Name-Nitesh Jaiswal
Roll No-2018400
Branch-CSB

Assignment -1 Report

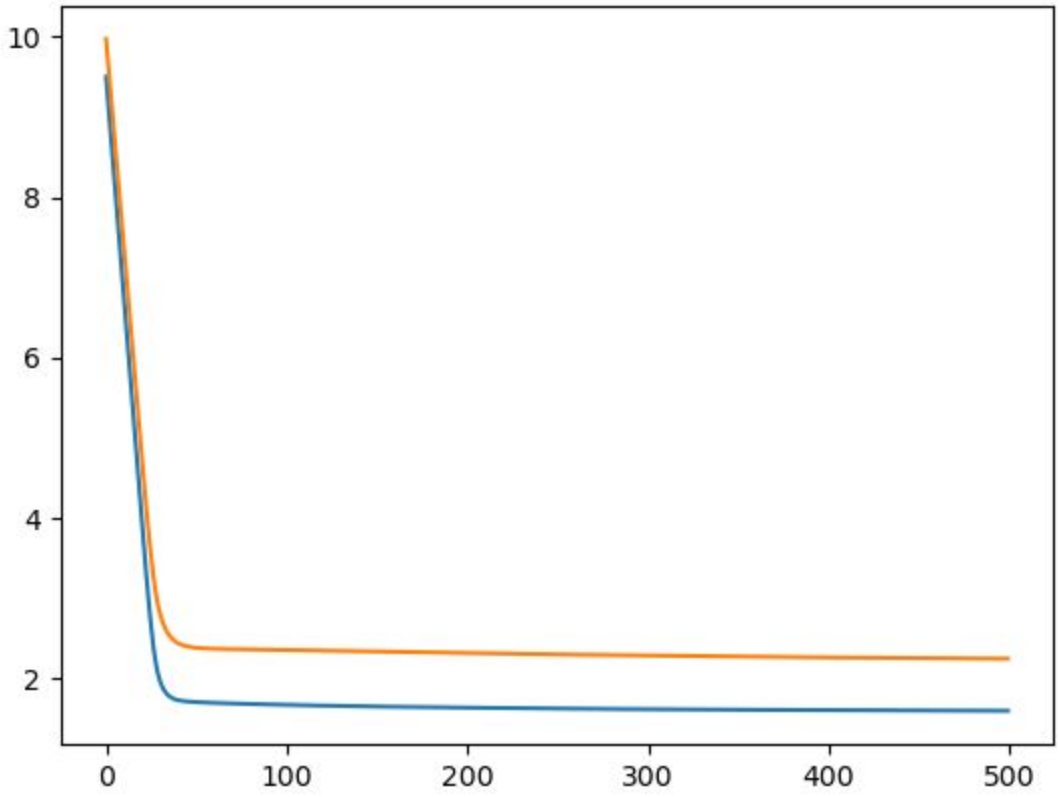
• Choose an appropriate value of K and justify it in your report along with the preprocessing strategy.

If the size of the dataset is smaller then the value of k should be higher to use more training data in each iteration and this will lower the bias but the dataset size is good enough in this assignment so I choose K=3 folds and also computation time is also lower in a low number of folds.

Preprocessing Strategy-
-Drop nan entries form data
-shuffle of dataset
-converting data entries to float 64

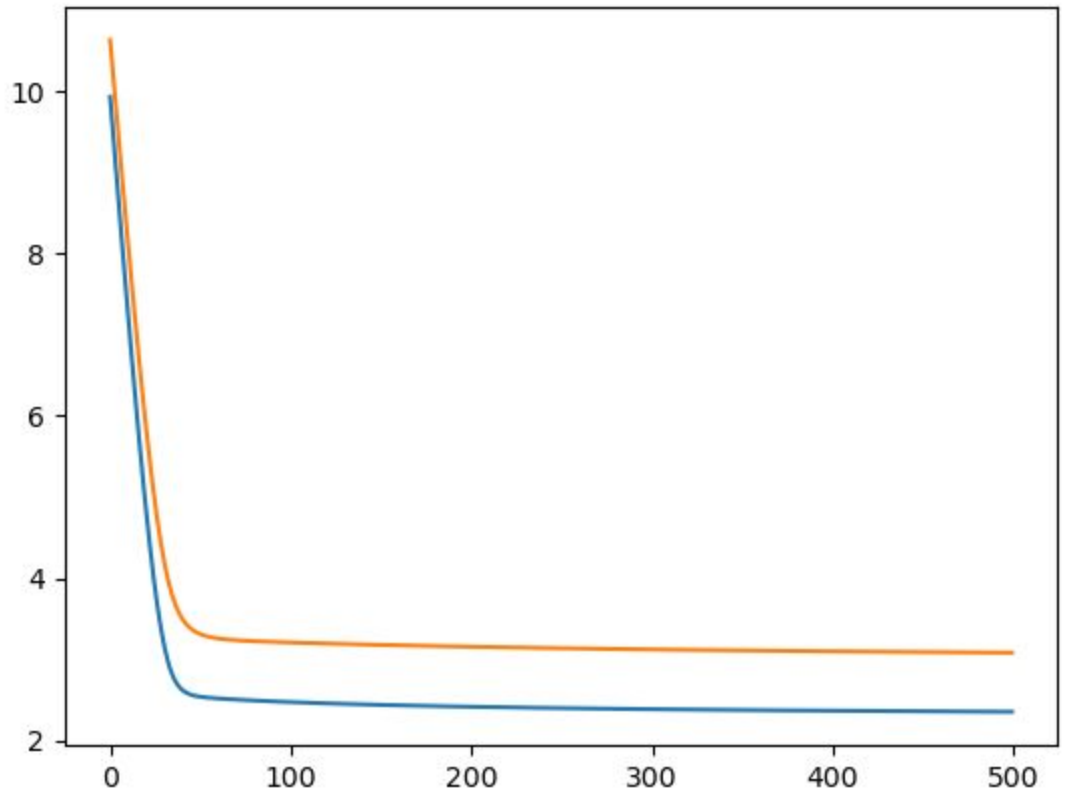
Analysis.
(a) Include plots between training loss v/s iterations and validation loss v/s iterations.(total 4 plots - 2 plots for each dataset).
Ans. Dataset-1 Plots
Fold -1
MAE Loss plot - 1) Training Loss - Blue 2) Validation Loss - Orange
MAE Train Loss , MAE Validation Loss= 1.6023556368946883 2.251300850611047

