**cebd1160\_project\_template**

| **Name** | **Date** |
| --- | --- |
| Liliana Romero Molina | 2019-06-15 |

**Research Question**

The objective of this exercise is to determine whether, based on the information of the dataset (physiological characteristics and blood serum measurements), it is possible to predict the diabetes progression one year after a baseline measurement.

**Abstract**

We have a dataset containing 442 samples from patients with diabetes, with 10 measurements established one year after the baseline, along with their respective qualitative assessment of disease progression.

Using this information as a base, we could identify if these factors have relevance in the progression of the disease (if they are predictors), and from there establish mitigation methods on them in an early way.

**Introduction**

Diabetes dataset

Ten baseline variables, age, sex, body mass index, average blood pressure, and six blood serum measurements were obtained for each of n = 442 diabetes patients, as well as the response of interest, a quantitative measure of disease progression one year after baseline.

Note: Each of these 10 feature variables have been mean centered and scaled by the standard deviation times n\_samples(i.e. the sum of squares of each column totals 1).

<https://scikit-learn.org/stable/datasets/index.html#diabetes-dataset>

According to an additional source, the real 10 variables names are (for the exercise we are keeping the original file descriptions):

age sex bmi map tc ldl hdl tch ltg glu y

<https://www4.stat.ncsu.edu/~boos/var.select/diabetes.html>

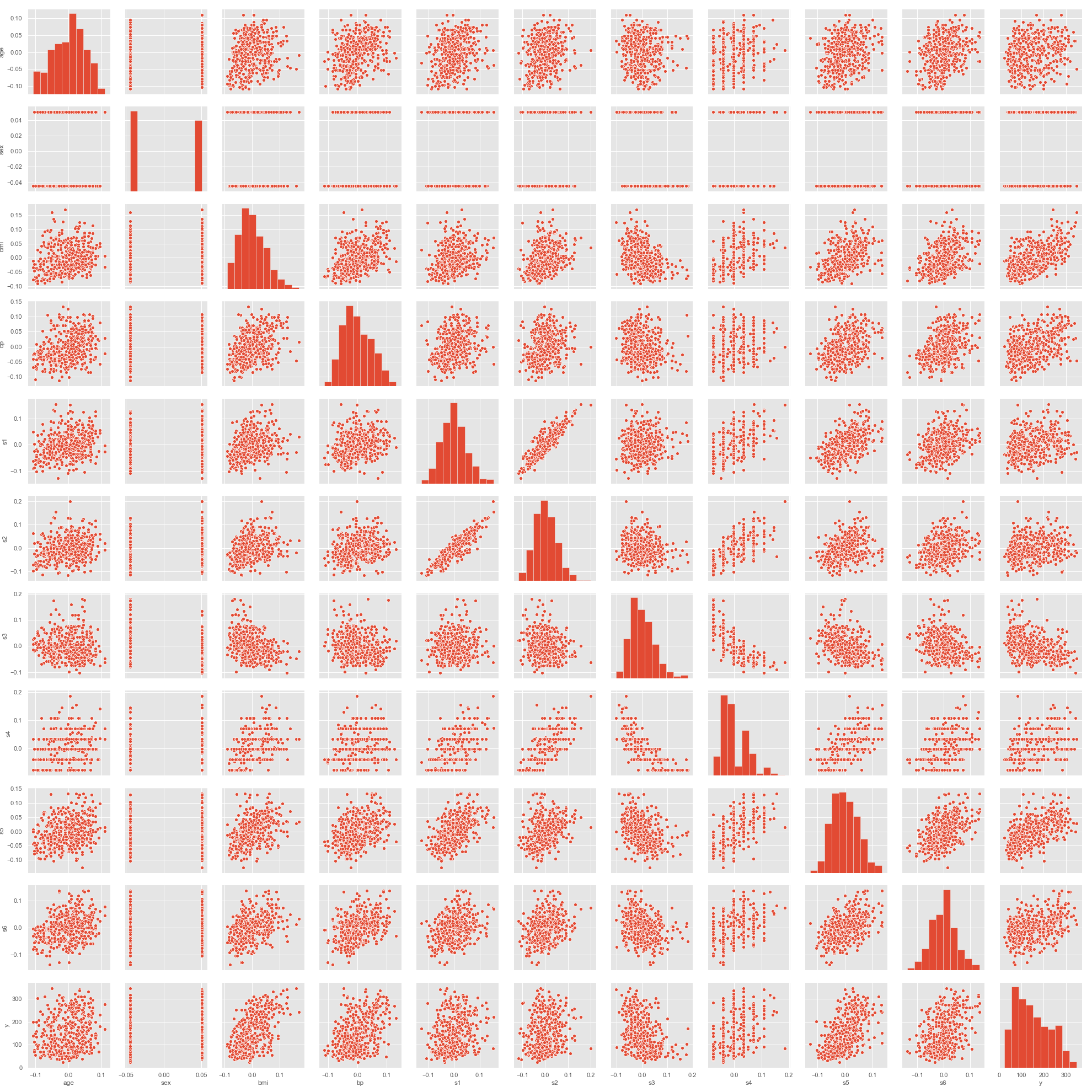
According to the initial revision, the information is complete, without null or missing data. In the same way, I do not see "atypical" data. It is understood that as a result of these steps the data is ready to start exploration and modeling.

