Report - Assignment 1

Machine Learning (CSE343), Monsoon 2021

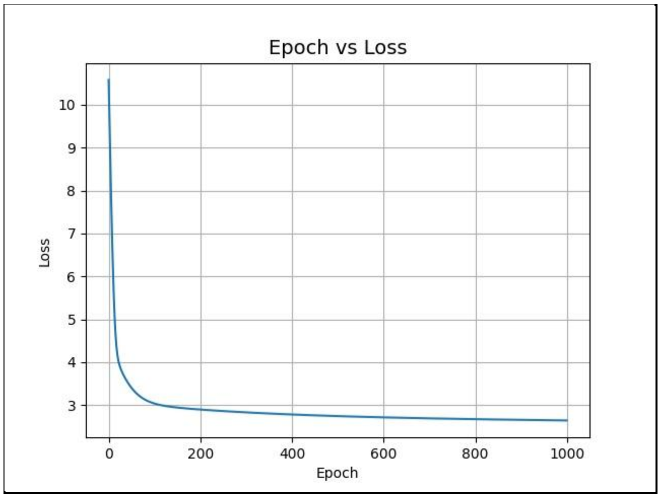
- Abhimanyu Gupta
2019226

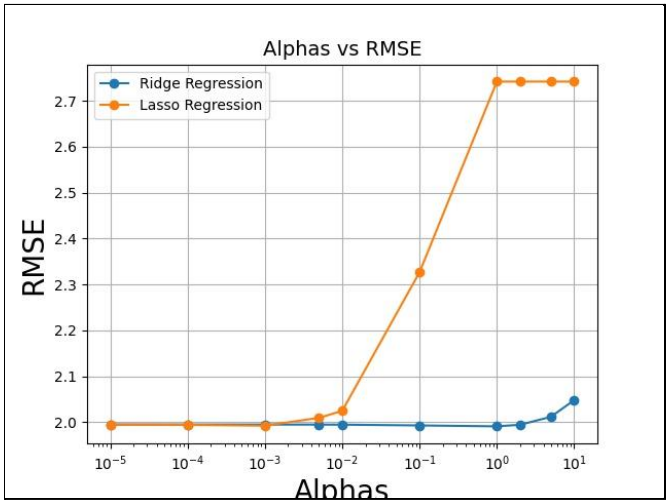
Question 1

Pre-Processing:

- Dataset is read into pandas dataframe
- Records are shuffled and converted into numpy arrays
- Following mappings are made, Male - 1, Female - 2, Infant - 3
- 8:2 train: test division is made on the dataset

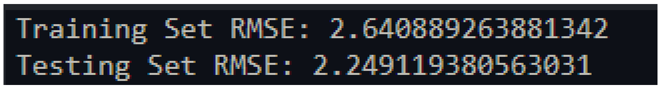
Plots:





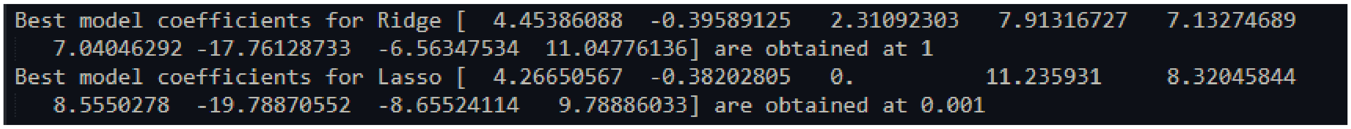
Outputs:

1.1



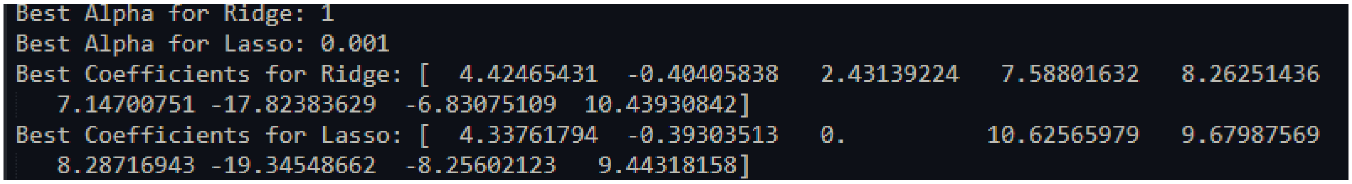
```
Training Set RMSE: 2.640889263881342
Testing Set RMSE: 2.249119380563031
```

