PART 1

REPORT

The logistic regression model built from scratch has been trained on the **"penguins' dataset".** This is considered as a binary classification problem and the target variable/column we chose from the dataset is the sex of the penguin. Our Best model(model_2) predicted the sex of the penguin with an accuracy of 90.9%. The values of the hyperparameters (i.e., maximum number of iterations and the learning rate) set to achieve this accuracy are 10000 and 0.1.

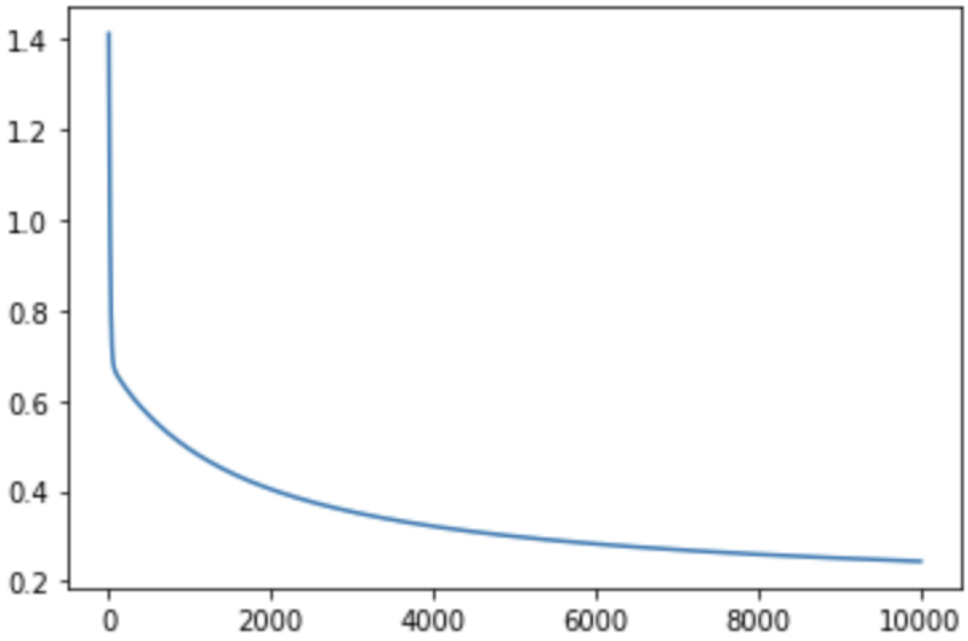

Fig: Loss graph (Test accuracy = 90.91%, iterations = 10000, learning rate = 0.1)

Analysis: Testing accuracy is close to training accuracy and can capture majority of the Variance in the data. The higher learning rate enabled the model to learn at a faster rate and was a ble to output highest accuracy among all the models. From the above plot, it can be concluded that ther e has been a rapid drop in the loss value obtained in each iteration in between 0-20000th iteration.

After that, there's been no significant change in the loss value at every iteration. The lowest loss f or this model is 0.244376.

Influence of hyperparameters on the accuracy of the model:

In order to explain how hyperparameters influence the accuracy of the model, six different models were considered with different combinations of learning rate and maximum iterations. The plots and inference to each setup is given below.

