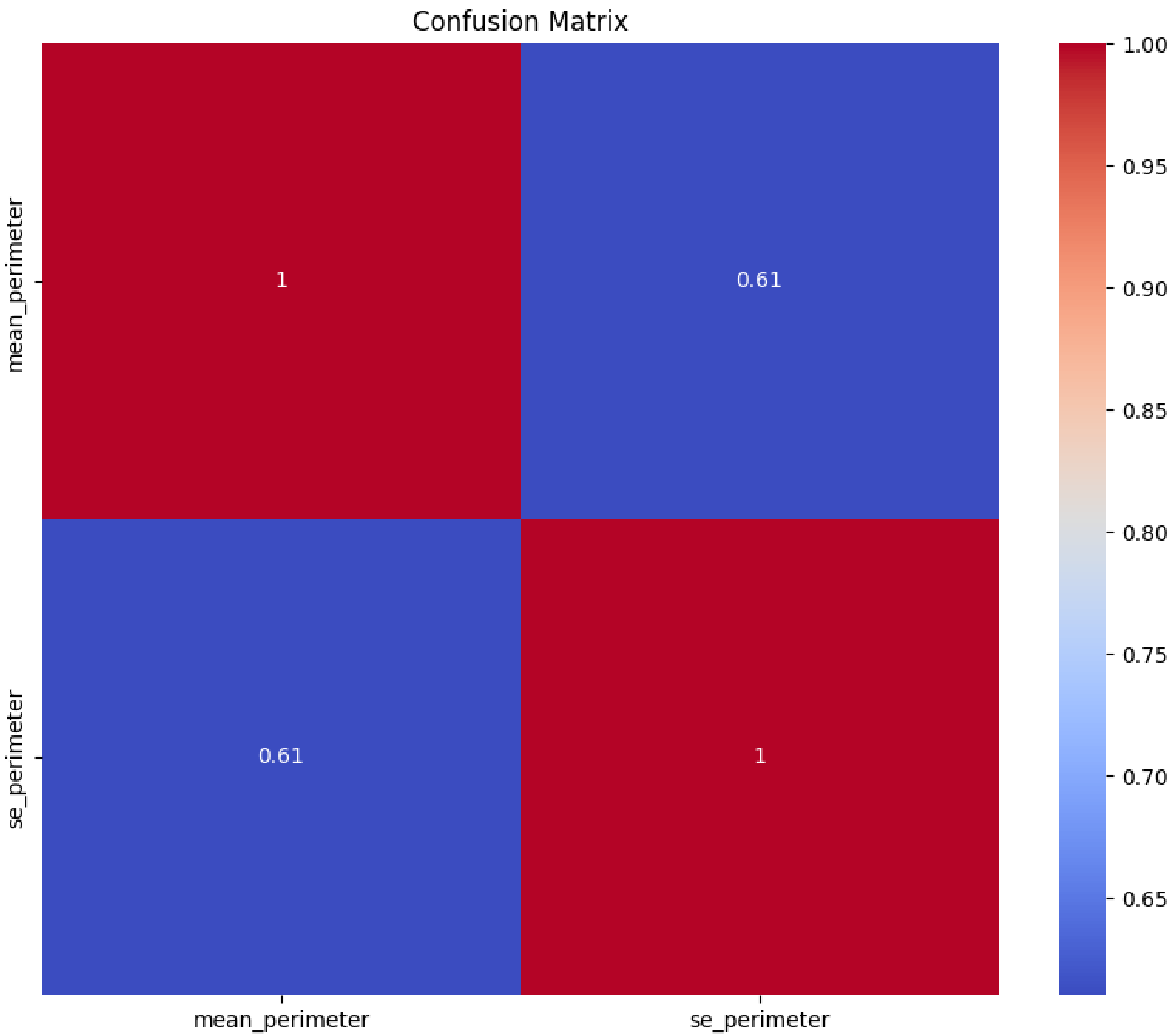
From this observation we can see perimeter and area are obtained using the radius. So removing these might not affect the analysis at a greater extent. For all these the radius and other two are highly corellated in all therr mean, se and worst sizes but i have given the maximum two pairs. If the pairs are area and perimeter if the analyisis focus on the outer or boundary go with perimeter if the analysis on surface area or mass then go with the area.



This is the correlation matrix.

