Homework 3

Intro to ML

Sophia Godfrey
Student ID: 801149485

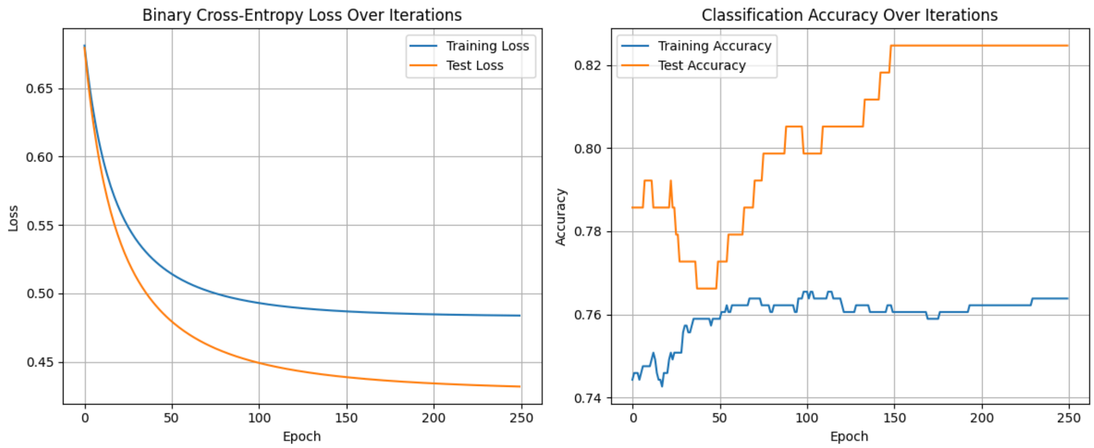
[Github URL Intro-To-ML/Homework 3]

Problem 1

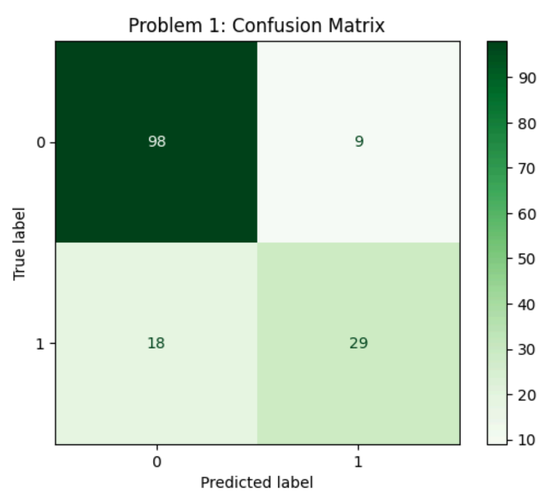
The *diabetes.csv* dataset was loaded and split into training and testing sets (80/20). Features were standardized to have zero mean and unit variance, and a bias term was added to include the intercept. Logistic regression was implemented with a sigmoid activation and binary cross-entropy loss, optimized via gradient descent (learning rate 0.1, 250 epochs). Training and testing losses and accuracies were recorded to monitor performance. The model was evaluated on the test set using accuracy, precision, recall, and F1 score, and a confusion matrix was plotted to visualize classification performance.

Results:

The loss curves showed a steady decrease and stabilization, indicating proper convergence. The accuracy curves for both training and testing increased and leveled off near 0.82, demonstrating consistent performance without significant overfitting.



The confusion matrix showed that the classifier correctly identified a majority of both positive and negative cases, with slightly lower recall indicating that some positive (diabetic) cases were missed.



The logistic regression model achieved good overall accuracy and balanced precision/recall for this dataset, seen below in the table. The moderate recall suggests that while the model is reliable at predicting non-diabetic cases, it may still miss a few true positives. Further improvement could be achieved by feature selection, adding regularization, or experimenting with different learning rates.

