A1

## 1. Exploration

After the dataset is loaded, we can take a look at the first 10 rows of data to develop our intuition of the dataset. The target has been added as a column named progression into a dataframe with the rest of the data.

	age	sex	bmi	bp	s1	s2	s3	s4	s5	s6	progression
0	0.038076	0.050680	0.061696	0.021872	-0.044223	-0.034821	-0.043401	-0.002592	0.019907	-0.017646	151.0
1	-0.001882	-0.044642	-0.051474	-0.026328	-0.008449	-0.019163	0.074412	-0.039493	-0.068332	-0.092204	75.0
2	0.085299	0.050680	0.044451	-0.005670	-0.045599	-0.034194	-0.032356	-0.002592	0.002861	-0.025930	141.0
3	-0.089063	-0.044642	-0.011595	-0.036656	0.012191	0.024991	-0.036038	0.034309	0.022688	-0.009362	206.0
4	0.005383	-0.044642	-0.036385	0.021872	0.003935	0.015596	0.008142	-0.002592	-0.031988	-0.046641	135.0
5	-0.092695	-0.044642	-0.040696	-0.019442	-0.068991	-0.079288	0.041277	-0.076395	-0.041176	-0.096346	97.0
6	-0.045472	0.050680	-0.047163	-0.015999	-0.040096	-0.024800	0.000779	-0.039493	-0.062917	-0.038357	138.0
7	0.063504	0.050680	-0.001895	0.066629	0.090620	0.108914	0.022869	0.017703	-0.035816	0.003064	63.0
8	0.041708	0.050680	0.061696	-0.040099	-0.013953	0.006202	-0.028674	-0.002592	-0.014960	0.011349	110.0
9	-0.070900	-0.044642	0.039062	-0.033213	-0.012577	-0.034508	-0.024993	-0.002592	0.067737	-0.013504	310.0

There are 442 rows, each with 11 columns, and of those columns there are 10 attribute variables and one target variable. A quick glance at the data shows it has been processed in
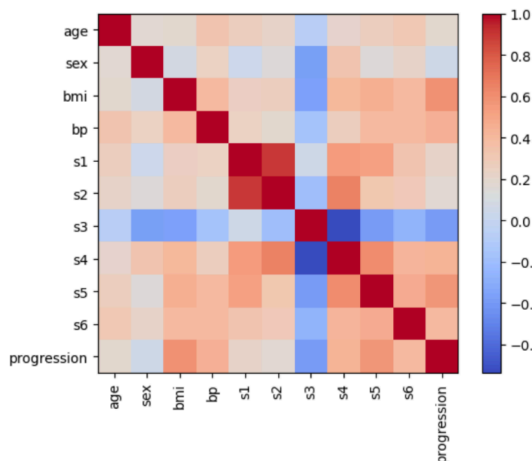
some way, and looking at the dataset documentation reveals that "each of these 10 feature variables have been mean centered and scaled by the standard deviation times the square root of n_samples". Based on the loading function, the features are real values between -0.2 and 0.2 and the target is an integer between 25 and 346. The target is stated to represent disease progression after a one year baseline. The dataset out of the box appears clean such that no values are missing and all the attributes have been normalized or processed in some manner. The features themselves correspond to data taken from patients about their health and demographic information between age, sex, blood pressure, and blood measurements.

To investigate the data further, let's look at a correlation matrix of the dataset.

This gives us a picture about some correlations in the data. Notably S3 has weak correlation with the rest of the features. In terms of correlations with the target, we see the most notable correlation involving s5 and bmi on the progression of the disease. Furthermore, we see notable correlations in the features such as (s4, s2), (s5, s4), and a very strong correlation with (s2, s1).

2 & 3. Model Fit and RFE

Before performing RFE, we find that there is an MSE of 5075.28 and an R^2: 0.05 for the initial model.

	coef
age	-0.402332
sex	-10.887657
bmi	25.723753
bp	16.215056
s1	-44.787686
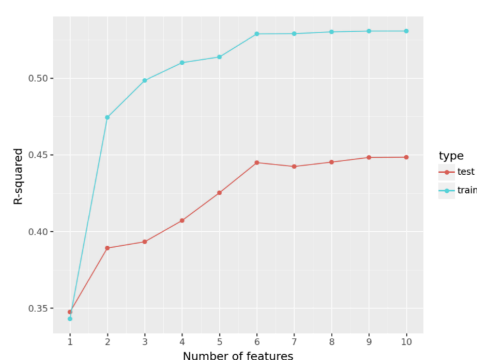
After performing RFE, we can find that the ideal number of features to choose is 6. There is a notable difference in performance

between train and test data, with a sharp drop in test data performance.

