Understanding Factors that Influence Diabetes in Adult Female Pima Indians

Aditya Pendyala

2023-11-09

Introduction:

This dataset "diabetes" has data from a study conducted by The National Institute of Diabetes and Digestive and Kidney Diseases, on 768 adult female Pima Indians living near Phoenix. The purpose of the study was to investigate factors related to diabetes. The dataset has 9 variables:
1. pregnant: number of times pregnant
2. glucose: plasma glucose concentration at 2 hours in an oral glucose tolerance
test 3. diastolic: diastolic blood pressure (mm Hg)
4. triceps: triceps skin fold thickness (mm)
5. insulin: 2-Hour serum insulin (mu U/ml)
6. bmi: body mass index (weight in kg/(height in meters squared))
7. diabetes: diabetes pedigree function
8. age: age (years)
9. test: test whether the patient shows signs of diabetes (coded 0 if negative, 1 if positive)

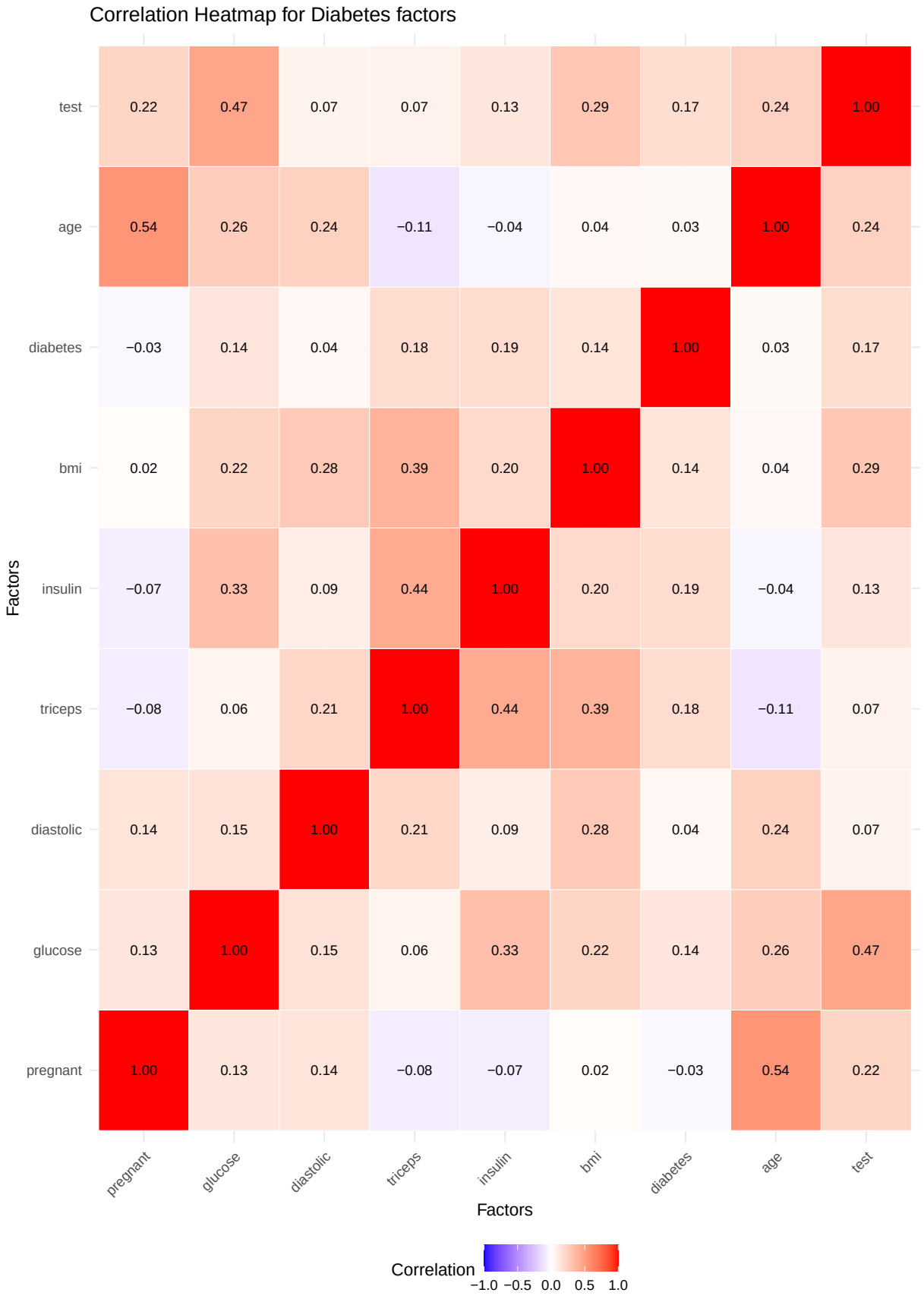
Purpose:

