Diabetes prediction

Authors:

Venkata Kanishk Chaganti (eh3762)

Pavan Kalyan Mashetti (gr6018)

Content

1.Introduction

Our body breaks down most of the food we eat into sugar (Glucose) and release it into bloodstream. When the sugar level goes up, it signals the pancreas to release insulin. Insulin acts like a key to let the blood sugar into your body's cells for use as energy. Diabetes is a chronic illness that affects how our body turns food into energy. Diabetes is mainly caused due to high glucose levels in a body. It may cause crucial problems like heart conditions, kidney problems, blood pressure, etc.

In this project we be focused on how different variables effect the Diabetes in a human body. It will also cover early prediction of diabetes in a patient with higher accuracy using some for the statistical methods like logistic regression and K-nearest neighbor.

2. Data Description

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage.

2.1. Data content

This dataset consists of 8 medical explanatory variables and a target variable.

Explanatory variables:

1) Pregnancies : Number of pregnancies
2) Glucose : Glucose levels in blood
3) BloodPressure : Blood pressure measurement
4) SkinThickness : Thickness of skin
5) Insulin : Insulin level in blood
6) BMI : Body mass index
7) DiabeticPedigreeFunction : Diabetes Percentage
8) Age : of a patient

Target variable:

Outcome : 1 if a person have diabetes and 0 if a person does not have diabetes.

In the dataset there are 768 observations of patients and 8 medical explanatory variables and 1 Outcome.

```
dim(diabetes)
```

```
## [1] 768   9
```

Figure: 2.1

3. Data Analysis

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0

Table: 3.1

Figure: 3.1 is a basic view of first 6 observations of a patient. We can observe that the variables do not have the units in which they are measured. We can also say that the data is not arranged properly.

	Pregnancies	Glucose_mg/dl	BloodPressure_mmHg	SkinThickness	Insulin_mL	BMI_W/H	DiabetesPedigreeFunction_%	Age	Outcome
1	1	89	66	23	94	28.1	0.167	21	0
2	1	73	50	10	0	23.0	0.248	21	0
3	2	84	0	0	0	0.0	0.304	21	0
4	1	80	55	0	0	19.1	0.258	21	0
5	2	142	82	18	64	24.7	0.761	21	0
6	0	125	96	0	0	22.5	0.262	21	0

Table:3.2

In table:3.1, by using rename() units for Glucose, Blood pressure, Insulin and BMI are been added to the variable names in the dataset. The data has been arranged by age of the patient.

```
## # A tibble: 2 x 16
##   Outcome count mean_Pregnancies SD_Pregnancies mean_Glucose SD_Glucose
##     <int> <int>            <dbl>          <dbl>        <dbl>      <dbl>
## 1       0   500             3.30           3.02         110.       26.1
## 2       1   268             4.87           3.74         141.       31.9
## # ... with 10 more variables: mean_BloodPressure <dbl>, SD_BloodPressure <dbl>,
## #   mean_SkinThickness <dbl>, SD_SkinThickness <dbl>, mean_Insulin <dbl>,
## #   SD_Insulin <dbl>, mean_BMI <dbl>, SD_BMI <dbl>,
## #   mean_DiabetesPedigreeFunction <dbl>, SD_DiabetesPedigreeFunction <dbl>
```

Table:3.3

Table:3.3 represents the mean, standard deviation and number of patients with and with out diabetes. There are total of 500 patients with diabetes and 268 patients without diabetes.

By observing the mean and standard deviation of variables we can say, medical explanatory variables of patients without diabetes have larger variability and average as compared to patients with diabetes.