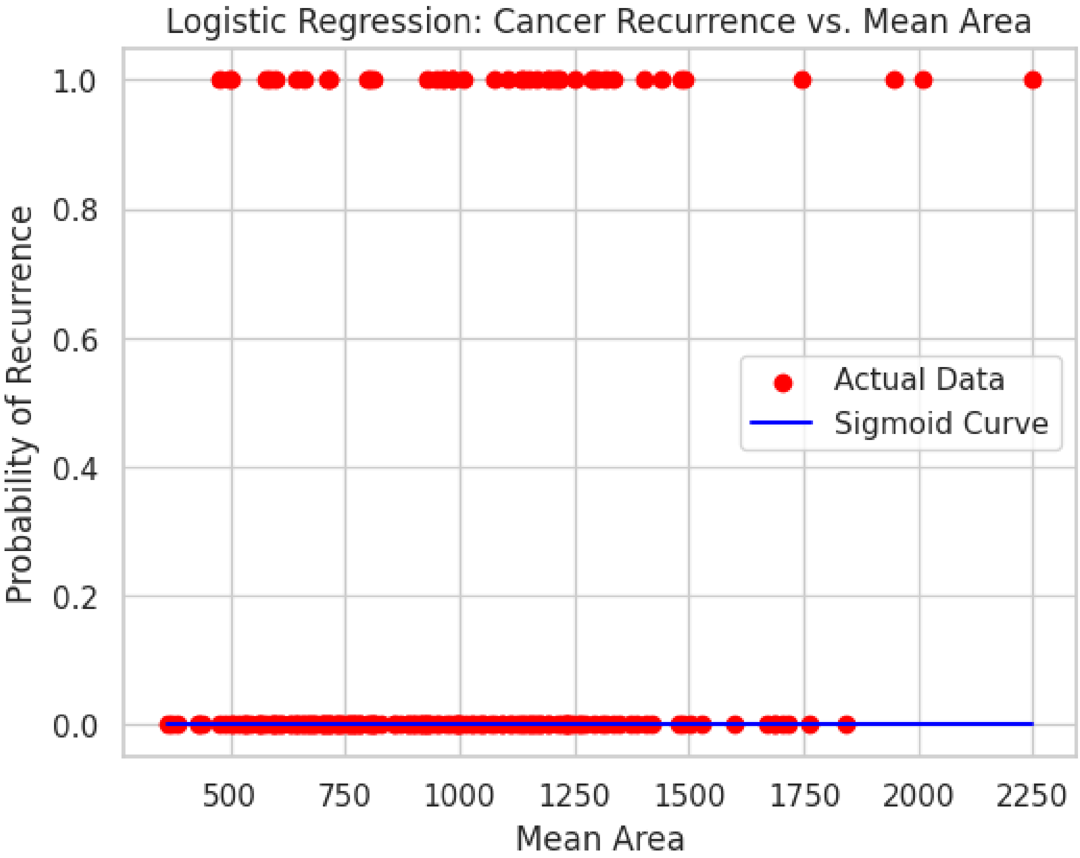
What is the correlation between mean_perimeter and se_perimeter?

If the value is between 0.5 and 0.7 they are indicated as moderaterly corellated but this doesnot affect the analysis.



2. Logistic Regression with One Variable **(20 points)**

a) Can you map the likelihood of breast cancer recurrence (outcome) based on "mean_area" feature from the dataset?



As you can see, here the sigmoid curve is flat and at 0. This might be due to the model not capturing the relationship properly