The goal of this project is to try and find variables that influence DPF(Diabetes Pedigree Function) in adult female Pima Indians. Furthermore, to formulate a prediction model, that predicts the DPF(Diabetes Pedigree Function) in an adult female Pima Indian.

Exploring the dataset:

We start by creating two data-sets for dependent variables and independent variable(DPF). We aim to find a prediction model for DPF which is the "diabetes" variable.

Here the "diabetes" variable is the value from the Diabetes Pedigree Function (DPF), which calculates diabetes likelihood depending on the subject's age and his/her diabetic family history.
First step in analyzing the dataset is to find the correlation of the factors. By doing this we can get a better understanding of which variables can be the possible factors of predicting DPF.


Correlation Heatmap for Diabetes factors
test 0.22 0.47 0.07 0.07 0.13 0.29 0.17 0.24 1.00
age 0.54 0.26 0.24 −0.11 −0.04 0.04 0.03 1.00 0.24
diabetes −0.03 0.14 0.04 0.18 0.19 0.14 1.00 0.03 0.17
bmi 0.02 0.22 0.28 0.39 0.20 1.00 0.14 0.04 0.29
insulin −0.07 0.33 0.09 0.44 1.00 0.20 0.19 −0.04 0.13
triceps −0.08 0.06 0.21 1.00 0.44 0.39 0.18 −0.11 0.07
diastolic 0.14 0.15 1.00 0.21 0.09 0.28 0.04 0.24 0.07
glucose 0.13 1.00 0.15 0.06 0.33 0.22 0.14 0.26 0.47
pregnant 1.00 0.13 0.14 −0.08 −0.07 0.02 −0.03 0.54 0.22
Factors
pregnant glucose diastolic triceps insulin bmi diabetes age test
Factors
Correlation
−1.0 −0.5 0.0 0.5 1.0

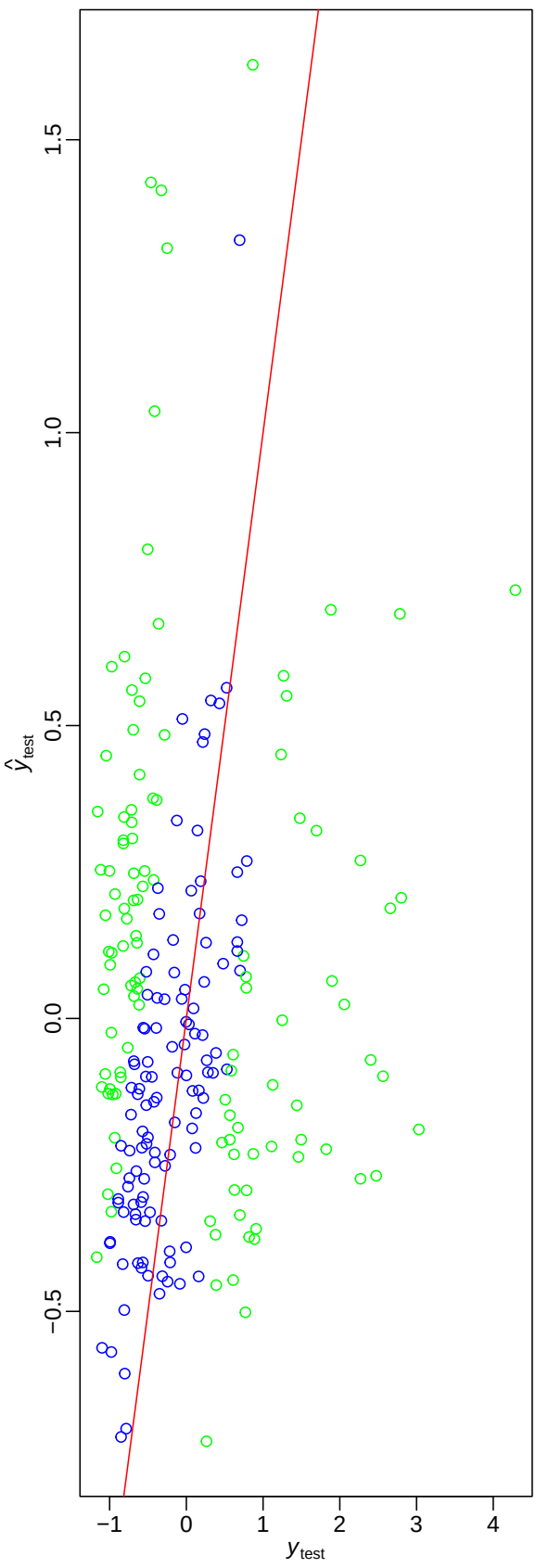
Looking at the heat map, we can see that 'diabetes' variable has some correlation with 'glucose', 'triceps', 'insulin', and 'bmi'. These are some variables that we can look out for while creating a prediction model for DPF.