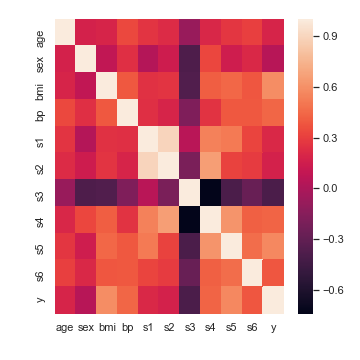
**Methods**

As drivers for defining the method, I took into account a visualization input (Heatmap) and the p-value of each of the independent variables obtained from an initial linear regression.

According to both observations, apparently only a handful of variables have a significant influence on the target variable, in this case the progression of diabetes.

Therefore, I opted to apply the Lasso model, understanding that it performs "eliminations" of non-significant variables within its process.



[](https://github.com/LilianaRomeroM/cebd1160_project_template/blob/master/diabetes7plots/heatseaborn.png)

OLS Regression Results

==============================================================================

Dep. Variable: y R-squared: 0.518

Model: OLS Adj. R-squared: 0.507

Method: Least Squares F-statistic: 46.27

Date: Fri, 14 Jun 2019 Prob (F-statistic): 3.83e-62

Time: 09:49:00 Log-Likelihood: -2386.0

No. Observations: 442 AIC: 4794.

Df Residuals: 431 BIC: 4839.

Df Model: 10

Covariance Type: nonrobust

==============================================================================

coef std err t P>|t| [0.025 0.975]

------------------------------------------------------------------------------

const 152.1335 2.576 59.061 0.000 147.071 157.196

x1 -10.0122 59.749 -0.168 0.867 -127.448 107.424

x2 -239.8191 61.222 -3.917 0.000 -360.151 -119.488

x3 519.8398 66.534 7.813 0.000 389.069 650.610

x4 324.3904 65.422 4.958 0.000 195.805 452.976

x5 -792.1842 416.684 -1.901 0.058 -1611.169 26.801

x6 476.7458 339.035 1.406 0.160 -189.621 1143.113

x7 101.0446 212.533 0.475 0.635 -316.685 518.774

x8 177.0642 161.476 1.097 0.273 -140.313 494.442

x9 751.2793 171.902 4.370 0.000 413.409 1089.150

x10 67.6254 65.984 1.025 0.306 -62.065 197.316

==============================================================================

Omnibus: 1.506 Durbin-Watson: 2.029

Prob(Omnibus): 0.471 Jarque-Bera (JB): 1.404

Skew: 0.017 Prob(JB): 0.496

Kurtosis: 2.726 Cond. No. 227.

==============================================================================

Warnings:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

“In the ideal world, we would be able to define a "perfectly" random sample, the most appropriate test and one definitive conclusion. We simply cannot. What we can do is try to optimise all stages of our research to minimise sources of uncertainty. When presenting P values some groups find it helpful to use the asterisk rating system as well as quoting the P value:

P < 0.05 \*

P < 0.01 \*\*

P < 0.001

Most authors refer to statistically significant as P < 0.05 and statistically highly significant as P < 0.001 (less than one in a thousand chance of being wrong).”