Metric	Value
Accuracy	0.8247
Precision	0.7632
Recall	0.6170
F1 Score	0.6824

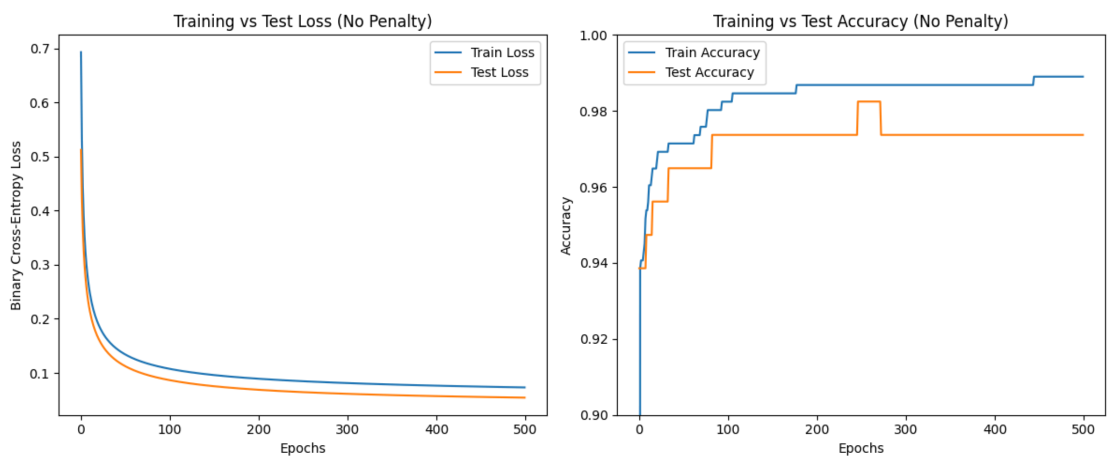
Problem 2

The Breast Cancer dataset from scikit-learn was used to build a logistic regression classifier for distinguishing malignant and benign tumors. It contains 30 numerical features describing tumor characteristics and a binary target variable. Like problem 1, the data was split 80/20 into training and testing sets. All features were standardized to zero mean and unit variance, and a bias term was added to include the intercept. Logistic regression was implemented with a sigmoid activation and binary cross-entropy loss, trained via gradient descent (learning rate 0.1, 500 epochs). Training and testing losses and accuracies were recorded to monitor convergence and evaluate generalization performance.

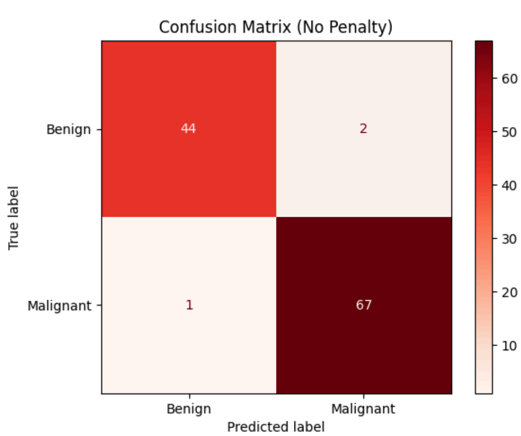
Part (a)

Logistic Regression without Penalty

In the first part, the model was trained without any regularization (i.e., $\lambda = 0$). The resulting loss and accuracy were plotted over iterations for both the training and test sets.



After training, predictions on the test set were compared with the true labels to compute accuracy, precision, recall, and F1 score. A confusion matrix was also generated to visualize the model's classification performance.