MAE Plot



RMSE Loss plot - 1) Training Loss - Blue 2) Validation Loss - Orange

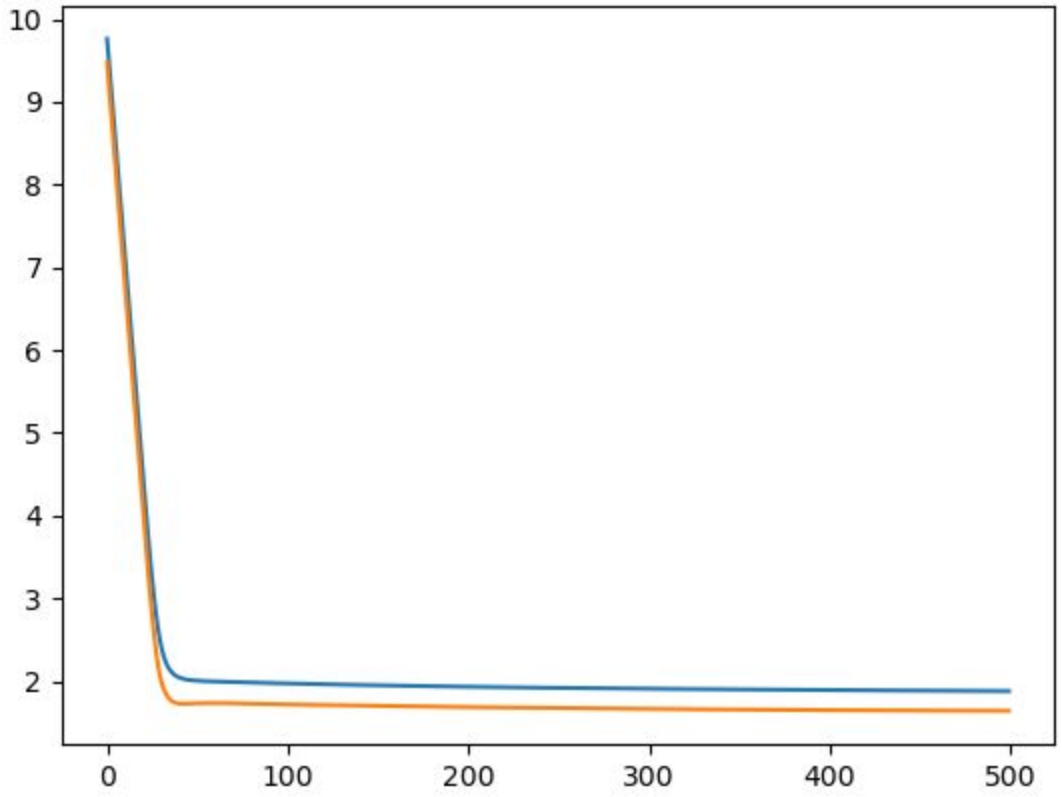
RMSE Train Loss , RMSE Validation Loss =2.351499269206243 3.079787354875813



