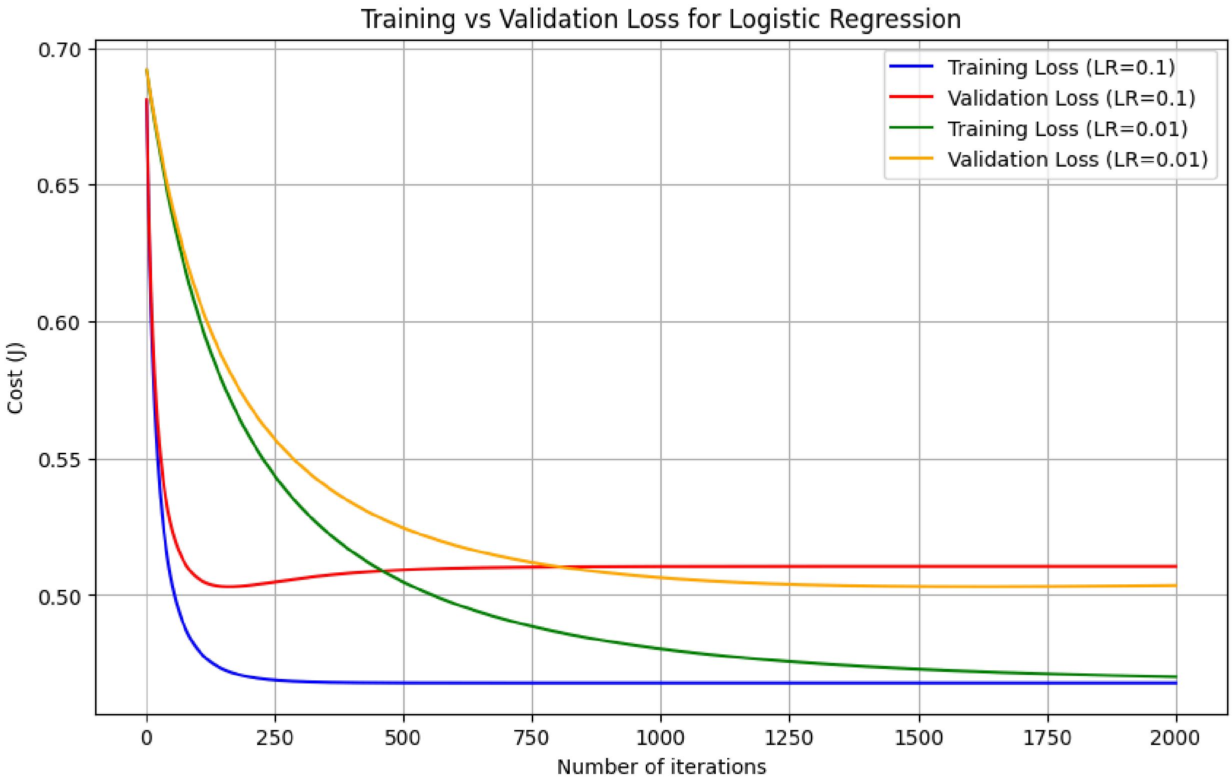
Homework 3 Report
Name: Viprav Lipare
Student ID: 801288922
Homework Number: 3
Github Repository: https://github.com/vipravlipare/ECGR-5105-Intro-to-Machine-Learning