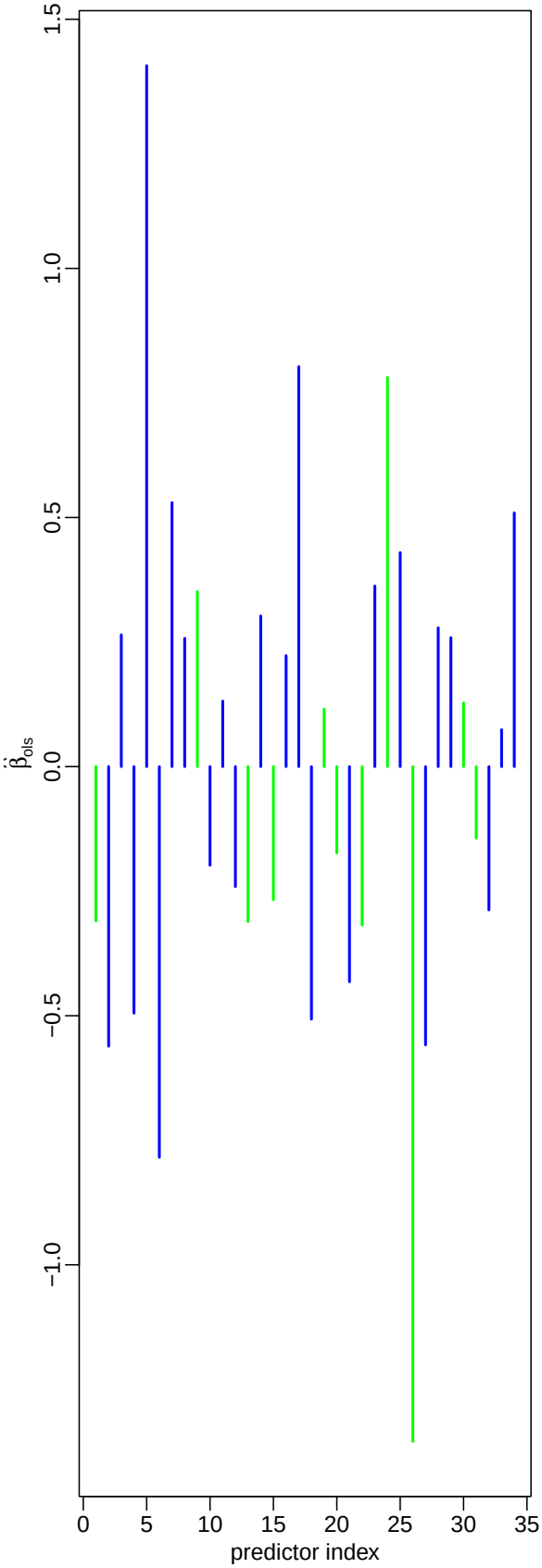
Prediction and analysis:

To continue with predictions and analysis, we standardize the data before regression and create new variables which are products of each variable with one another. We also include squares of each variable.