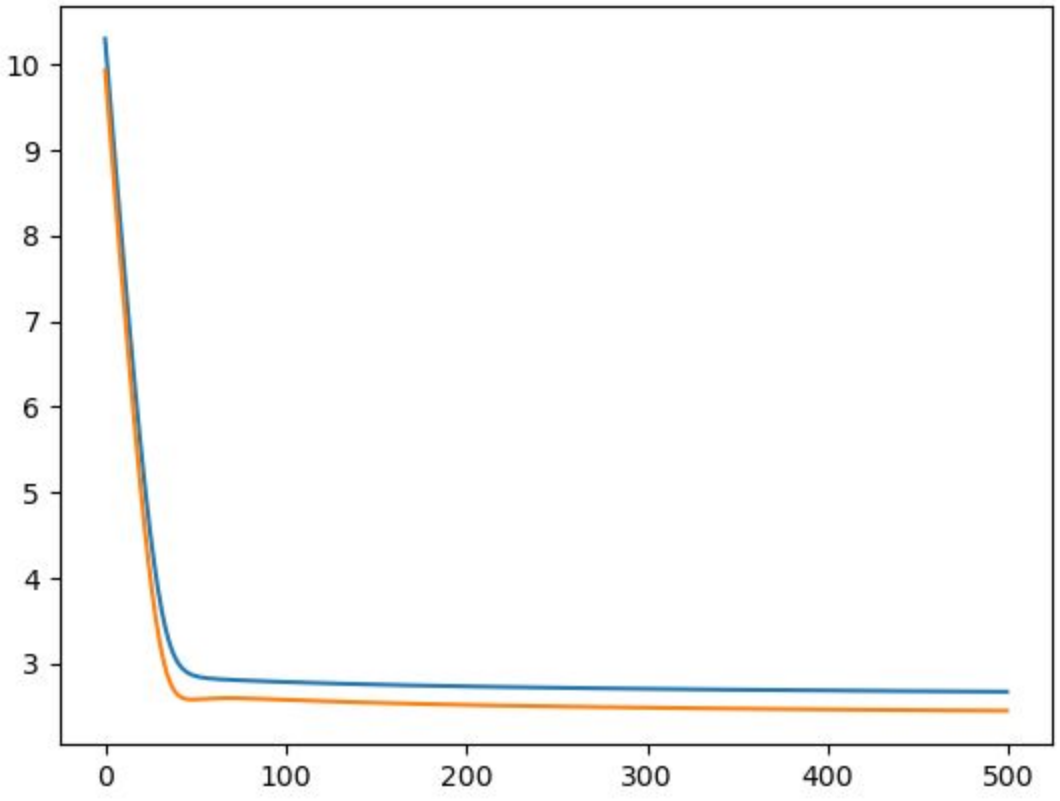
Fold -2

MAE Loss plot - 1) Training Loss - Blue 2) Validation Loss - Orange

MAE Train, Validation Loss- 1.8801890282795275, 1.6421272207974913

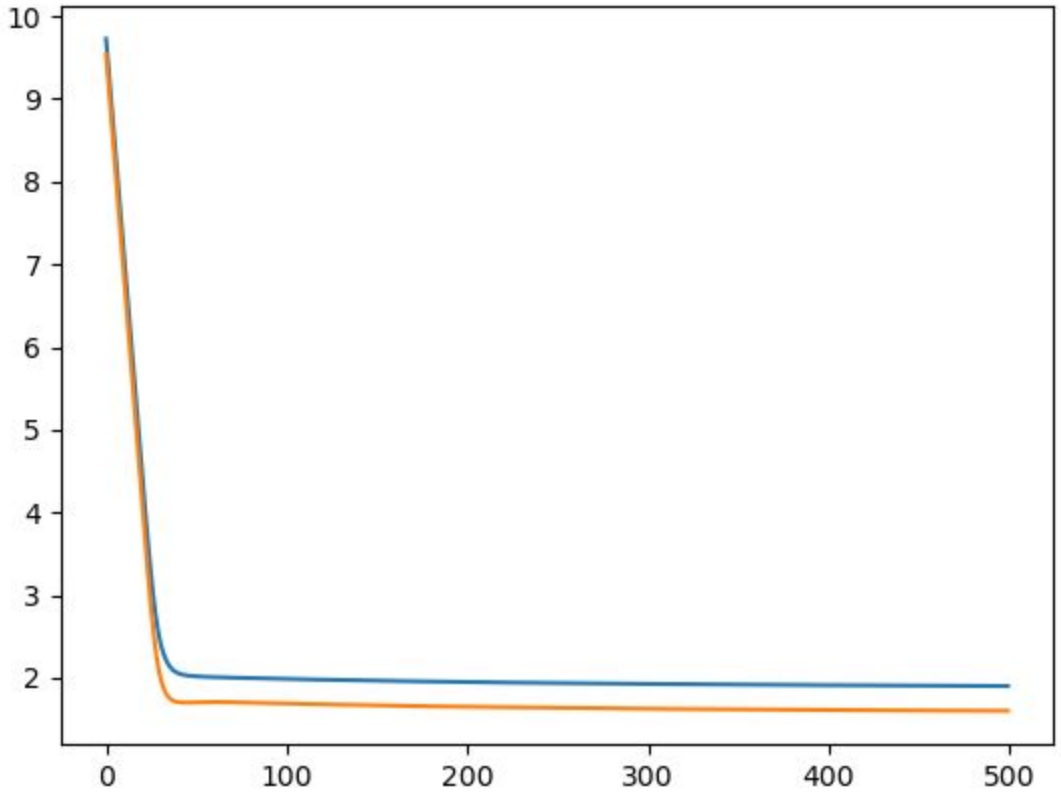


RMSE Loss plot - 1) Training Loss - Blue 2) Validation Loss - Orange
RMSE Train , Validation Loss -2.6724357546824895 2.4498871157634414