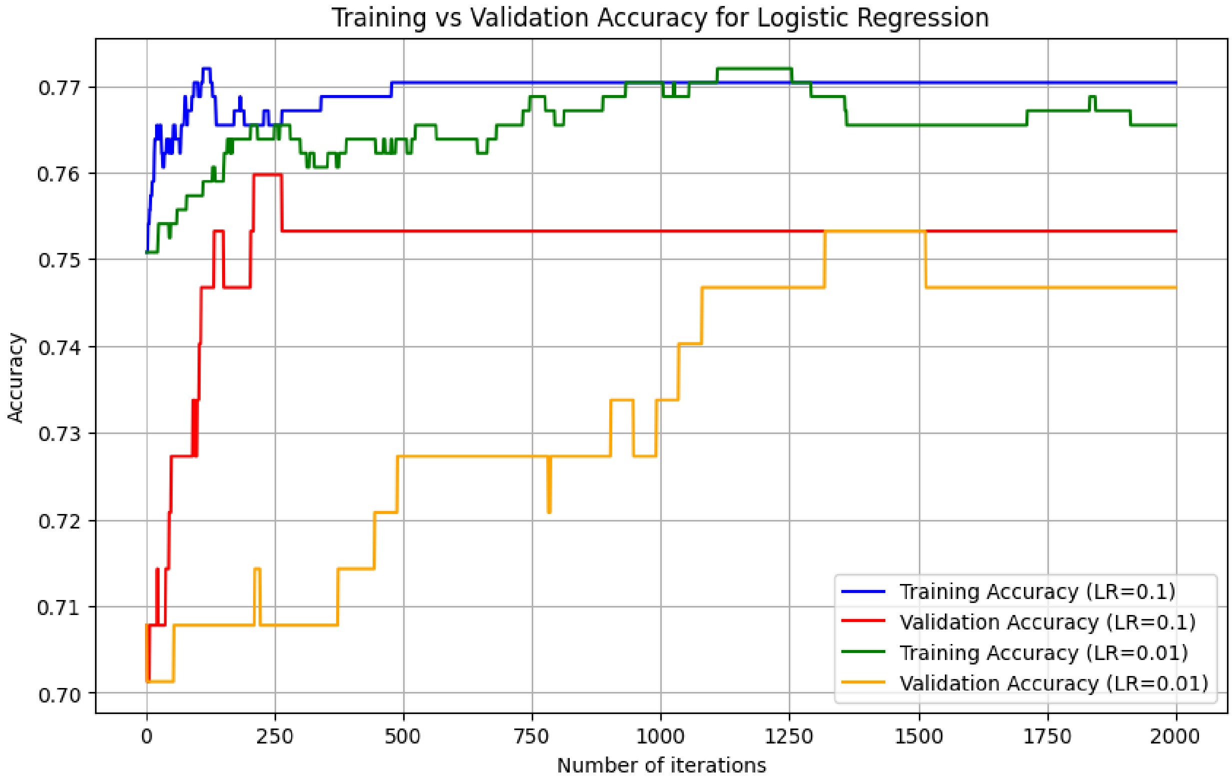
Problem 1

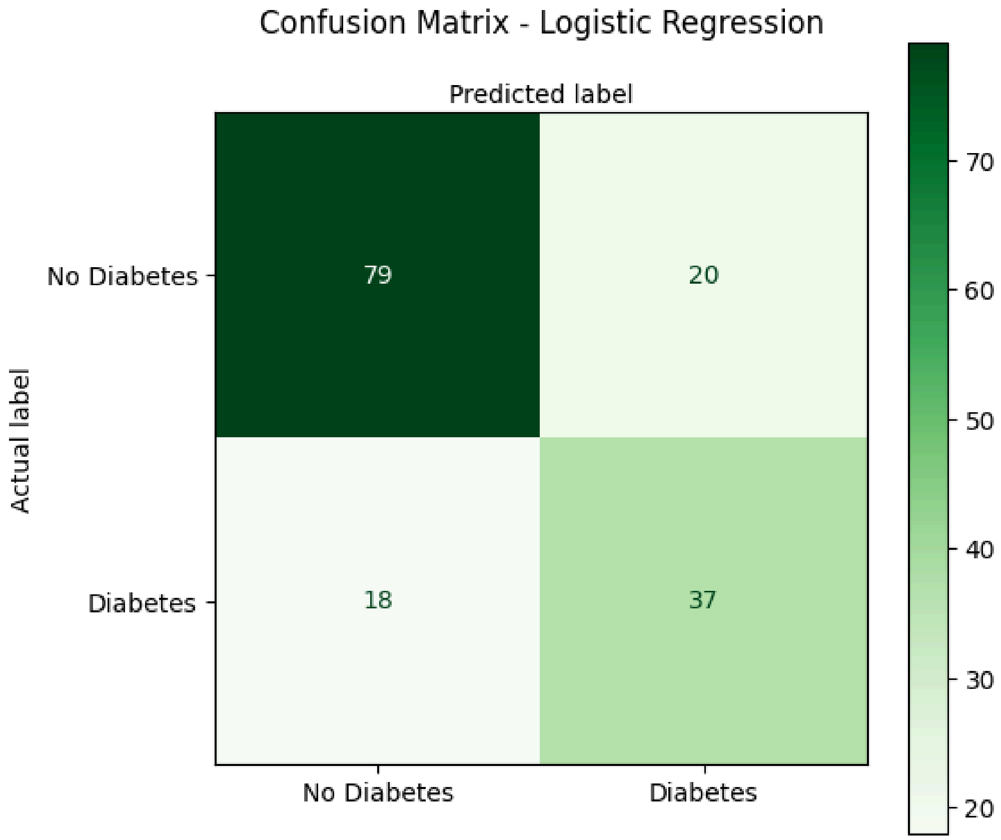
In Problem 1, a logistic regression binary classifier algorithm was implemented to predict the presence of diabetes using the inputs from the diabetes dataset. The dataset was split into 80% training and 20% test sets, and all inputs were normalized. The training and validation losses were plotted, and then the accuracy of the training and validation sets were also plotted. Some results were gathered from the resulting training sets, accuracy, precision, recall, and F1 score. The final step was plotting a confusion matrix to see the final results of the algorithm.

The logistic regression model for problem 1 achieved an accuracy of 75.32%, a precision of 64.91%, a recall of 67.27%, and a F1 score of 0.6671. This means that the model correctly classified approximately three quarters of all cases in the test set, hence the 75.32% accuracy. The model predicted a positive diabetes case approximately two thirds of the times, hence the 64.91% precision. The model correctly predicted a positive diabetes case approximately two thirds of the times, hence the 67.27% recall. The F1 score gives a score based on the overall performance of the model to classify the groups correctly, and the F1 score for problem 1 was 0.6671 meaning an overall grade of approximately 66.71%. The model performs moderately well and classifies whether the patients have diabetes or not.

Results:
Accuracy : 0.7532467532467533
Precision: 0.6491228070175439
Recall   : 0.6727272727272727
F1 Score : 0.6607142857142857


Training vs Validation Loss for Logistic Regression
Training Loss (LR=0.1)
Validation Loss (LR=0.1)
Training Loss (LR=0.01)
Validation Loss (LR=0.01)
Cost (J)
0.70
0.65
0.60
0.55
0.50
0
250
500
750
1000
1250
1500
1750
2000
Number of iterations


Training vs Validation Accuracy for Logistic Regression
Accuracy
0.77
0.76
0.75
0.74
0.73
0.72
0.71
0.70
Training Accuracy (LR=0.1)
Validation Accuracy (LR=0.1)
Training Accuracy (LR=0.01)
Validation Accuracy (LR=0.01)
0
250
500
750
1000
1250
1500
1750
2000
Number of iterations