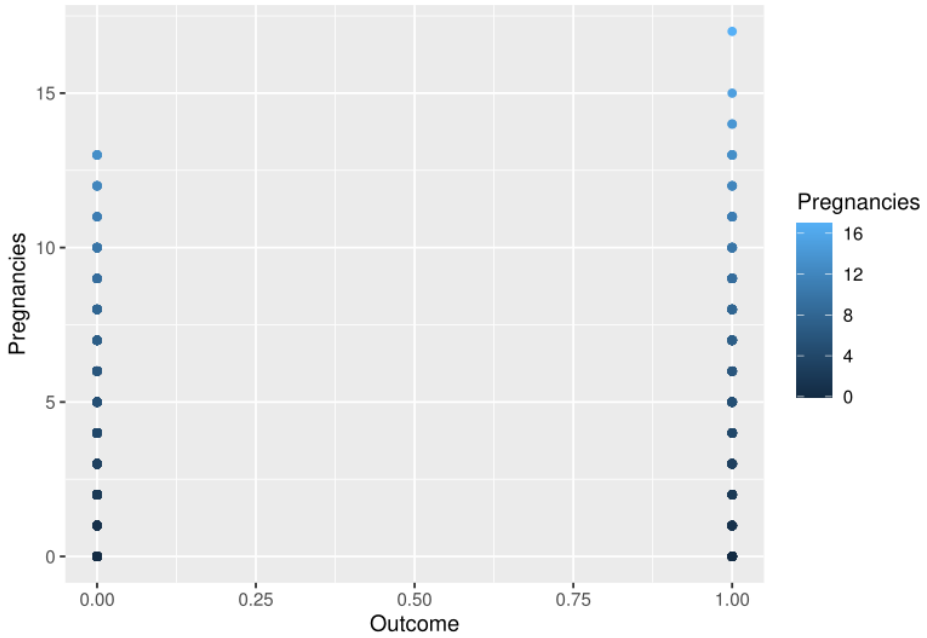

Figure:3.1

Figure: 3.1 represents patients with more the 13 pregnancies have high chance of getting diabetes.

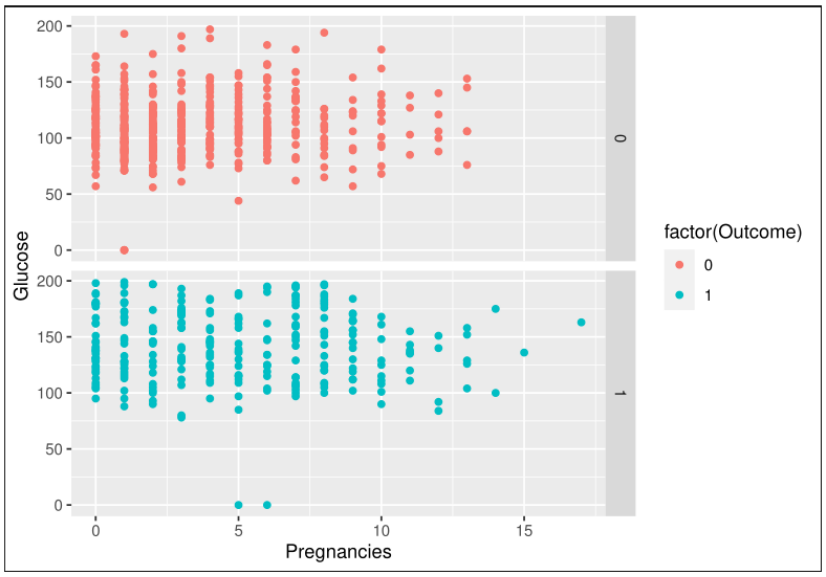

Figure: 3.2

Figure: 3.2 represents patients with diabetes and a smaller number of pregnancies have glucose levels more that 100mg/dl as compared to patients without diabetes.

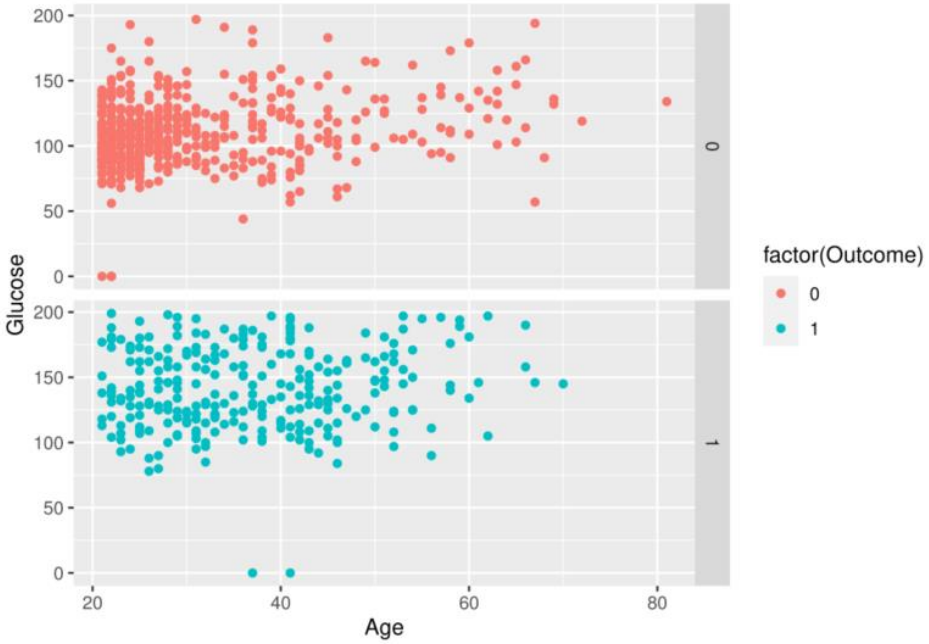

Figure: 3.3

Figure:3.3 represents that the patients with diabetes and age between 20 and 40 have greater Glucose level as compared to patients without diabetes.

