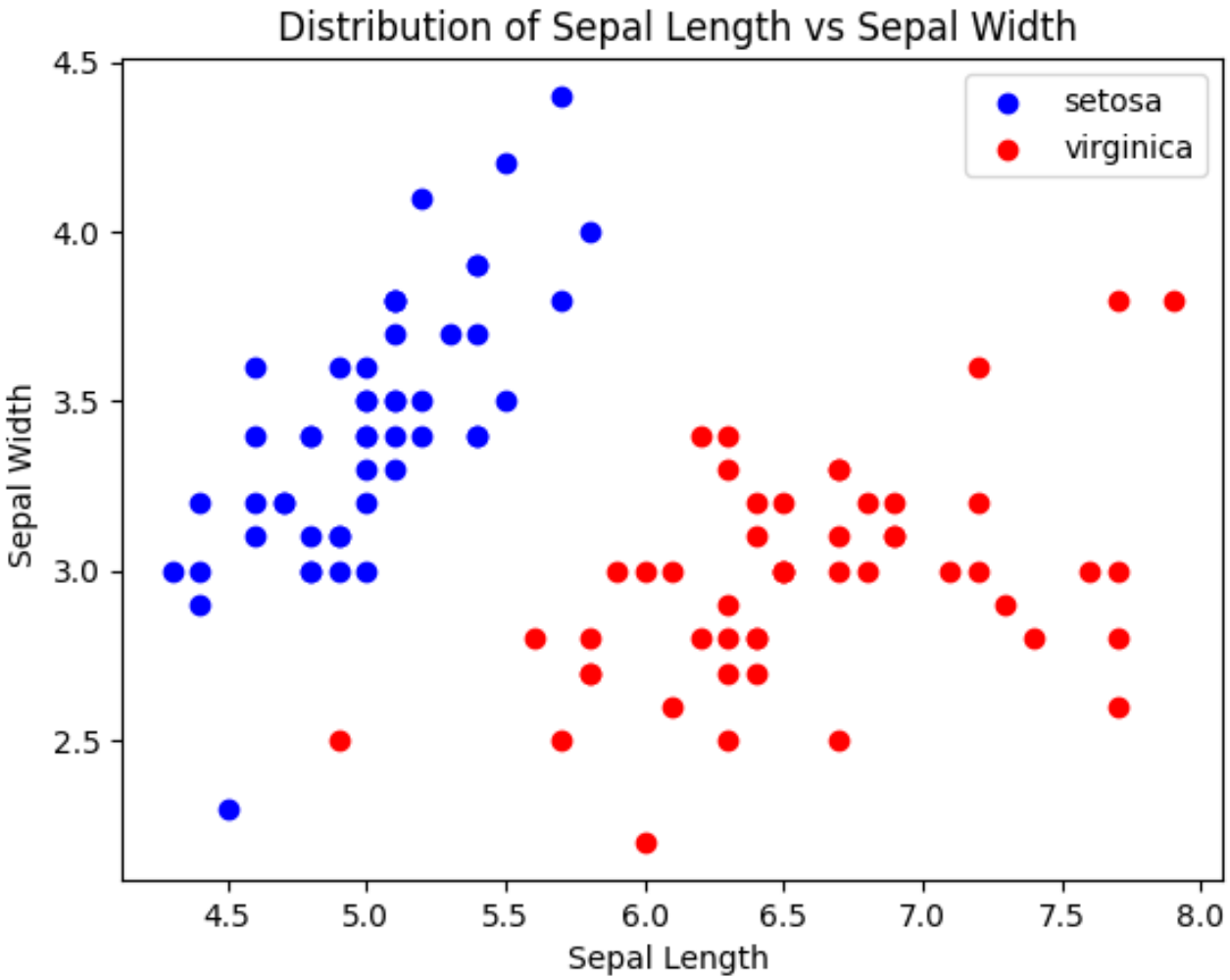
Assignment-2 Report

1.Segregate the data of my choice and plot its distribution.

- I choose iris-setosa and iris-virginica as the label
- Sepal length and sepal width as the input features



2. 80:20 train test split

- I split the data %80 train and %20 as split using sklearn train_test_spilt function both separated 50:50 randomly

```
Training dataset:
Num of setosa samples: 40
Num of virginica samples: 40
Total data in training set: 80
Test Set:
Num of setosa samples: 10
Num of virginica samples: 10
Total data in test set: 20
```

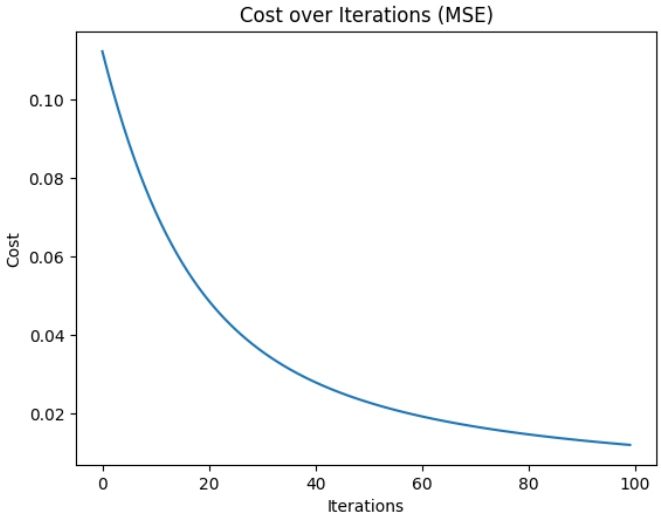
3. Logistic regression model accuracy and cost of the models through iterations for my models