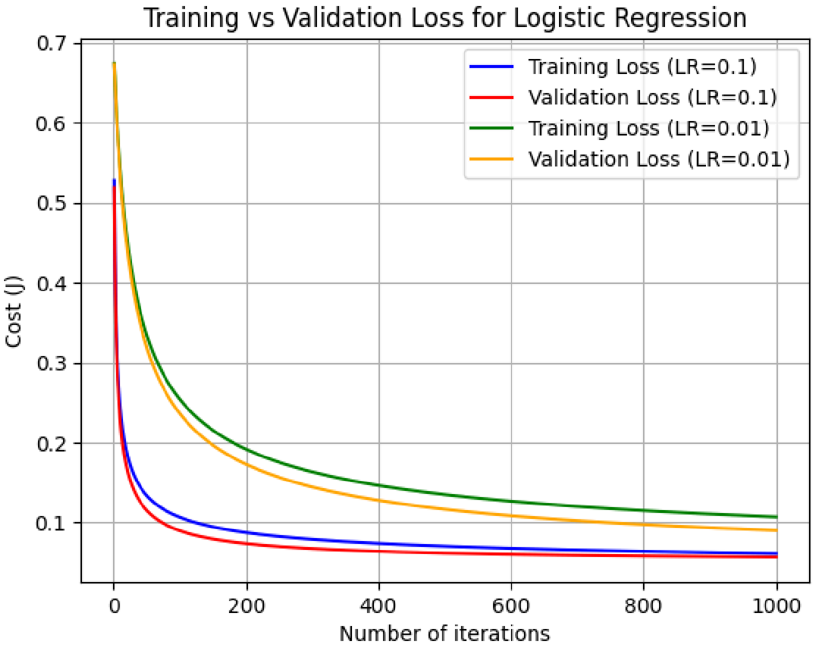
Problem 2A

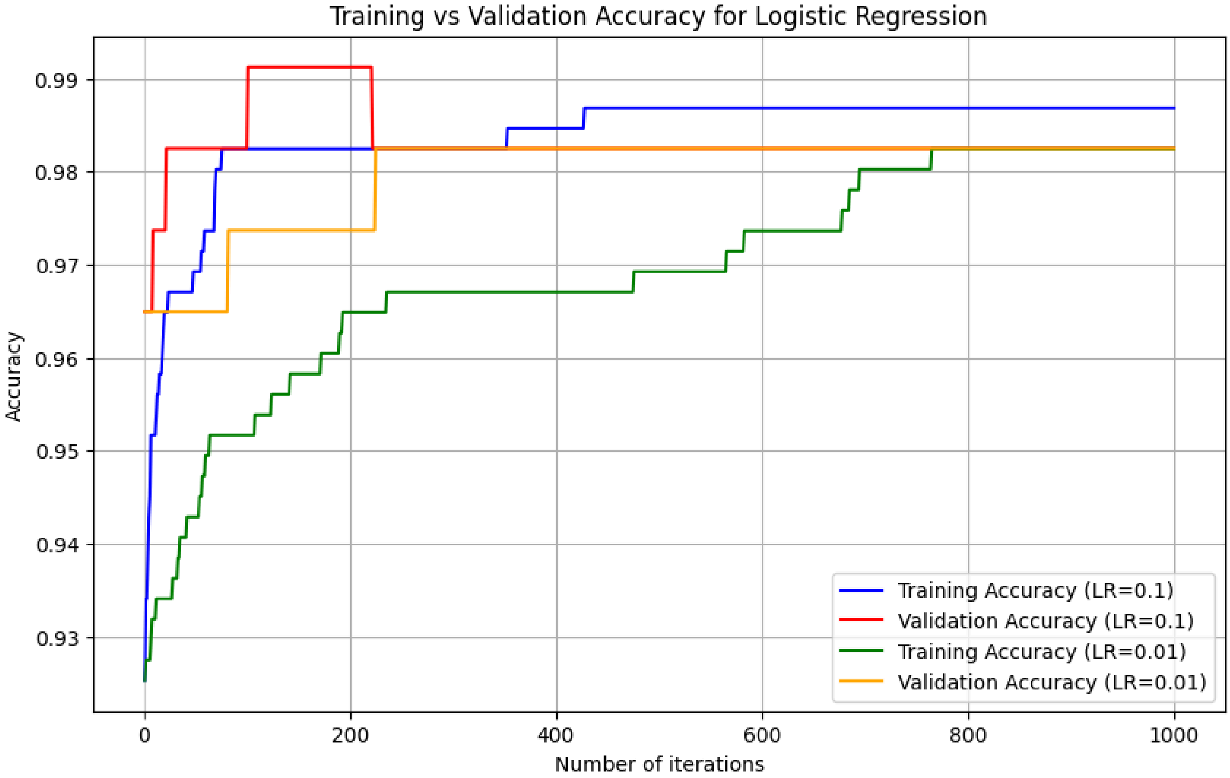
In Problem 2A, a logistic regression binary classifier algorithm was implemented to predict the type of cancer (Malignant vs. Benign) using all input features from the breast cancer dataset. The dataset was split into 80% training and 20% test sets, and all inputs were standardized. The training and validation losses were plotted, and then the accuracy of the training and validation sets were also plotted. Some results were gathered from the resulting training sets, including accuracy, precision, recall, and F1 score. The final step was plotting a confusion matrix to see the final results of the algorithm.