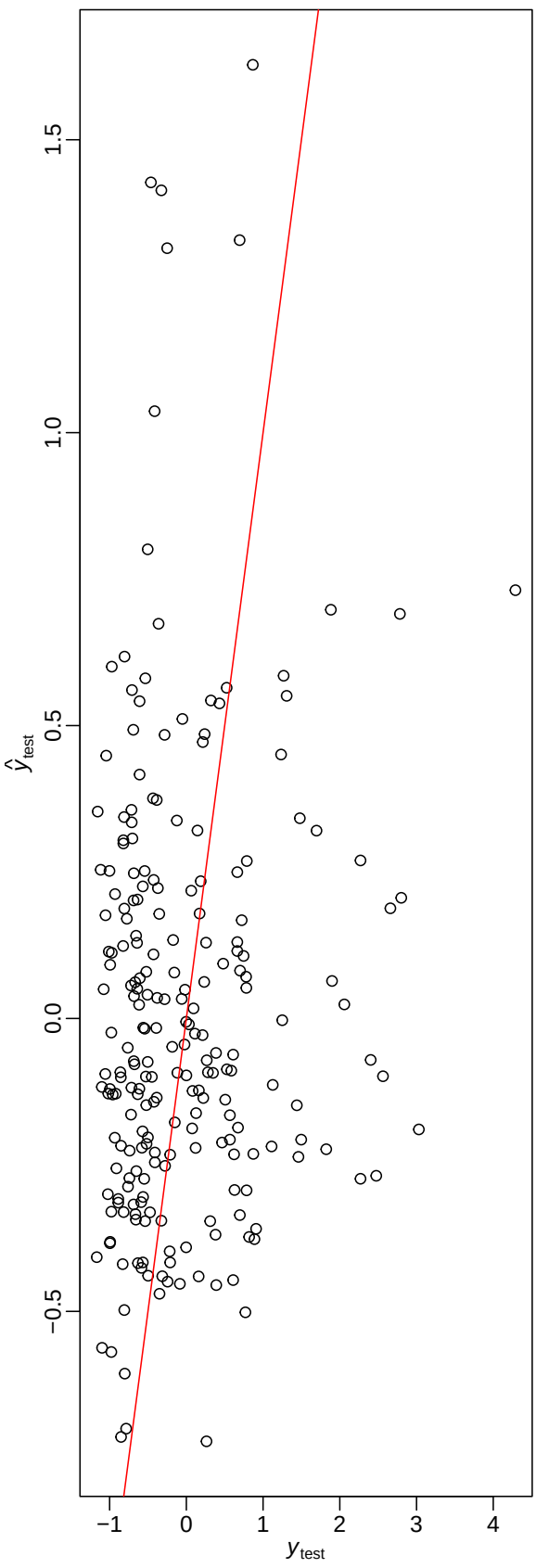
After creating the new data, we continue by standardizing the data and continue by splitting the standardized data into testing and training data-sets(30-70 ratio).