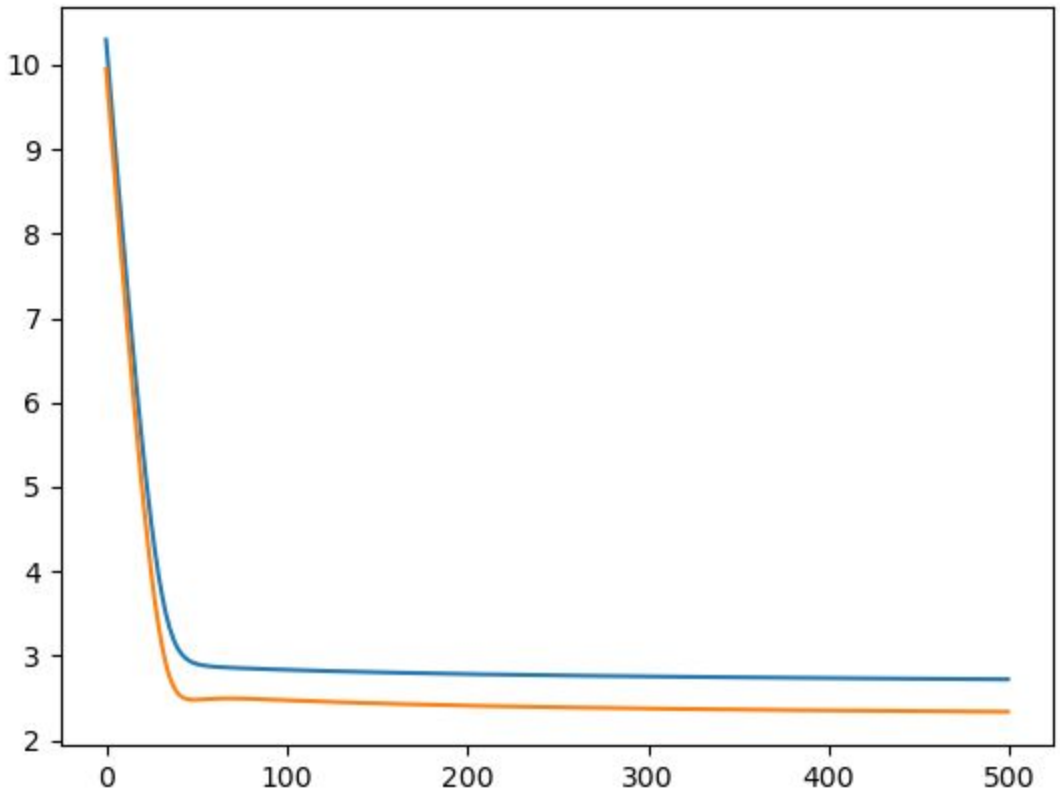


Fold -3
MAE Loss plot - 1) Training Loss - Blue 2) Validation Loss - Orange

MAE Train, Validation Loss -1.9011295780493653 1.601826115956705



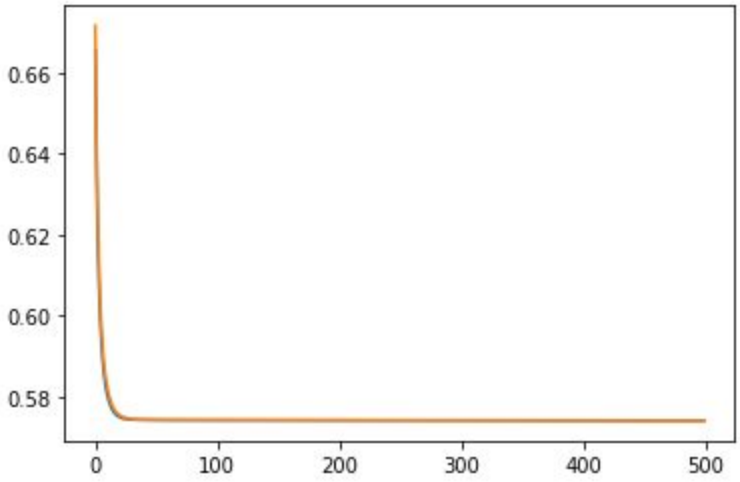
RMSE Loss plot - 1) Training Loss - Blue 2) Validation Loss - Orange
RMSE Train, Validation Loss - 2.7222705483975287, 2.338049447532514



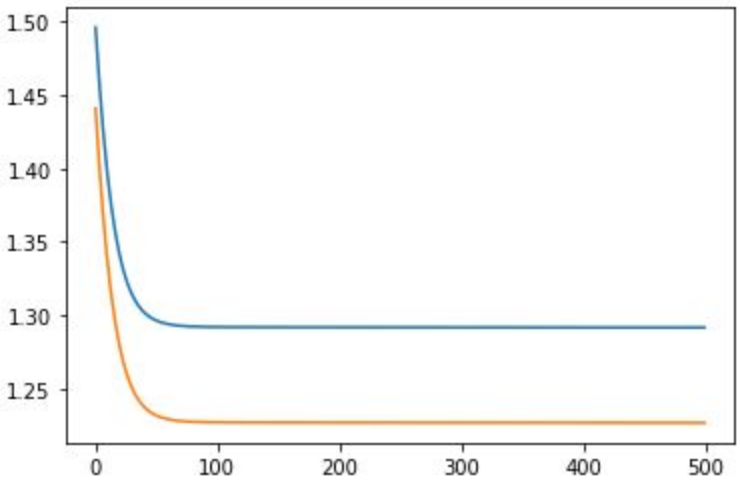
Dataset-2

Fold -1

MAE Loss plot - 1) Training Loss - Blue 2) Validation Loss - Orange
MAE train,Validate loss - 0.5740304276144139 , 0.5739678447667742



