# Q5

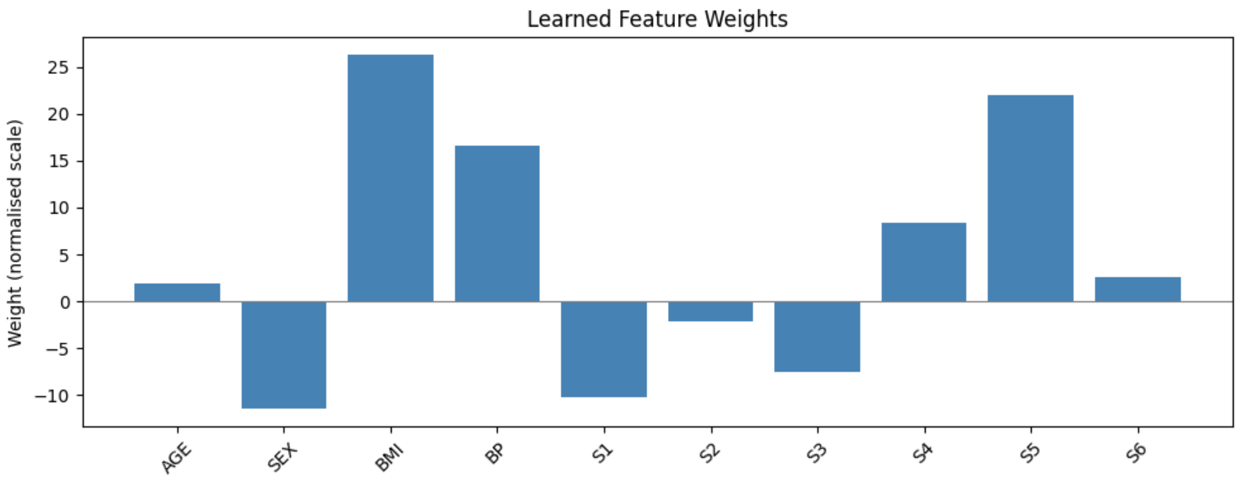


Figure 1: Feature values, before and after normalization

The model assigns the largest positive weights to BMI , the S5 serum marker, and BP, indicating these features most strongly drive disease progression. The most negative weights belong to SEX, S1, and S3. These weights suggest that**Higher BMI is the strongest predictor of faster diabetes**.

With a coefficient of approximately -10, and with the encoding of females as -1 and males as +1, the model predicts that **female patients experience faster progression than male patients**. When other factors are constant, women are predicted to progress roughly 20 points higher on the progression scale compared to men.

The predicted progression values for the examples given were 112.35 and 232.6, for the 25 year old female and the 50 year old male respectively.

# Q6

| $\alpha$ | Train MSE | Test MSE |
|---|---|---|
| 0.001 | 19767.20 | 18346.47 |
| 0.010 | 3351.02 | 3425.63 |
| 0.100 | 2890.00 | 2886.15 |
| 1.000 | Diverged (NaN) | Diverged (NaN) |
| 10.000 | Diverged (NaN) | Diverged (NaN) |

Table 1: Training and test MSE after 100 iterations for different learning rates.

As learning rate increases from 0.001 to 0.01 to 0.1, the train and test errors after 100 iterations can be seen to decrease significantly. **A learning rate too small results in slow convergence**. For example, at $\alpha$ =0.001 the model takes very small gradient steps, resulting in a very slow convergence and an under-trained model after 100 iterations.

**Excessively large learning rates cause divergence**. At $\alpha = 1.0$ and $\alpha = 10.0$, gradient descent repeatedly overshoots the minimum so that weights explode, resulting in NaN values due to numerical overflow.

# 1 Q8

Various learning rates were tested with $\lambda = 0$ to find the optimal $\alpha$. Testing showed that $\alpha = 1.4$ achieved a cost of 0.0079, while $\alpha = 1.5$ resulted in a higher cost of 0.026. Further testing with $\alpha = 1.41$ yielded an even lower cost, indicating the optimal learning rate lies in the range $1.4 < \alpha < 1.5$ when $\lambda = 0$.

Various values of $\lambda$ were then tested. As $\lambda$ increased, the hypothesis curve became smoother and simpler, withoutout any sharp bends. This is because **regularization penalizes large polynomial coefficients**, preventing the model from overfitting to the training data. This trade-off between fitting the training data closely and maintaining model simplicity is essential for generalization to unseen data.