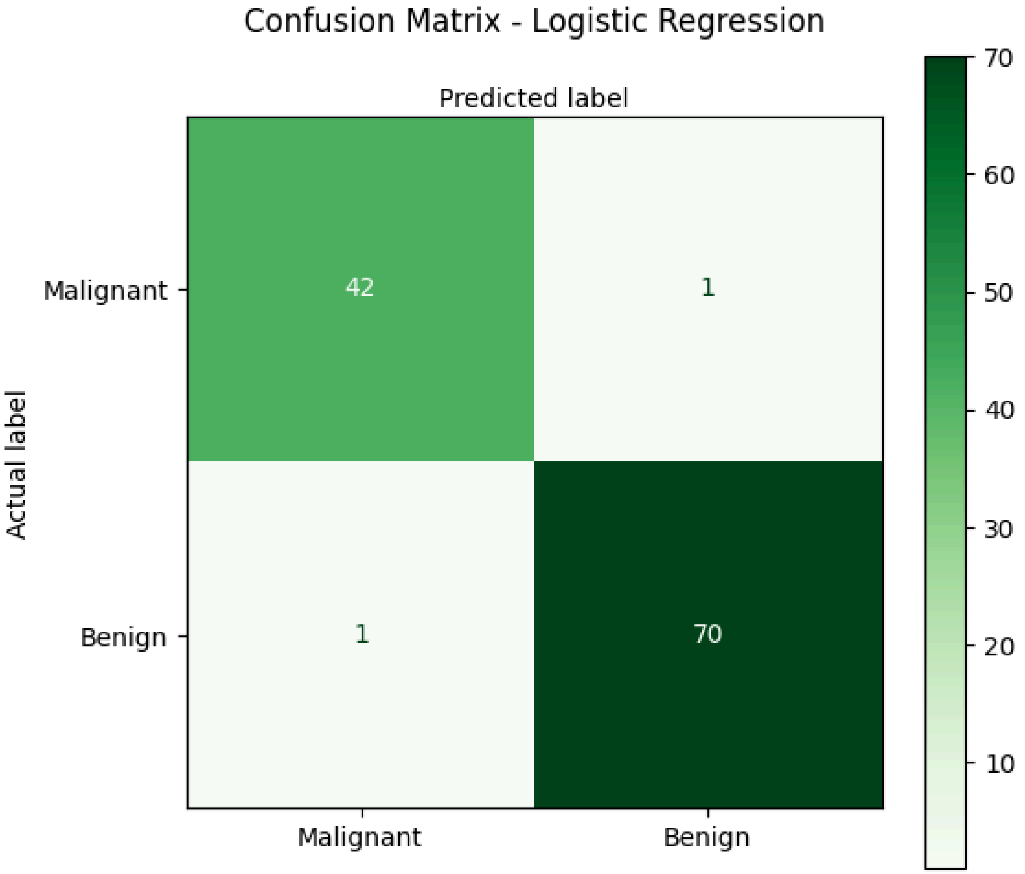
The logistic regression model for Problem 2A achieved an accuracy of 98.24%, a precision of 98.59%, a recall of 98.59%, and a F1 score of 0.9859. This means that the model correctly classified nearly all cases in the test set, hence the 98.24% accuracy. The model predicted a malignant cancer case correctly approximately 98.59% of the times it made a positive prediction, hence the precision. The model correctly identified approximately 98.59% of actual malignant cancer cases, hence the recall. The F1 score gives a score based on the overall performance of the model to classify the groups correctly, and the F1 score for problem 2A was 0.9859 meaning an overall grade of approximately 98.59%. The model performs exceptionally well in distinguishing malignant cancer from benign cancer, correctly identifying most cases.