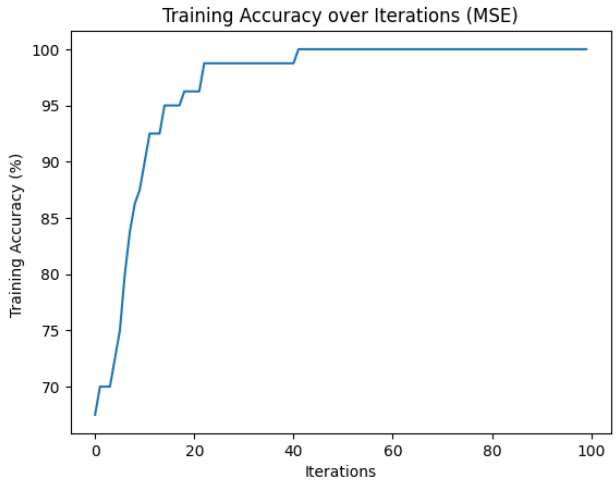
- I used two different cost functions one is log-loss the other is MSE
- For the log losses logistic regression the weights are: final weights: [2.07841092 -0.73141164]

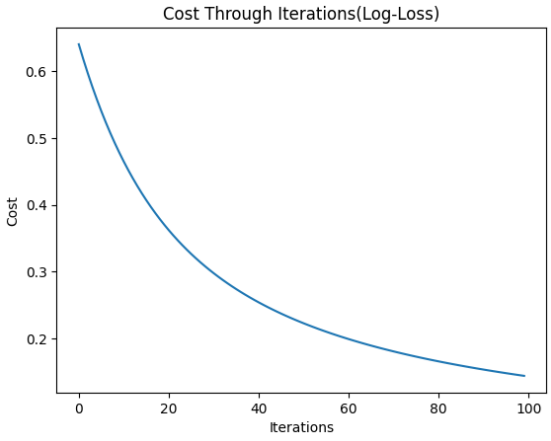
final bias: 0.07620664959869698
final accuracy: % 100.0
final cost: 0.14395353918143203

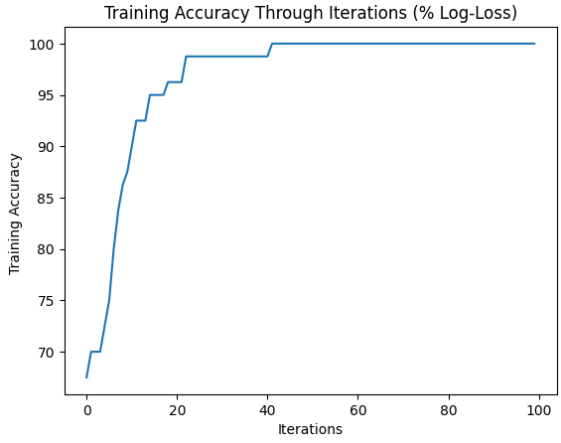
- For the MSE my weight are: Final weights: [2.07841092 -0.73141164]

Final bias: 0.07620664959869698
final accuracy: % 100.0
final cost: 0.011958838540524909









4.Model testing and accuracy for test data

- After training logistic regression model for both MSE cost and Log-loss and both had %100 training accuracy but both performed %95 accuracy in the unseen data
- The test data wasn't as good as training data but it still performs really well and didn't overfit or underfoot

```
test accuracy (Log-Loss) is: % 95.0
test accuracy using OldLogisticRegression (MSE) is: % 95.0
```

5. Summary

In my model i applied the selction of the input features and classes for prediciton as sepal length and sepal width and for lables as virginica and setosa. I looked at the distirbution of the data and it look like linearly seperetable but two points are close to each other so it could effect the training or test accuracy it went to the test set and thats why it couldnt seperated well thus it make the test accuracy as % 95.

I used normalization even though the data is not that big for better model result. For linear regression I applied sigmoid function and applied inside the logistic regression to classify the data. I plotted the cost through iterations to see if the model is working as it should be and plotted accuracy to see if the models performance is good.

For cost function and update the parameters I used log loss for LogisticRegression model and OldLogisticRegression model I used MSE and see their performance. They both performed well in this case but for performance log loss was better computationally as O(N). Also to compute the gradient of the log loss, I differentiated the log loss function with respect to each weight, focusing on how changes in each weight impact the prediction error. By transposing the feature matrix, I aligned it with the error term, enabling efficient calculation of the gradient for all weights at once, which allowed for simultaneous updates during each iteration.

for the log lossed logistic regression the cost weights bias and accuracy are:
final weights: [3.73976568 -1.44467846]
final bias: 0.33752877424124866
final accuracy: % 100.0
final cost: 0.04657999054823778

for the MSE logistic regression the cost weights bias and accuracy are:
Final weights: [2.07841092 -0.73141164]
Final bias: 0.07620664959869698
final accuracy: % 100.0
final cost: 0.011958838540524909