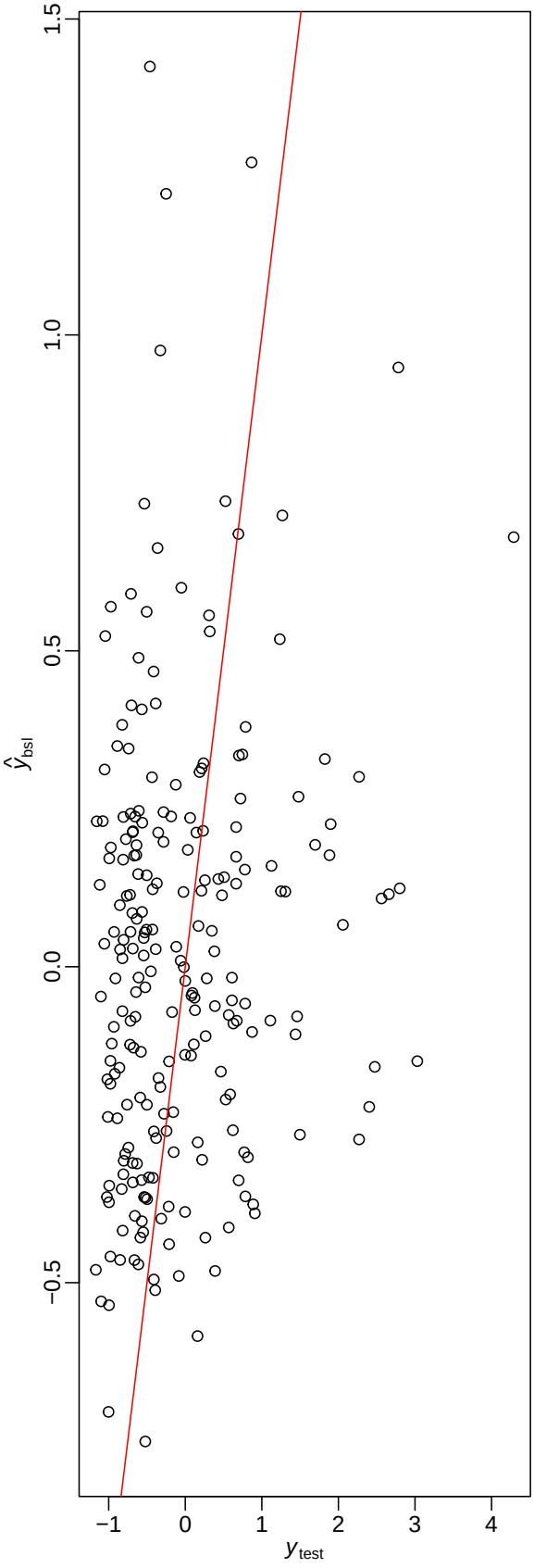
First prediction model that we consider is, OLS. We perform OLS for DPF vs all the dependent variables. We now plot y-test and estimated-y-test. We are also plotting the predictor index for variables.


1.5
1.0
0.5
0.0
−0.5
$\hat{y}_{\text{test}}$
−1
0
1
2
3
4
y_{test}


1.5
1.0
0.5
0.0
−0.5
−1.0
$\hat{\beta}_{\text{ols}}$
0
5
10
15
20
25
30
35
predictor index

Looking at the 2nd plot, it's visible that many variables may not be truly associated with DPF.
So, to remove variables that aren't helpful for the plot, we perform a backward elimination procedure.


1.5
1.0
0.5
0.0
−0.5
$\hat{y}_{\text{test}}$
−1
0
1
2
3
4
y_{test}


1.5
1.0
0.5
0.0
−0.5
$\hat{y}_{\text{bsl}}$
−1
0
1
2
3
4
y_{test}

```
## MSE for OLS: 0.9197586
```

```
## MSE for backward elimination model: 0.8701973
```

```
## Number of Variables that remain after backward elimination process: 9
```

```
## Number of Variables that remain after backward elimination process: 25
```

Looking at the results we can see that the mean square error for the OLS with all the variables is 0.92. After backward elimination, the mean square error is 0.87. This shows that the regression model's prediction performance has improved.

The variables that remain after the backward elimination process are:

```
## insulin
## bmi
## age
## glucose&bmi
## glucose&age
## triceps&bmi
## insulin&bmi
## insulin&age
## bmi&bmi
```

Now we consider a different prediction model, Bayesian Model Selection.
We are performing Bayesian Model Selection that finds the average beta regression variables. We perform this selection on the training data first and then predict values using the testing data. We are using regression_gprior.R file that contains the necessary functions to perform a Bayesian Model Selection using a g-prior.

```
##  [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

Looking at the average BETA values, we can see that none of the variables have a posterior probability greater than 0.5. This means that the Bayesian Model Selection doesn't select any variables. One of the reasons for this could be the very low correlation values of the variables.

Conclusion:

After comparing the different models, we can see that the backward elimination process helped develop the best performing prediction model. This prediction model uses "insulin", "bmi", "age", "glucose&bmi", "glucose&age", "triceps&bmi", "insulin&bmi", "insulin&age", "bmi&bmi" as it's predictor variables. These variables have a high influence on DPF, which helps determine the likelihood of diabetes in adult female Pima Indians. These findings are in-tune with other studies that mention most of these variables as factors for DPF.