1.2



```
Best model coefficients for Ridge [  4.45386088  -0.39589125   2.31092303   7.91316727   7.13274689
   7.04046292 -17.76128733  -6.56347534  11.04776136] are obtained at 1
Best model coefficients for Lasso [  4.26650567  -0.38202805   0.          11.235931     8.32045844
   8.5550278  -19.78870552  -8.65524114   9.78886033] are obtained at 0.001
```

1.3



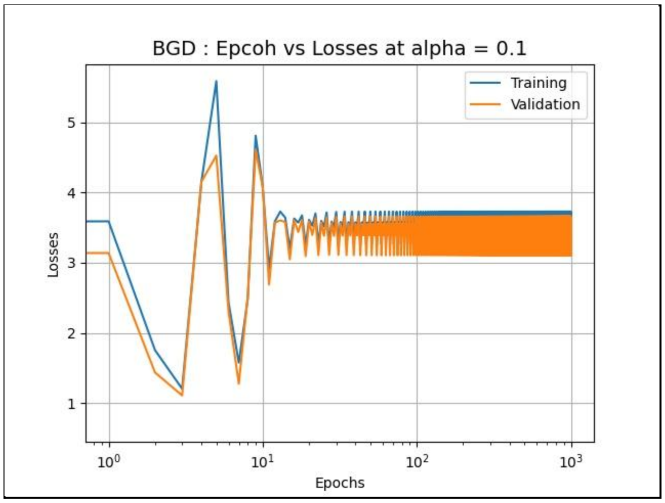
```
Best Alpha for Ridge: 1
Best Alpha for Lasso: 0.001
Best Coefficients for Ridge: [  4.42465431  -0.40405838   2.43139224   7.58801632   8.26251436
   7.14700751 -17.82383629  -6.83075109  10.43930842]
Best Coefficients for Lasso: [  4.33761794  -0.39303513   0.          10.62565979   9.67987569
   8.28716943 -19.34548662  -8.25602123   9.44318158]
```

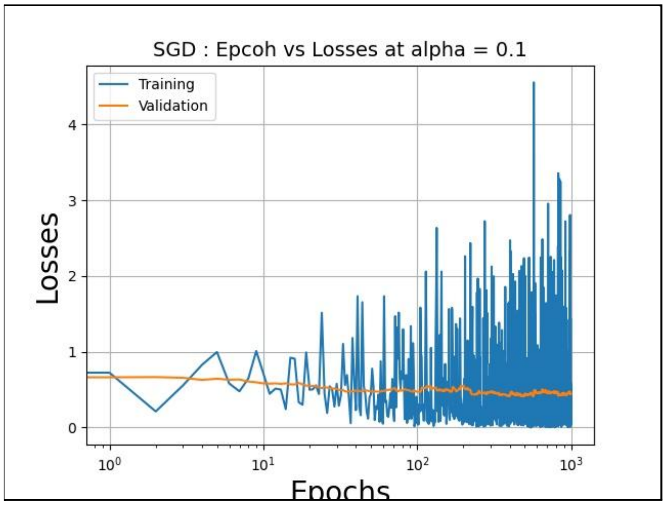
Question 2

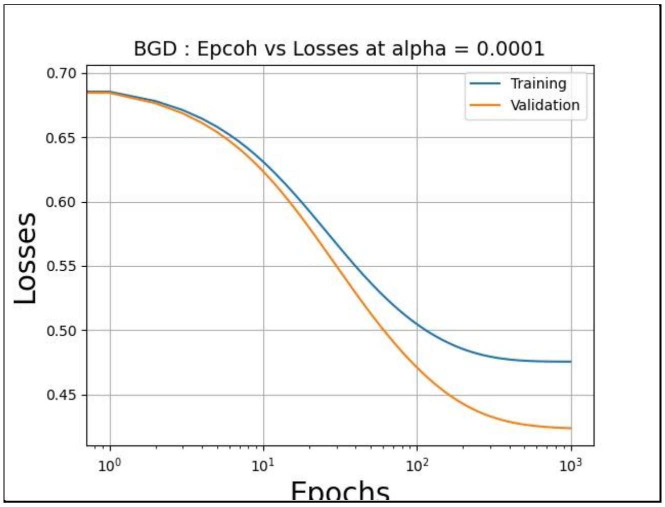
Pre-Processing:

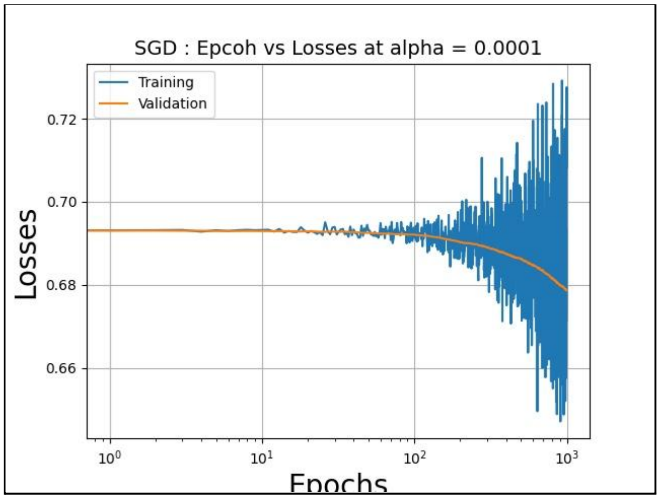
- Dataset is read into pandas dataframe
- Records are shuffled and converted into numpy arrays
- Since columns like Blood Pressure can’t contain zero, such zeroes are replaced by columns respective medians
- Columns are normalized
- 7:2:1 train:val:test division is made on the dataset

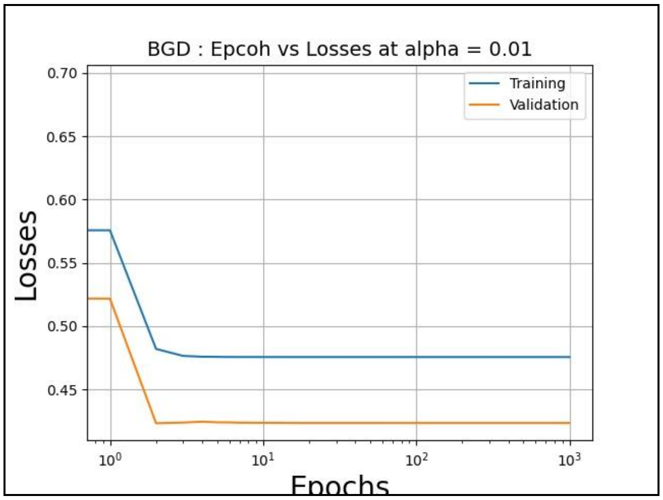
Plots:

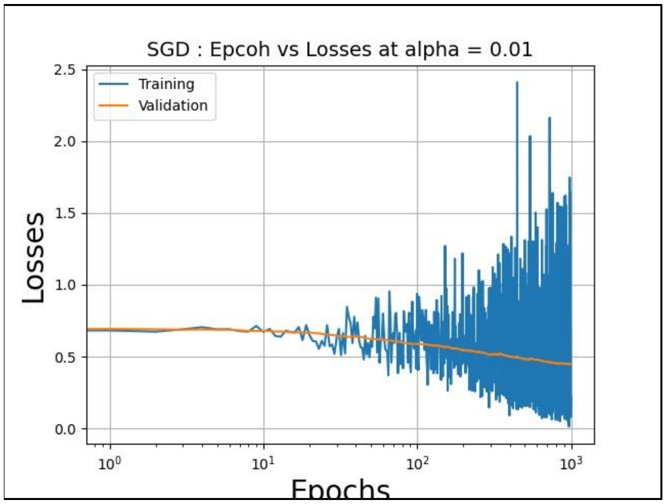


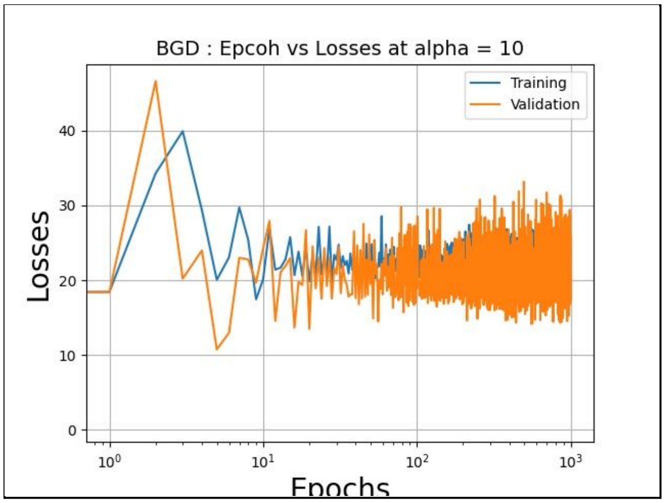


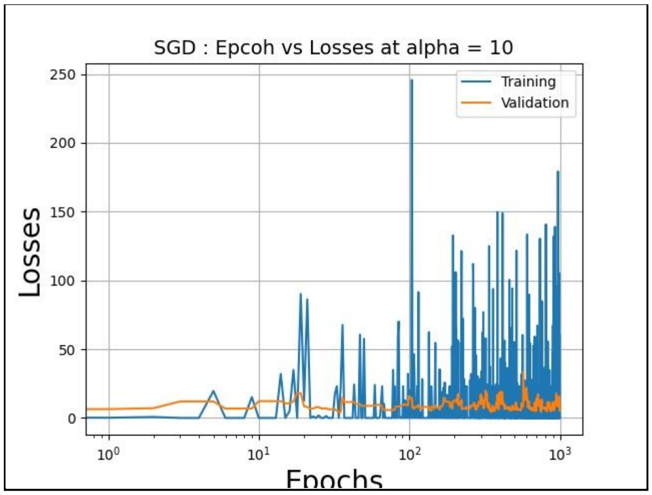






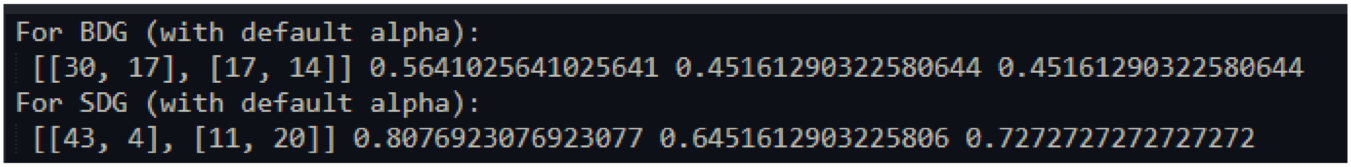






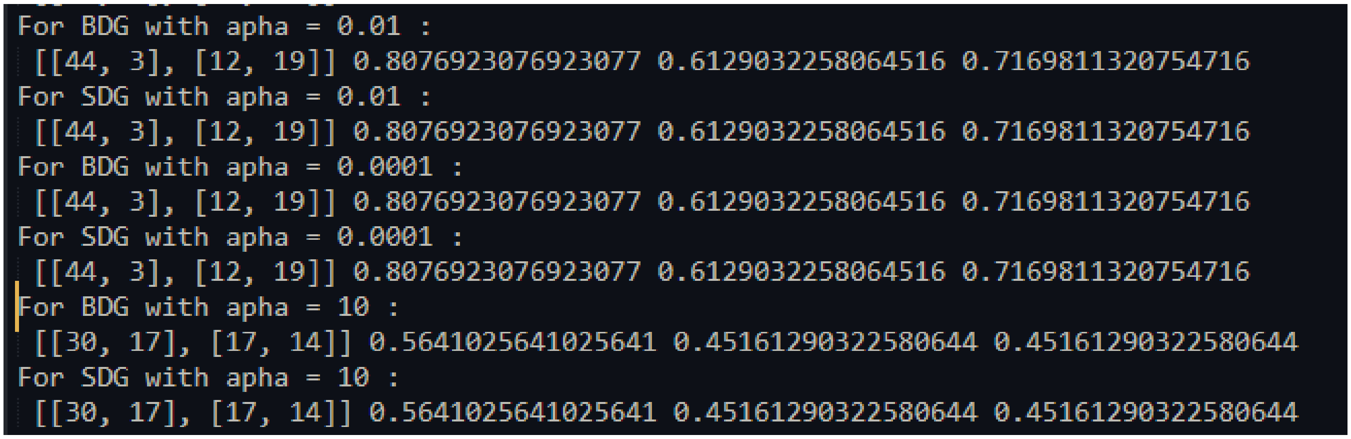
Outputs:

1.a



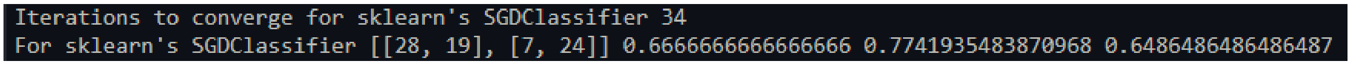
```
For BDG (with default alpha):
 [[30, 17], [17, 14]] 0.5641025641025641 0.45161290322580644 0.45161290322580644
For SDG (with default alpha):
 [[43, 4], [11, 20]] 0.8076923076923077 0.6451612903225806 0.7272727272727272
```