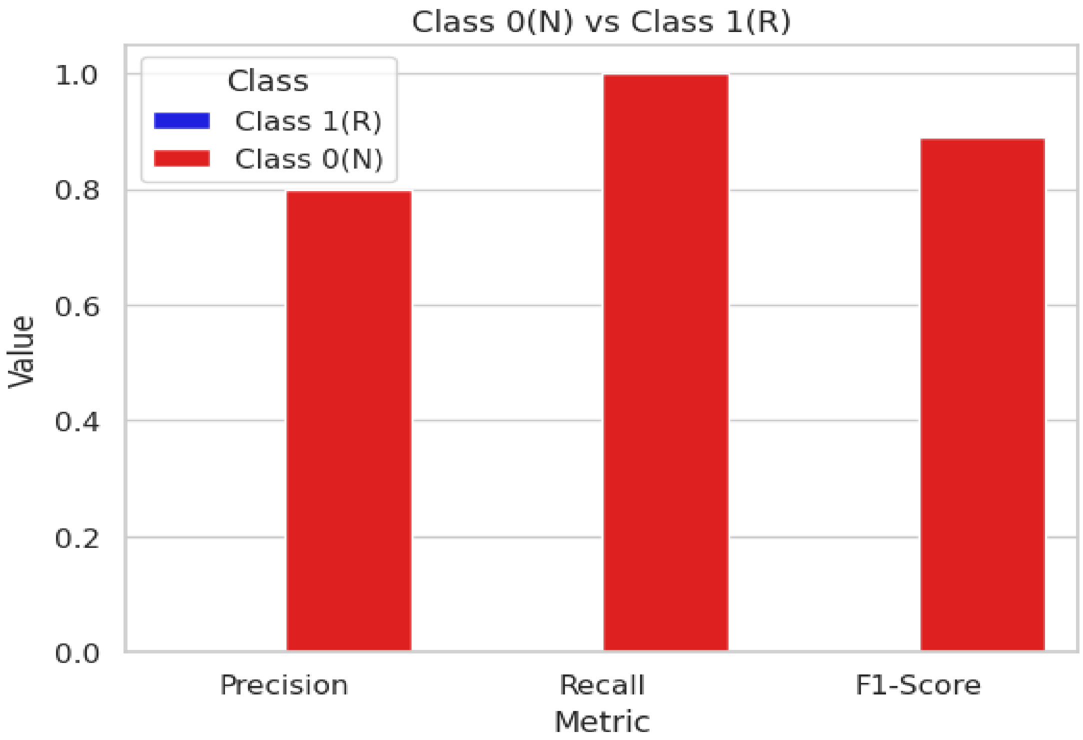
b) Evaluate performance using a metric discussed in class (such as confusion matrix). You may also use graphs to explain your observations.

Confusion Matrix:

TP: 0, TN: 32, FP: 0, FN: 8

Metric Value

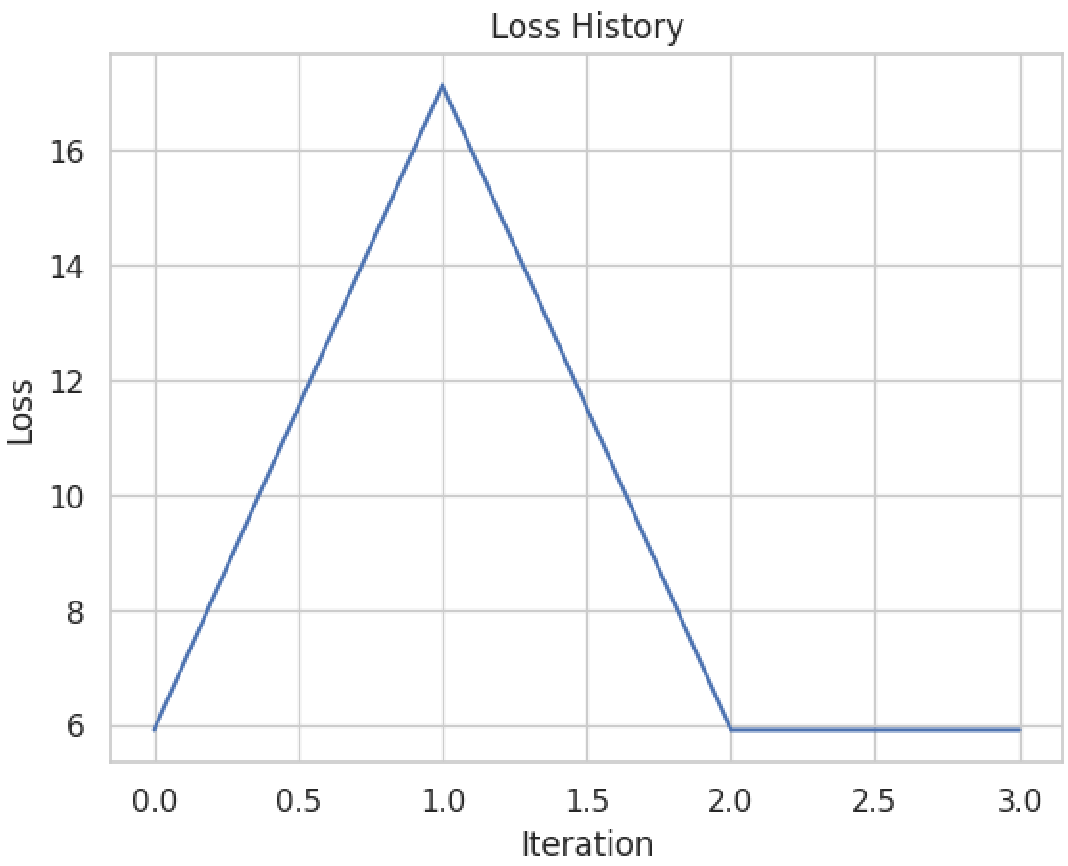
```
0            Precision_1   0.00
1               Recall_1   0.00
2             F1-Score_1   0.00
3            Precision_0   0.80
4               Recall_0   1.00
5             F1-Score_0   0.89
6               Accuracy   0.80
7    Macro Average Precision   0.40
8       Macro Average Recall   0.50
9     Macro Average F1 Score   0.44
10 Weighted Average precision   0.64
11   Weighted average Recall   0.80
12       Weighted Average F1   0.71
```



