Name: Bheki Maenetja
Student Number: 230382466
Assignment: No.1 Part 1

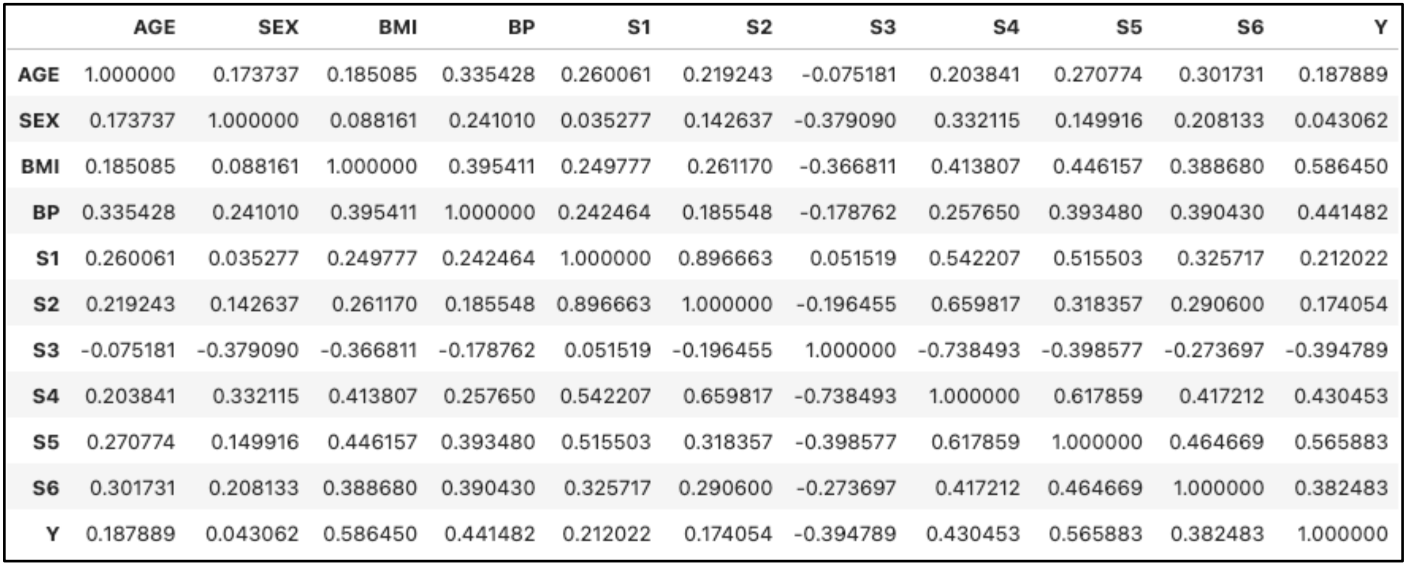
Question 5

What conclusion if any can be drawn from the weight values?

Looking at the weights we can see that BMI (weight = 26.3) is the most influential predictor of Y. Second to it is S5 (weight = 21.98). BMI and S5 have a relatively large influence on Y in comparison to the other features. Another interesting takeaway is that gender has a significant negative weight (-11.4) which suggests that gender may have little to no effect on Y; this is further evidenced by the correlation matrix (see figure 1). The rest of the features appear to only have a more modest influence on Y, given the lower absolute values of their weightings.

How does gender and BMI affect blood sugar levels?

Looking at the correlation matrix created in the notebook, see that both BMI and gender have a positive but weak correlation with blood sugar (column S6). This suggests that as these variables increase, so too will blood sugar. BMI has stronger correlation (0.39) with blood sugar than gender (0.21). However, this correlation is still fairly modest.



	AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6	Y
AGE	1.000000	0.173737	0.185085	0.335428	0.260061	0.219243	-0.075181	0.203841	0.270774	0.301731	0.187889
SEX	0.173737	1.000000	0.088161	0.241010	0.035277	0.142637	-0.379090	0.332115	0.149916	0.208133	0.043062
BMI	0.185085	0.088161	1.000000	0.395411	0.249777	0.261170	-0.366811	0.413807	0.446157	0.388680	0.586450
BP	0.335428	0.241010	0.395411	1.000000	0.242464	0.185548	-0.178762	0.257650	0.393480	0.390430	0.441482
S1	0.260061	0.035277	0.249777	0.242464	1.000000	0.896663	0.051519	0.542207	0.515503	0.325717	0.212022
S2	0.219243	0.142637	0.261170	0.185548	0.896663	1.000000	-0.196455	0.659817	0.318357	0.290600	0.174054
S3	-0.075181	-0.379090	-0.366811	-0.178762	0.051519	-0.196455	1.000000	-0.738493	-0.398577	-0.273697	-0.394789
S4	0.203841	0.332115	0.413807	0.257650	0.542207	0.659817	-0.738493	1.000000	0.617859	0.417212	0.430453
S5	0.270774	0.149916	0.446157	0.393480	0.515503	0.318357	-0.398577	0.617859	1.000000	0.464669	0.565883
S6	0.301731	0.208133	0.388680	0.390430	0.325717	0.290600	-0.273697	0.417212	0.464669	1.000000	0.382483
Y	0.187889	0.043062	0.586450	0.441482	0.212022	0.174054	-0.394789	0.430453	0.565883	0.382483	1.000000

Figure 1 Correlation matrix of features in diabetes dataset. See notebook.