4. Analyze Feature Importance

	age	sex	bmi	bp	s1	s2	s3	s4	s5	s6
1	-	-	45.650347	-	-	-	-	-	-	-
2	-	-	31.611380	-	-	-	-	-	31.552411	-
3	-	-	28.007185	13.631823	-	-	-	-	27.762012	-
4	-	-	28.550848	13.833697	-9.855891	-	-	-	32.577782	-
5	-	-	27.106807	13.974239	-21.949705	12.079356	-	-	35.765062	-
6	-	-10.636346	25.861744	16.072454	-32.420861	21.862475	-	-	39.728338	-
7	-	-10.523376	25.891648	16.028773	-37.379165	26.063952	2.195756	-	41.870962	-
8	-	-10.816302	25.971188	16.503181	-44.864372	28.461003	9.122351	8.092169	42.596858	-
9	-	-10.944739	25.727677	16.134032	-44.775003	28.249702	8.891999	7.690111	42.091842	2.110479
10	-0.402332	-10.887657	25.723753	16.215056	-44.787686	28.319852	8.925173	7.660856	42.138026	2.168386

The three most important features for predicting the progression of diabetes appear to be bmi, blood pressure, and s5. This aligns with what we saw in the correlation matrix. The most significant predictor is bmi closely followed by s5. Age was a surprisingly weak predictor.

We can compare the initial and final models as follows:
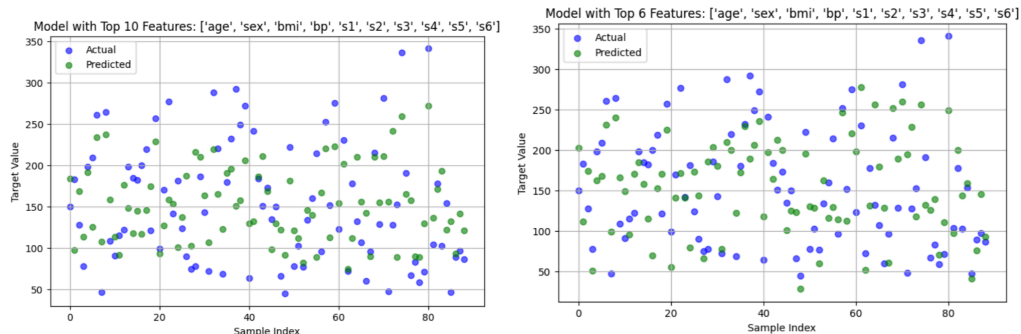
We can see pretty similar performance between the initial and final models. Both models perform pretty poorly and have similar values of R-squared as based on our earlier plot of model R-squared performance. The 6-featured model, by virtue of being simpler while keeping the same performance, is a better fit for the data.

5. Reflection

RFE appears to be a very effective method for iteratively selecting important features. It's ease of use when wanting to target a specific number of features provides flexibility of tuning. Given situations where the number of features is not large, RFE seems like a good choice.

LASSO compared to RFE includes a penalty term which means it might be better for optimizing for sparsity. It might also be a better choice given large numbers of features when compared to RFE as it doesn't have to recursively iterate to find the best features. Lasso does however require plugging in alpha values to test for the best alpha, which is not as intuitive.

BMI, BP, and S5 seem to be the most important features for predicting diabetes. Intuitively, body mass index and blood pressure should be associated with one's health, and one might even go as far as to assume that an increased BMI or BP would be a sign of worse health, so their association with diabetes is not surprising. S5 is meant to be serum triglycerides, which according to mayo clinic is a measure of a certain type of fat in the blood, so it'd make sense how it might be associated with diabetes. There is a blood sugar feature (s6), however it did not have a positive association with diabetes, which was a surprising find.