3. Logistic Regression with Multiple Variables **(50 points)**

a) Design a Logistic Regression model to predict breast cancer recurrence (outcome) using the following 12 variables from the dataset as input features:

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



Classification Report for Logistic Regression Before Regularization and Feature Scaling

Confusion Matrix:

TP: 0, TN: 41, FP: 0, FN: 9

```
                     Metric  Value
0              Precision_1   0.00
1                 Recall_1   0.00
2               F1-Score_1   0.00
3              Precision_0   0.82
4                 Recall_0   1.00
5               F1-Score_0   0.90
6                 Accuracy   0.82
7     Macro Average Precision   0.41
8        Macro Average Recall   0.50
9      Macro Average F1 Score   0.45
10  Weighted Average Precision   0.67
11     Weighted Average Recall   0.82
12         Weighted Average F1   0.74
```

b) Design a Logistic Regression model to predict breast cancer recurrence (outcome) using forward selection to select the most significant variables in the dataset as input features. Which subset of features gave you the best performance? What are your thoughts on these features getting selected? (Use 12 features from 3a as Input features)

```
Confusion Matrix:
 TP: 0, TN: 41, FP: 0, FN: 9
```

ACC value obtained: 0.7364864864864865 which is smaller than the best Acc obtained before so the loop stoped

```
2
mean_radius
mean_texture
```

```
Best training accc: 0.7567567567567568

Test accuracy using forward stepwise regression: 0.82


Classification Report for Logistic Regression Before Regularization and Feature Scaling

                  Metric  Value
0              Precision_1   0.00
1                 Recall_1   0.00
2               F1-Score_1   0.00
3              Precision_0   0.82
4                 Recall_0   1.00
5               F1-Score_0   0.90
6                 Accuracy   0.82
7     Macro Average Precision   0.41
8        Macro Average Recall   0.50
9      Macro Average F1 Score   0.45
10  Weighted Average Precision   0.67
11     Weighted Average Recall   0.82
12         Weighted Average F1   0.74
```

c) Compare the performance of the full model built using all the features in (3a) with the resultant accuracies of the full model using the selected features (3b). Which set of features performed better?

Both gives same accuracy both are good in terms of accuracy that is the metric used mainly for comparison. While working with the stepwise I changed the code to check how well the features work for 12 iteration even if the accuracy is same for ith and i+1th iteration. This gives the result where modt of the iteration the accuracy value remains the same which is 0.82. This might be due to the high correlation(Multi collinearity) between the features.

4. Experimenting with regularization and Cost function. **(40 points)**

   a) Regularization and Feature Scaling: **(20 points)**

      I. For the best performing model in Q.3 (Model from 3c), does regularization improve the performance? II. Does Feature Scaling improve the performance for the model in Q 3c?

After seeing the result both models in Q3 have same accuracy so I choose the Logistic regression model with 12 features.