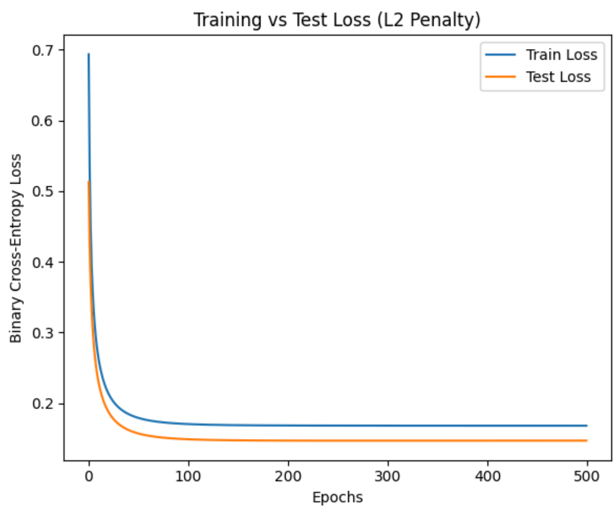


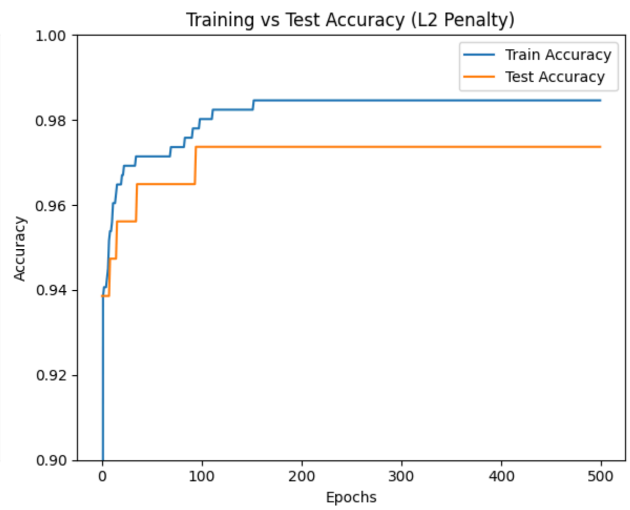
Metric	Value
Accuracy	0.9737
Precision	0.9710
Recall	0.9853
F1 Score	0.9781

Part (b)

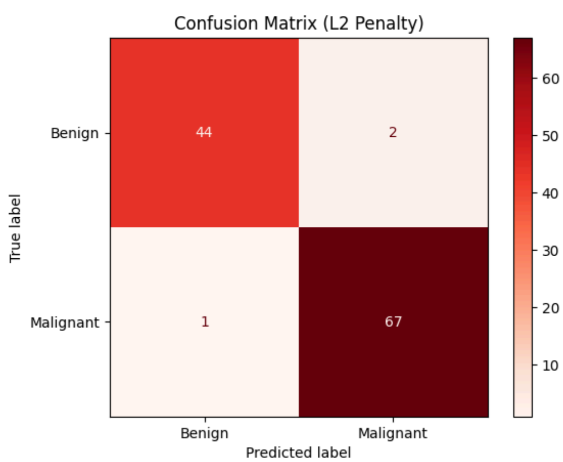
Logistic Regression with L2 Regularization

To evaluate the effect of regularization, the logistic regression model was retrained with an L2 penalty term (λ = 0.05). The weight penalty helped reduce overfitting by constraining parameter magnitudes. Similar to Part (a), training and testing losses and accuracies were plotted, and performance metrics—including accuracy, precision, recall, F1 score, and confusion matrix—were calculated and compared against the unregularized model.





Adding the L2 penalty produced nearly identical results to the unregularized model. This happened because the penalty value (λ = 0.05) was small and the dataset is already well-separated, meaning the model performed well even without regularization. The L2 penalty mainly helped control the size of the weights but did not significantly change the accuracy or other metrics.



Metric	Value
Accuracy	0.9737
Precision	0.9710
Recall	0.9853
F1 Score	0.9781

The logistic regression models for both ‘No Penalty’ and ‘L2 Penalty’ achieved high overall accuracy (97%), as well as strong precision (97%) and recall (99%) for this dataset, as shown in the table above. The results indicate that the model reliably distinguishes between malignant and benign tumors.