<https://www.statsdirect.com/help/basics/p_values.htm>

Variables with p-values < 0.05 (relationship to the response variable is signifcant): x2(sex), x3(bmi), x4(bp), x9(ltg).

**LASSO CV**

Lasso linear model with iterative fitting along a regularization path.

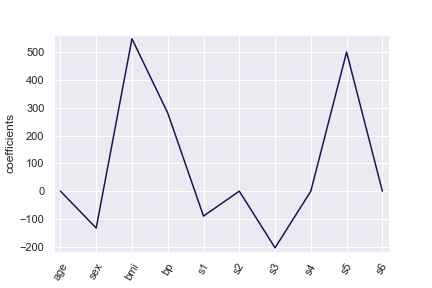
The best model is selected by cross-validation.

The optimization objective for Lasso is:

(1 / (2 \* n\_samples)) \* ||y - Xw||^2\_2 + alpha \* ||w||\_1

<https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LassoCV.html>

**Results**

[](https://github.com/LilianaRomeroM/cebd1160_project_template/blob/master/EXPERPLOTS/lassores.png)

Under the same test, the performance is:

LINEAR REGRESSION

MSE : 2973.0061322166034

RIDGE CROSS-VALIDATION alpha 0.0466301673441609 (in all cases low value, high bias, close to LR) MSE: 2864.841165528976

LASSO CROSS-VALIDATION (6 var)

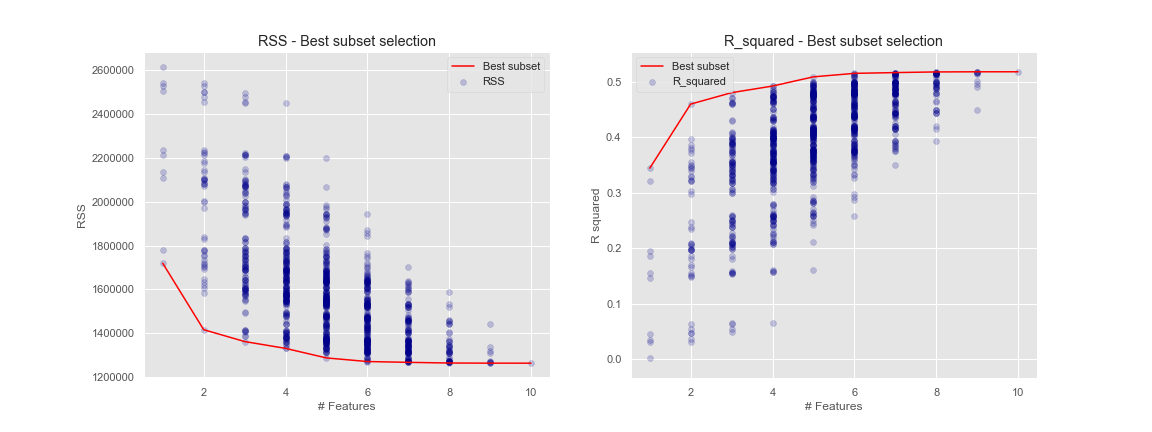
MSE: 2994.9584054969678

Scores <.55 for all cases

In general, the Ridge model generates an improvement although not very significant versus the linear regression.

However, since our initial exploration of data indicates that few variables have relevance to the target variable, in essence the Lasso method is able to deliver a more adequate/realistic result.

In order to perform a lateral validation, I separately developed a short subset exercise, observing that the reliability of the model "jumps" notoriously only with the passage from one variable (bmi) to 2(bmi-s5); this would indicate me that even finishing and assigning a subset, we would be approximating the same scenarios described above.

[](https://github.com/LilianaRomeroM/cebd1160_project_template/blob/master/SUBSETS/graph.png)

**Discussion**

I consider that the revised methods (especially Lasso) do not provide a reliable or statistically acceptable result.

In spite of this, I think that the Dataset does have valuable predictive information that can be explored through more elaborate, probably non-linear methods.

Another alternative study could focus on the body mass index, where surely there will be a better reading on the target variable, and only by initial observations, I believe that this same factor could be predictive with respect to other relevant variables (i.e. bp or s5)