1. When using just regularization the accuracy didn’t improved it declined.

Test Accuracy: 0.34

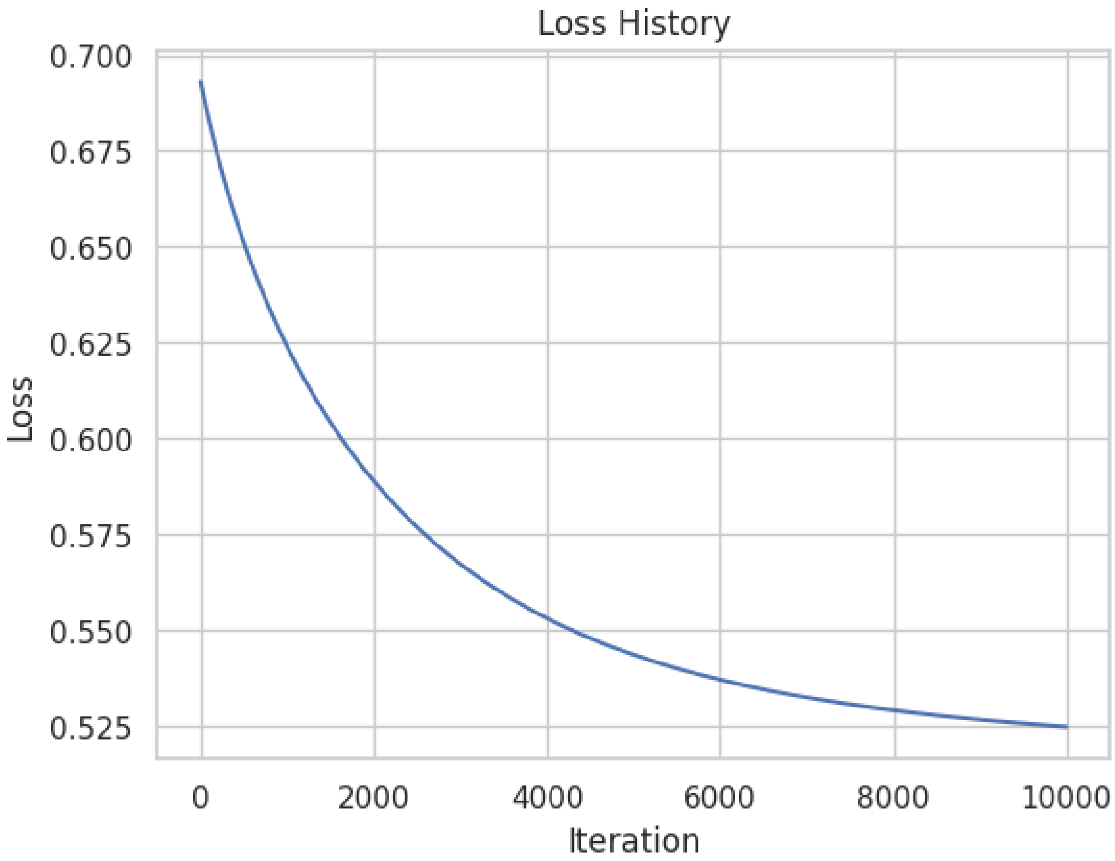
Confusion Matrix:

TP: 9, TN: 8, FP: 33, FN: 0

Classification Report for Logistic Regression after regularization

```
                Metric  Value
0             Precision_1   0.21
1                Recall_1   1.00
2              F1-Score_1   0.35
3             Precision_0   1.00
4                Recall_0   0.20
5              F1-Score_0   0.33
6                Accuracy   0.34
7     Macro Average Precision   0.61
8        Macro Average Recall   0.60
9      Macro Average F1 Score   0.34
10  Weighted Average Precision   0.86
11     Weighted Average Recall   0.34
12         Weighted Average F1   0.33
```

II. When using the feature scaling the accuracy didn’t increase. The test accuracy is 0.84.



But when using them both at the same time the accuracy increased to 0.84

Confusion Matrix:

TP: 1, TN: 41, FP: 0, FN: 8

Classification Report for Logistic Regression after Feature Scaling

```
                   Metric  Value
0             Precision_1   1.00
1                Recall_1   0.11
2              F1-Score_1   0.20
3             Precision_0   0.84
4                Recall_0   1.00
5              F1-Score_0   0.91
6                Accuracy   0.84
7  Macro Average Precision   0.92
```