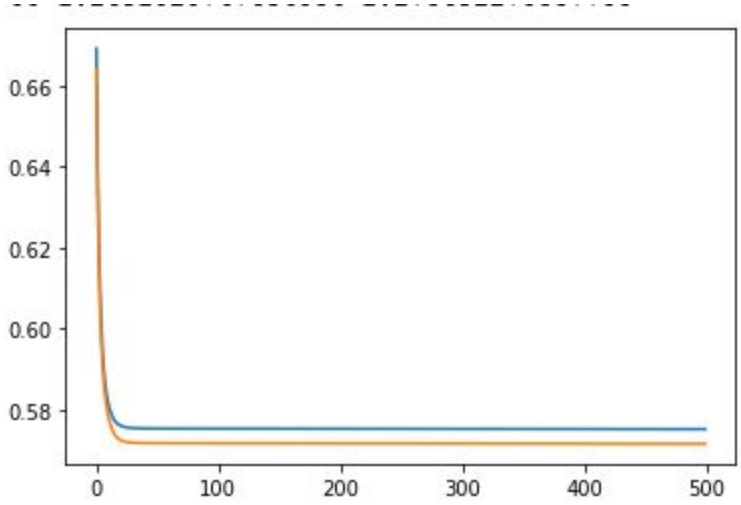
RMSE Loss plot - 1) Training Loss - Blue
2) Validation Loss - Orange
RMSE train, validate loss-1.2918328575567648, 1.2270463547663608



Fold -2
MAE Loss plot - 1) Training Loss - Blue

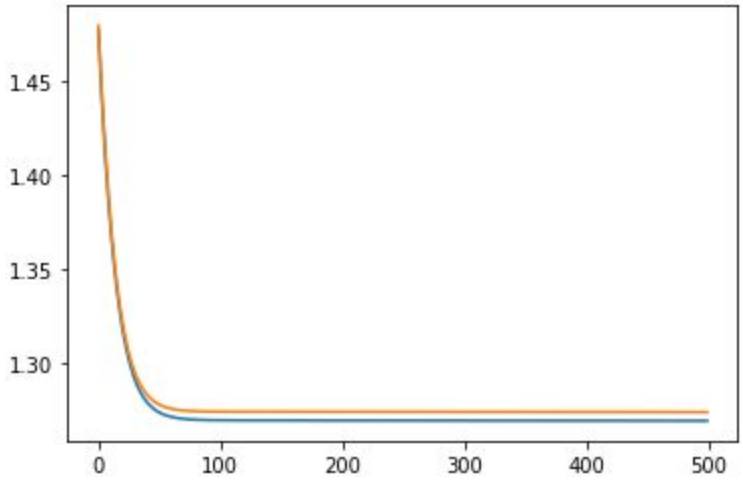
2) Validation Loss - Orange

MAE train,Validate loss - 0.5751938998887718 , 0.5715699758119687



RMSE Loss plot - 1) Training Loss - Blue 2) Validation Loss - Orange

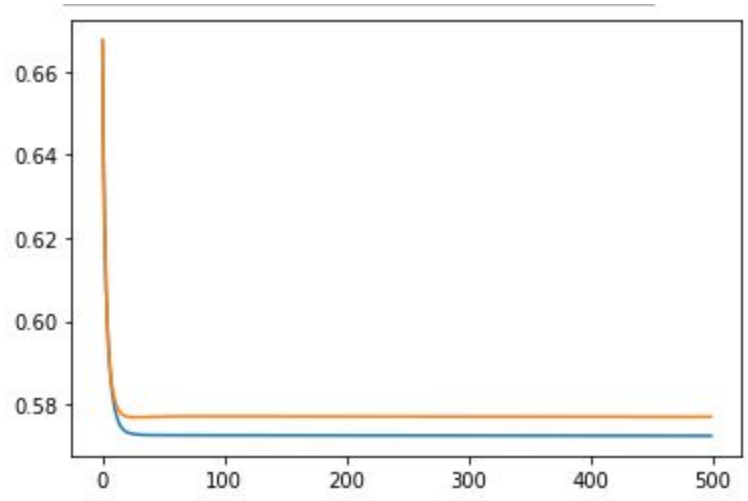
RMSE train, validate loss -1.2691020747684034 , 1.273692270337755