1.b



```
For BDG with apha = 0.01 :
 [[44, 3], [12, 19]] 0.8076923076923077 0.6129032258064516 0.7169811320754716
For SDG with apha = 0.01 :
 [[44, 3], [12, 19]] 0.8076923076923077 0.6129032258064516 0.7169811320754716
For BDG with apha = 0.0001 :
 [[44, 3], [12, 19]] 0.8076923076923077 0.6129032258064516 0.7169811320754716
For SDG with apha = 0.0001 :
 [[44, 3], [12, 19]] 0.8076923076923077 0.6129032258064516 0.7169811320754716
For BDG with apha = 10 :
 [[30, 17], [17, 14]] 0.5641025641025641 0.45161290322580644 0.45161290322580644
For SDG with apha = 10 :
 [[30, 17], [17, 14]] 0.5641025641025641 0.45161290322580644 0.45161290322580644
```

2.c



```
Iterations to converge for sklearn's SGDClassifier 34
For sklearn's SGDClassifier [[28, 19], [7, 24]] 0.6666666666666666 0.7741935483870968 0.6486486486486487
```

Analysis:

1.a

In the BDG both training and validation curve converges between 3 and 4, while for SGD although the training curve oscillates more as epochs are increase, the validation curve converges to a value between 0 and 1.

1.b

For smaller values of alpha (0.01, 0.001), in BGD both training and validation curves smoothly converge to a value around 0.45. Similarly, while in SGD although the training curve oscillates more as epochs are increase, the validation curve converges and reduces as epochs increase to a value less than 1.

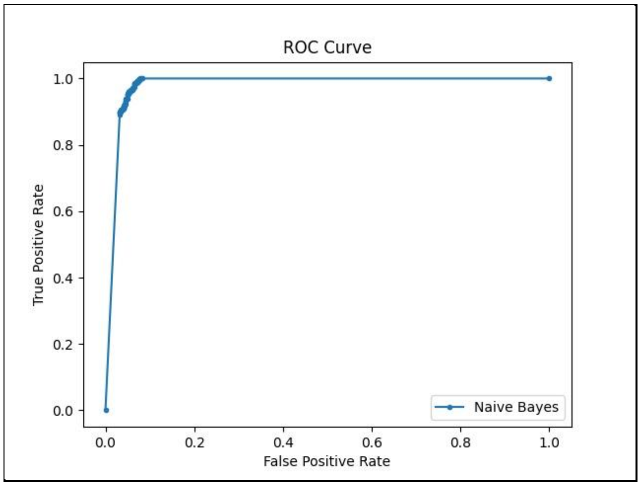
But, for alpha = 10, cost function consciously oscillates and never really appears to converge, this is because, for high learning rate, gradient descent tends to skip minima and hence overshoots.

Question 3

Pre-Processing:

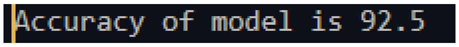
- Dataset is read into pandas dataframe
- Dataset is filtered for Trouser and Pullover
- Binarization of the extracted dataset

Plots:



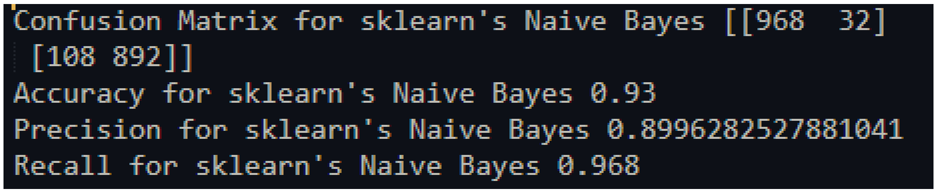
Outputs:

3.1



```
Accuracy of model is 92.5
```

3.3



```
Confusion Matrix for sklearn's Naive Bayes [[968  32]
 [108 892]]
Accuracy for sklearn's Naive Bayes 0.93
Precision for sklearn's Naive Bayes 0.8996282527881041
Recall for sklearn's Naive Bayes 0.968
```

Analysis:

3.2

K is chosen to be 5, as it is not too low or not high and moreover having 5 folds divide the dataset into 8:2 train:test split which is appropriate for a machine learning model.

Question 4

1. All three cases and can be covered, by incorporating men and women in our model by adding a new feature X_2 and its corresponding parameter B_2. That is $W = B_0 + x_1 * B_1 + x_2 * B_2 + u$
    a. We can compare the value of B_0 to detect suspicion
    b. We can compare the value of B_1 to detect suspicion
    c. We can run Linear Regression and can check if the error is low and accuracy is high to detect suspicion.
2.
3.