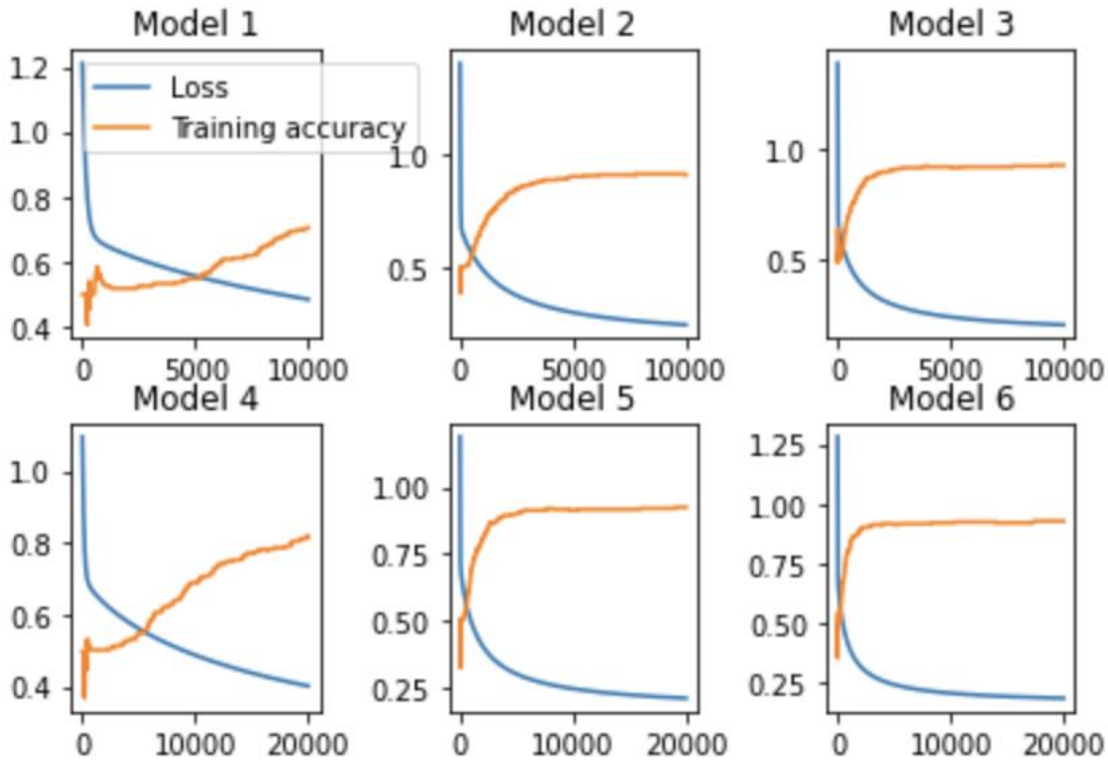

Fig: Loss and Accuracy plots for the logistic regression model for different values of iterations and learning rate.

From the above plots, the following conclusions can be made:

1. In models 1 and 4, the model **took more iterations to make higher quantity of correct predictions than incorrect predictions.** This is because of the lower learning rate (i.e;0.01).

2.The remaining models (2,3,5 and 6) the models were able to make **higher quantity of correct predictions than incorrect predictions at an early stage compared to models 1 and 3**. This is because of the higher learning rate greater than 0.01.

3. This means models with learning rate 0.01 took more iterations than the models with learning rate > 0.01 for the accuracy to become higher than the loss. This is relatable as with a higher learning rate, the model makes bigger leaps during gradient descent and approaches global minimum value in the initial iterations.

	Training Accuracy	Testing Accuracy	LOSS
0	0.704120	0.818182	0.483200
1	0.913858	0.909091	0.244376
2	0.925094	0.893939	0.208373
3	0.820225	0.878788	0.401342
4	0.925094	0.893939	0.208741
5	0.928839	0.878788	0.187672

Fig: Table consisting of testing accuracy, training accuracy and loss of the 6 models

The following inferences can be made from the above table:

1.**Model 1(max_iterations=10000, learning_rate=0.01):** The training accuracy is lower than the testing accuracy. This means that the model is underfitting. The model could not capture all the relationships within the data.

2. **Model 2(max_iterations=10000, learning_rate=0.1):** Testing accuracy is close to training accuracy and the model is able to capture majority of the variance within the data. This change is caused by the increased learning rate. The higher learning rate enabled the model to learn at a faster rate and was able to output a higher accuracy than model 1.

3. **Model 3(max_iterations=10000, learning_rate=0.2):** The testing accuracy is almost same as model 2 but with increased learning rate of 0.2 the model managed to reduce the loss even further to 0.20

4. **Model 4(max_iterations=20000, learning_rate=0.01):** Model 3 is like model 1 in terms of model performance. Despite increasing the iterations, it did not output any significant difference because of the low learning rate (i.e., 0.01)

5. **Model 5(max_iterations=20000, learning_rate=0.1):** model 4 performed similar model 2. Despite running the model for more iterations, the testing accuracy failed to improve. It is observed that the loss value got decreased (i.e., lower than the loss value obtained for model 2).

6. **Model 6(max_iterations=20000, learning_rate=0.2)**: With increased iterations and higher learning rate than model2, the model 5 gave a higher training accuracy but lower testing accuracy which is the most important thing when evaluating a model. But we got the lowest loss value for model 5 among all the five models.

Conclusion 1: Increasing the learning rate keeping the Iterations constant helped to decrease the loss value.

Conclusion 2: The model 2 can be considered as the best since it has good accuracy and low loss value.

Benefits and drawbacks of Logistic Regression:

Benefits:

There are a plenty of real-world scenarios where the value that is to be predicted is not continuous. In such a case, logistic regression works well compared to linear regression. In the dataset provided to us, we were asked to predict either the species that a penguin belongs to(multi-class) or the sex of the penguin(binary). Logistic regression performs better than linear regression for such classification problems as the outcome of the logistic regression is probabilistic rather than continuous. Another advantage of a logistic regression model is that is it sensitive to imbalance data as the regression line is a sigmoid curve.

Drawbacks:

Logistic regression assumes that all the features in the dataset are independent of each other. The model does not work well if any of the features are dependent on any of the other features. Real world problems are mostly non-linear in nature and a logistic regression model does not work well with non-linear data.

CHECKPOINT CONTRIBUTION:

Team Member	Assignment part	Contribution
Manish Chava	Steps 1 – 8 (creating data matrix, splitting the data) Steps 9-12 (gradient descent, fit, predict functions, training the model)	55%
Sriinitha Chinnapatlola	Steps 1 – 8 (preprocessing, normalizing) Steps 9 – 12 (sigmoid, cost, predict functions)	45%