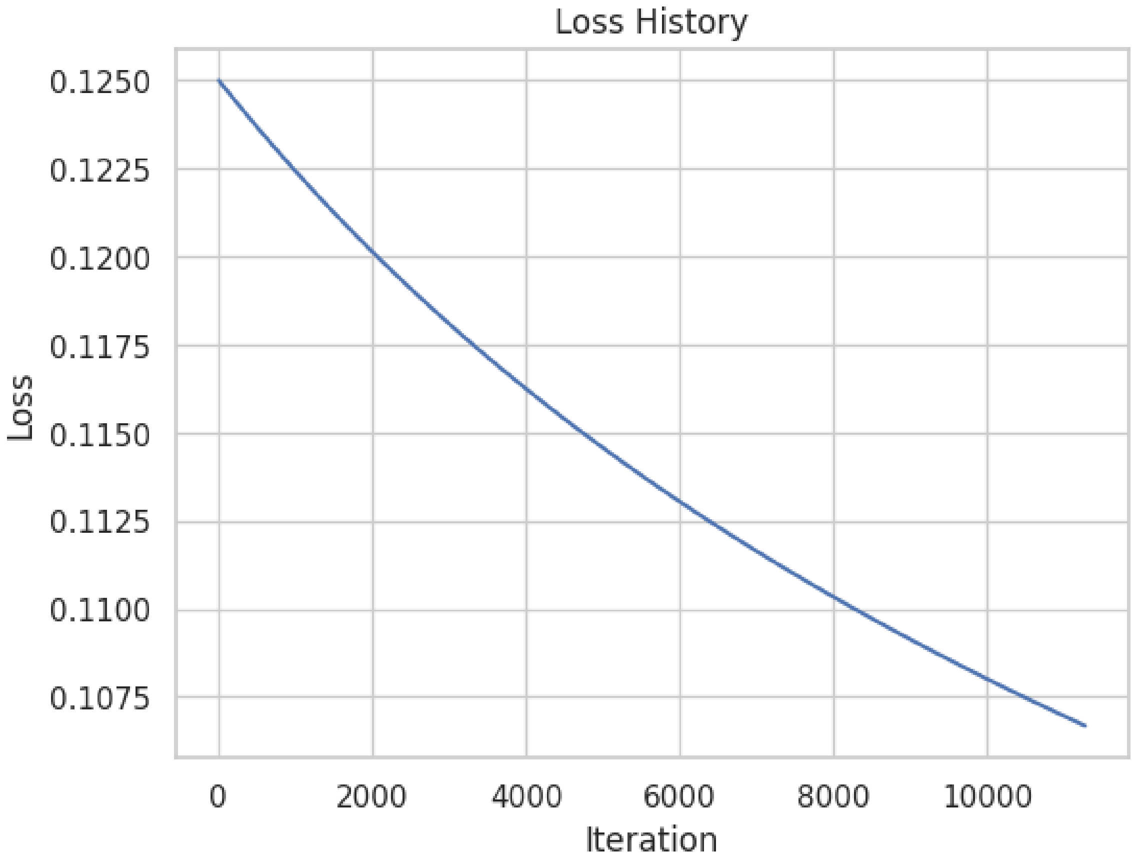
8	Macro Average Recall	0.56
9	Macro Average F1 Score	0.56
10	Weighted Average Precision	0.87
11	Weighted Average Recall	0.84
12	Weighted Average F1	0.78

b) Cost Function: **(20 points)**

I. Keeping the best model after the experiments from Q.4a, design a Logistic Regression model to predict breast cancer recurrence (outcome) by changing the cost function to the following (Mean Squared Error).

$$J(\Theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\Theta(x^{(i)}) - y^{(i)})^2$$

By using the costfunctin asMSE is used to create a logistic regression model with 12 features.



II. Compare the performance of both the models (4.b.i and 4.a.). Do they give the same solution with a difference in cost function?

Yes they gve the different result but using MSE didn’t improve but it reduced the accuracy significantly. This is due to the non convex nature of MSE which means this can give more than one local minima, which will be hard for gradient descent to converge properly. Even with regularization and normalization the Accuracy declined to 0.18.

Classification Report for Logistic Regression after Regularization and Feature Scaling

```
                      Metric  Value
0                Precision_1   0.18
1                   Recall_1   1.00
2                 F1-Score_1   0.31
3                Precision_0   0.00
4                   Recall_0   0.00
5                 F1-Score_0   0.00
6                   Accuracy   0.18
7     Macro Average Precision   0.09
8        Macro Average Recall   0.50
9      Macro Average F1 Score   0.15
10 Weighted Average Precision   0.03
11    Weighted Average Recall   0.18
12        Weighted Average F1   0.05
```