Question 6

Try the code with a number of learning rates that differ by orders of magnitude and record the error of the training and test sets. What do you observe on the training error? What about the error on the test set?

Looking at the graph below (see figure 2) we observe that the error on the training set seems to be more sensitive to the choice of learning rate.). The error is very high for large alpha values, suggesting that the model might be failing to converge on an optimal solution; overshooting instead. As the learning rate decreases the training set error generally decreases before rising again. The optimal value for alpha with regard to the training set error appears to be somewhere between 0.1 and 0.01. The test set error follows a similar pattern to the training set error: very high for high values of alpha, then decreasing before increasing again as alpha gets very small. Like the training set error, the optimal value for alpha is somewhere between 0.1 and 0.01.

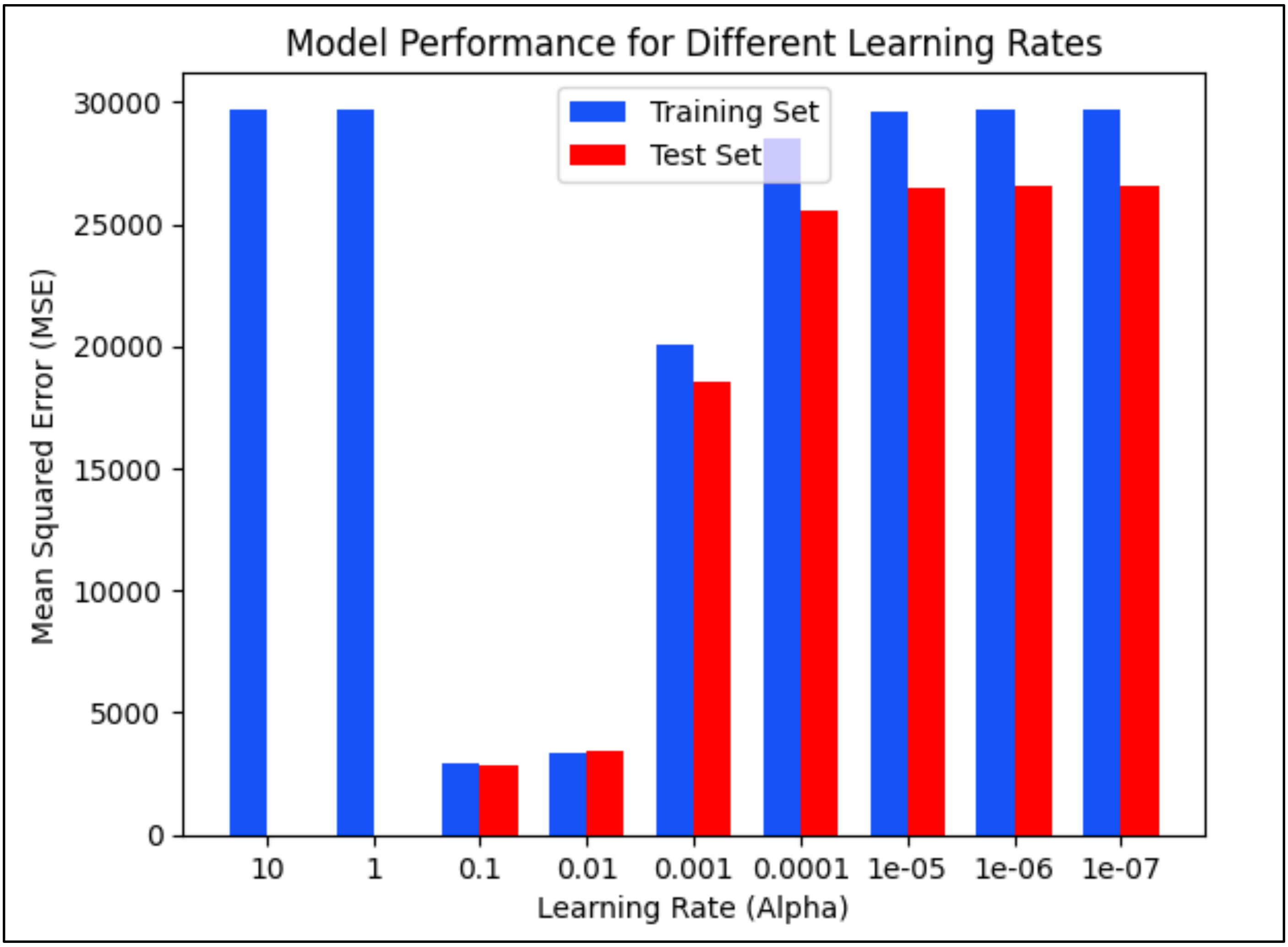


Figure 2 Graph showing the MSE on the training and test sets for various learning rate values.