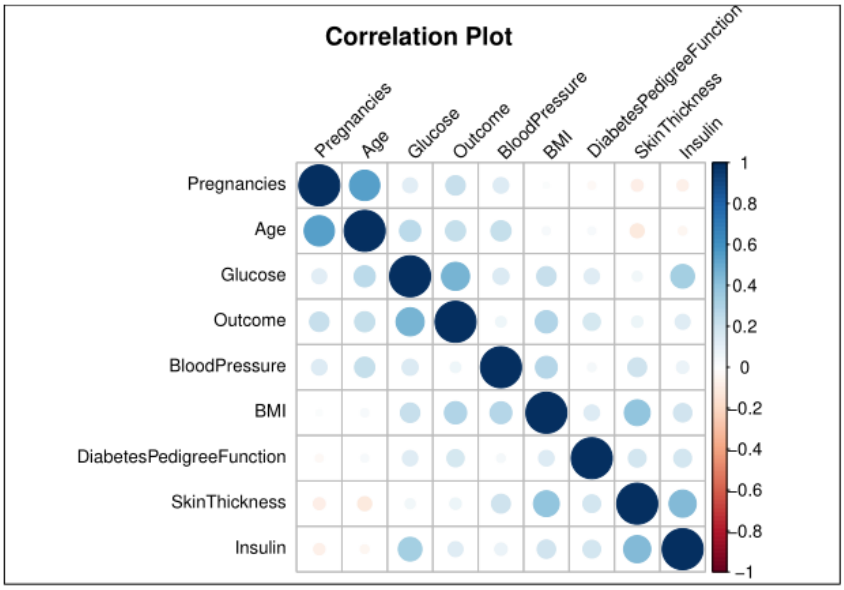

Figure:3.4

Correlation plot in Figure: 3.4 concludes that there is not much correlation between the variables.

4.Statistical Methods

4.1 Logistic Regression

```
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -7.954249   0.842148  -9.445  < 2e-16 ***
## Pregnancies              0.126841   0.037517   3.381 0.000723 ***
## Glucose                  0.031758   0.004341   7.315 2.57e-13 ***
## BloodPressure           -0.010548   0.005974  -1.766 0.077444 .
## SkinThickness            0.001865   0.008279   0.225 0.821742
## Insulin                 -0.001344   0.001068  -1.258 0.208296
## BMI                      0.092683   0.017440   5.314 1.07e-07 ***
## DiabetesPedigreeFunction 0.935424   0.345216   2.710 0.006735 **
## Age                      0.005734   0.010964   0.523 0.600975
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 700.47  on 539  degrees of freedom
## Residual deviance: 523.39  on 531  degrees of freedom
## AIC: 541.39
```

Figure: 4.1

```
##
## Coefficients:
##                           Estimate Std. Error z value Pr(>|z|)
## (Intercept)             -7.618032   0.793376  -9.602  < 2e-16 ***
## Pregnancies              0.140881   0.032719   4.306 1.66e-05 ***
## Glucose                  0.030248   0.003863   7.831 4.86e-15 ***
## BloodPressure           -0.010086   0.005721  -1.763  0.07791 .
## BMI                      0.089820   0.016399   5.477 4.32e-08 ***
## DiabetesPedigreeFunction 0.889556   0.339302   2.622  0.00875 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 700.47  on 539  degrees of freedom
## Residual deviance: 525.51  on 534  degrees of freedom
## AIC: 537.51
```

Figure:4.2

We first performed a full model logistic regression (model name: lm), the results are showed in Figure:4.1. In lm model we can observe that variables like SkinThickness, Insulin, Age have high p-values which means these variables are not contribution enough and we can just remove them from the model.

Using step() we built another model (lm1) removing the unnecessary variables. By comparing Akoike information criterion (AIC) values of both the models we can say that lm1 models has the lowest AIC value.

fitted.results	0	1
0	136	36
1	14	42

Figure: 4.3

Figure: 4.3 represents the confusion matrix. We observed that out of 228 observations 50 predicted values are wrongly predicted.

By above confusion matrix we can say that the accuracy of the model is **78%.**

5. Conclusion

We can say that people with higher Glucose level, Blood pressure, BMI and pregnancies tend to have diabetes and they need to take extra precautions to lead a healthy life. AS our data had only 768 observations, we acquired 78% of accuracy using logistic regression. If it was a larger dataset we might have obtained higher accuracy.

6.code appendix

Githud link: https://github.com/chvkanishk/STAT-650