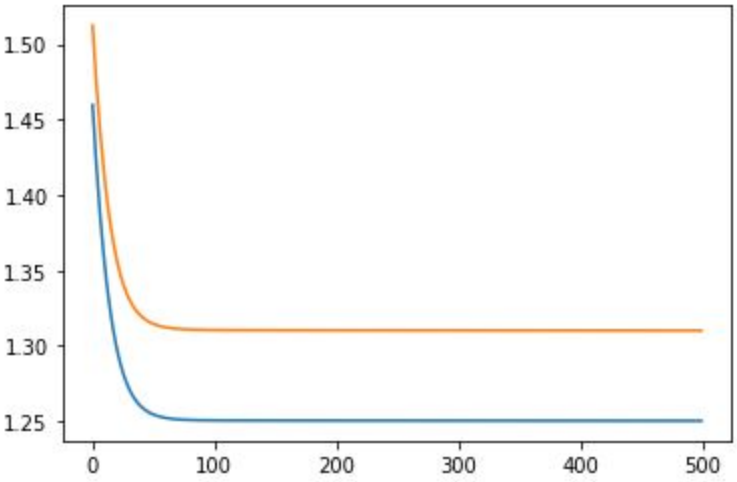
Fold -3

MAE Loss plot - 1) Training Loss - Blue 2) Validation Loss - Orange

MAE train,Validate loss - 0.5724933219724972 , 0.5770681130115273



RMSE Loss plot - 1) Training Loss - Blue 2) Validation Loss - Orange
RMSE train, validate loss - 1.2503991272438881 1.3101976037639023



(b) Include the best RMSE and MAE value achieved (as well as which fold achieves this) in your report.

Dataset-1
Best MAE train loss- 1.6023556368946883 in kfold=1
Best MAE val Loss-0.5715699758119687 in fold =2
Best RMSE train loss -2.351499269206243 in kflod=1
Besr RMSE val lass -2.338049447532514 in kfold =3

Dataset-2
Best MAE train loss-0.5724933219724972 in Kfold-3
Best MAE Val Loss-0.5715699758119687 in fold-3
Best RMSE train loss -1.2503991272438881 in fold-3
Besr RMSE val lass -1.2270463547663608 in kfold 1

(c) For each dataset, analyze and describe which of the loss leads to better performance. (Hint: Compare the values of RMSE and MAE).

Ans-MAE Loss leads to better performance as it gives lower validation loss than RMSE.

(d) What is the relationship between MAE and RMSE? Under what conditions are RMSE and MAE expected to give similar values? Which loss will you prefer in such a case and why? (1+1+3 = 5 points)

Ans-
Both MAE and RMSE gives average model prediction and both show the same behaviours (graph plots) to the direction of errors..

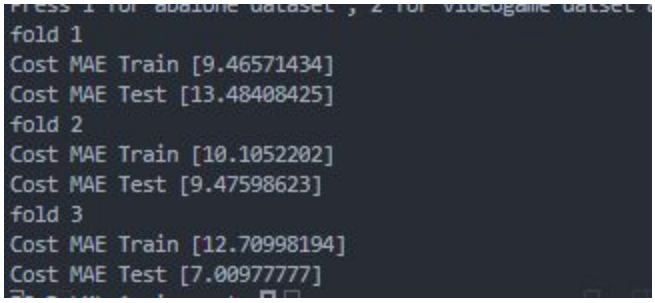
RMSE and MAE can give similar values when the predicted value becomes equal to the original value.

In this case, I prefer RMSE because if predicted values becomes equal to the original value then derivative of MAE is undefined.

(e). Implement the normal equation form (closed form) of linear regression and get the

optimal parameters directly. Consider the Dataset 1 and the most appropriate loss function you've described for this dataset in part(c). Compute the training and validation loss for the best fold for this loss described in part(b) using these optimal parameters.

Ans-



2. Logistic Regression

Analyze the class distributions and comment on the feature values for each of the given features.

	variance	skewness	curtosis	entropy	class
count	1372.000000	1372.000000	1372.000000	1372.000000	1372.000000
mean	0.433735	1.922353	1.397627	-1.191657	0.444606
std	2.842763	5.869047	4.310030	2.101013	0.497103
min	-7.042100	-13.773100	-5.286100	-8.548200	0.000000
25%	-1.773000	-1.708200	-1.574975	-2.413450	0.000000
50%	0.496180	2.319650	0.616630	-0.586650	0.000000
75%	2.821475	6.814625	3.179250	0.394810	1.000000
max	6.824800	12.951600	17.927400	2.449500	1.000000

one count of class: 610
Zero counts of class:762

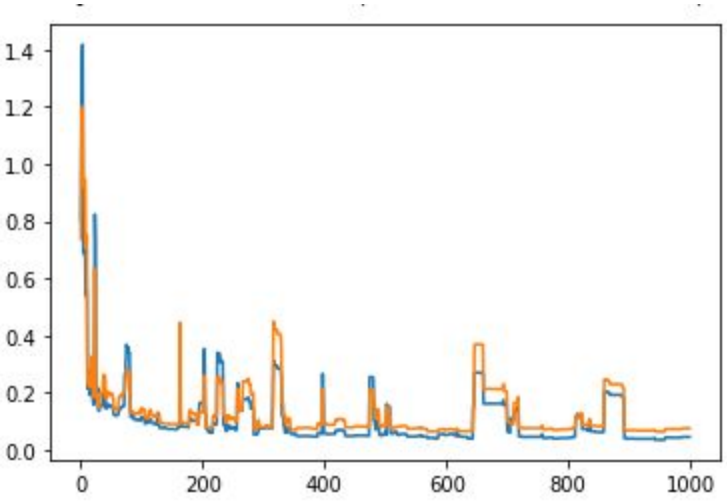
(a) Using Stochastic Gradient Descent (SGD), choose an appropriate learning rate and the number of epochs (iterations). Report the accuracy obtained on both the training and test set.