Results:

Accuracy : 0.9824561403508771
Precision: 0.9859154929577465
Recall   : 0.9859154929577465
F1 Score : 0.9859154929577465







Problem 2B

In Problem 2B, a weight penalty was added to the logistic regression model to account for the number of parameters. The dataset was again split into 80% training and 20% test sets, and all inputs were standardized. The training and validation losses were plotted, along with the accuracy of the training and validation sets. The model was then evaluated using accuracy, precision, recall, and F1 score. The final step was plotting a confusion matrix to see the final results of the algorithm.

The logistic regression model with the weight penalty achieved an accuracy of 98.25%, a precision of 97.26%, a recall of 100%, and an F1 score of 0.9861. This indicates that the model correctly classified nearly all cases in the test set, achieving even higher performance than the previous model. The precision shows that almost all positive predictions for malignant tumors were correct, while the recall of 100% indicates that all actual malignant cases were successfully identified. The F1 score gives a score based on the overall performance of the model to classify the groups correctly, and the F1 score for problem 2A was 0.9861 meaning an overall grade of approximately 98.61%. The addition of the weight penalty improved the classifier’s ability to generalize and detect all malignant cases while maintaining very few false positives.

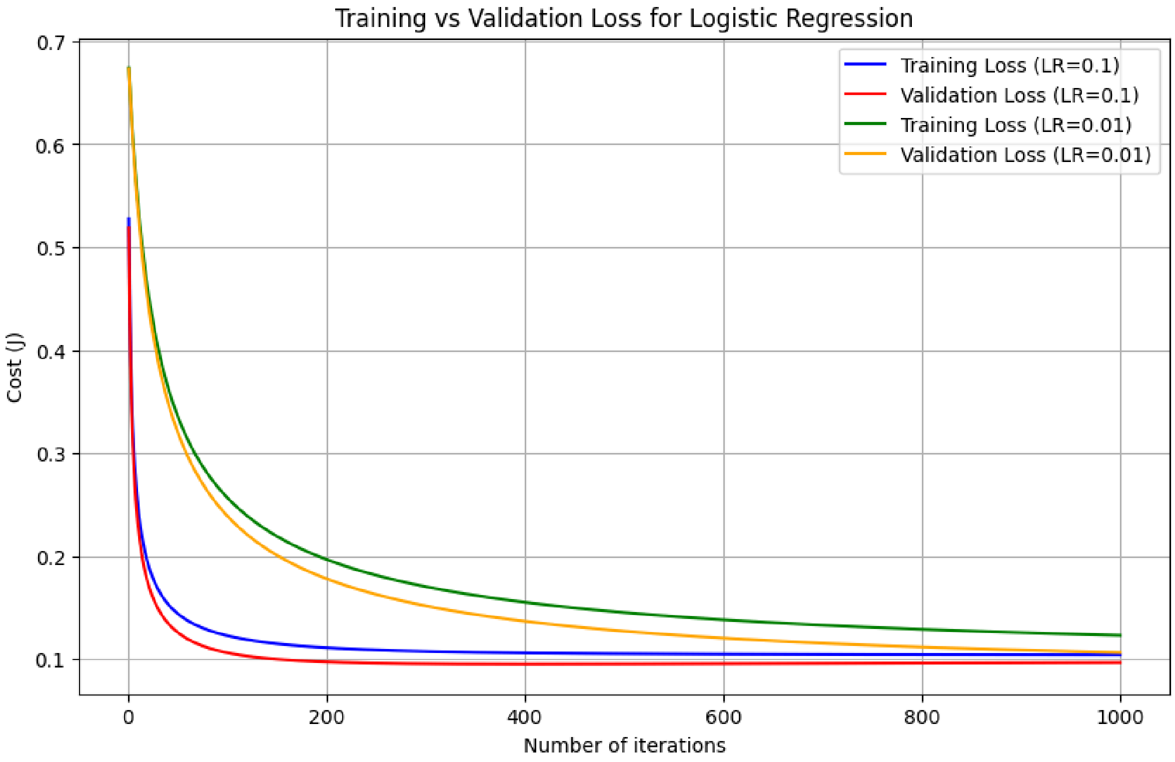
Results:

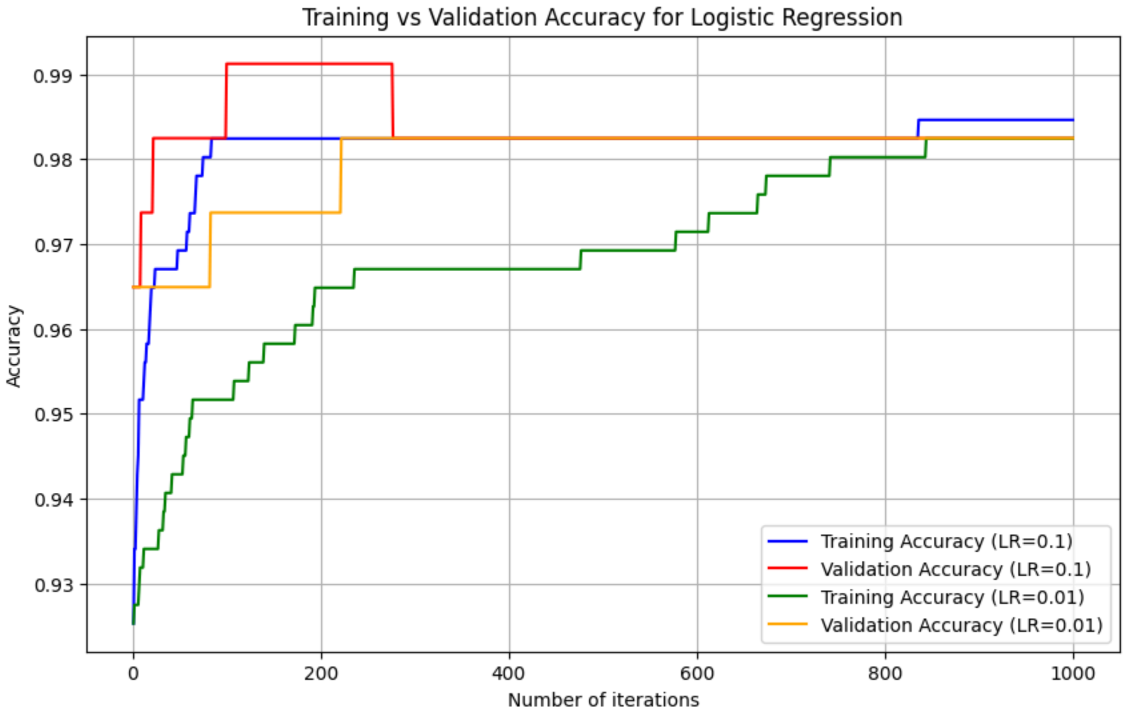
Accuracy : 0.9824561403508771

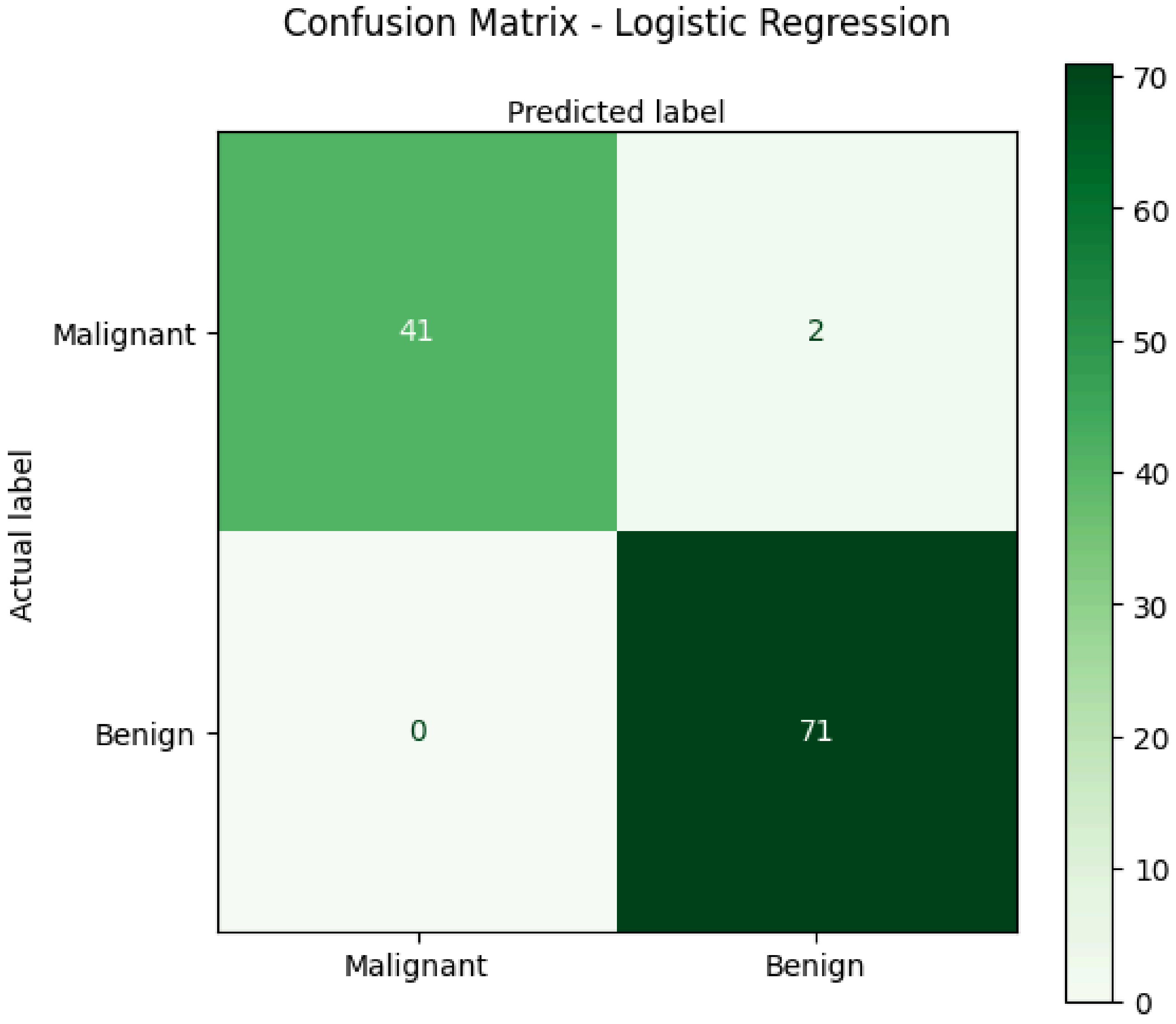
Precision: 0.9726027397260274

Recall : 1.0

F1 Score : 0.9861111111111112






Confusion Matrix - Logistic Regression
Predicted label
Actual label
Malignant
Benign
41
2
0
71
Malignant
Benign
70
60
50
40
30
20
10
0