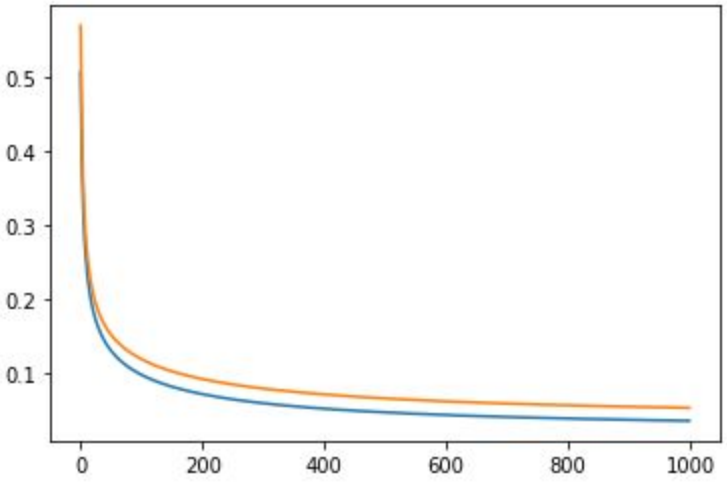
Ans-

```
sgd training_set_accuracy 94.47340980187695
bgd training_set_accuracy 98.74869655891554
sgd testing_set_accuracy 94.92753623188406
bgd testing_set_accuracy 98.91304347826086
```

(b) Include plots between training loss v/s iterations and validation loss vs iterations.

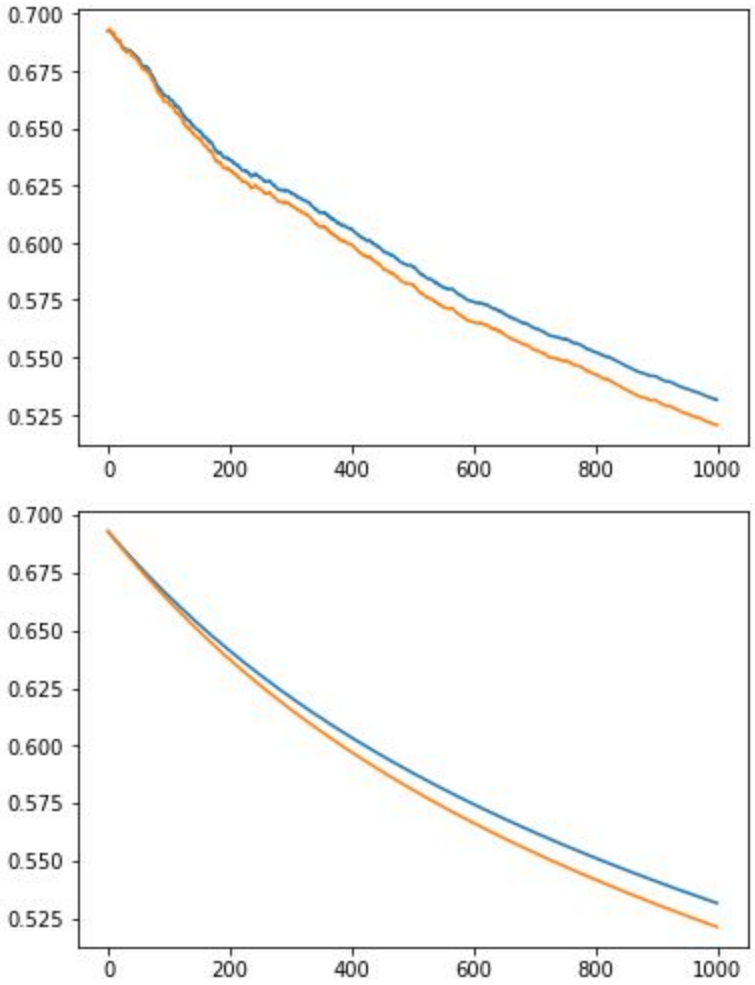
Sgd loss,bgd loss


1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000

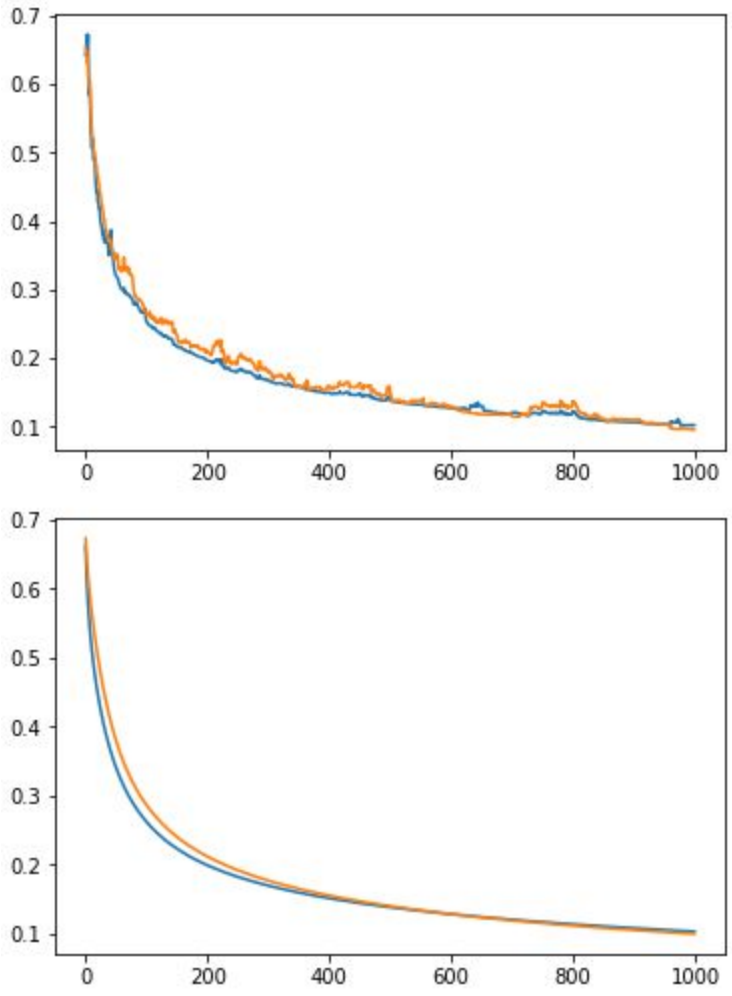

0.5
0.4
0.3
0.2
0.1
0
200
400
600
800
1000

(c) Re-run your implementation for 3 variations in learning rates - 0.0001, 0.01, 10
Sgd and bgd graph

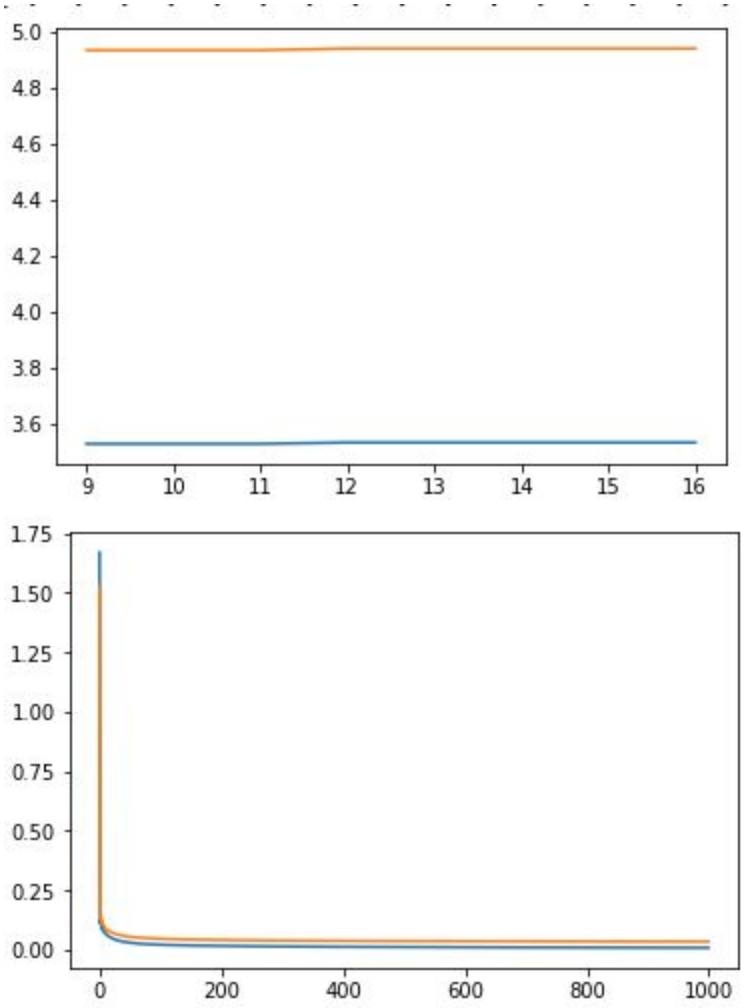
apha=0.001



alpha=0.01



alpha=10



(a) Loss plots

above

(b) Number of epochs taken to converge.

SGD loss train-970 epochs
SGD loss test-660 epochs

BGD loss -1000 epochs
BGD loss -1000 epochs

(d) Report the accuracy obtained on both the training and test set. Use the same hyper-parameters as in the SGD implementation in 2(a), and compare sklearn's performance with that.

```
sgd_training_loss 0.9812304483837331
sgd_testing_loss 0.9855072463768116
```