Diabetes and related factors

ansari.shekoofeh@stud.hs-fresenius.de

07.06.2023

Source of dataset

The Diabetes dataset originates from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) in the United States. It was specifically created for research purposes, aiming to facilitate the development and evaluation of predictive models for diagnosing diabetes.

The dataset consists of records from 768 individuals, representing a diverse group of female patients aged 21 years and above, primarily of Pima Indian heritage. Each record includes a set of feature values and an associated diabetes outcome

Features

The Diabetes dataset comprises a range of numerical and categorical features relevant to predicting diabetes. These features include:

- **Pregnancies**: Number of times pregnant
- **Glucose**: Plasma glucose concentration (mg/dL)
- **BloodPressure**: Diastolic blood pressure (mm Hg)
- **SkinThickness**: Triceps skinfold thickness (mm)
- **Insulin**: 2-Hour serum insulin (mu U/ml)
- **BMI**: Body mass index (weight in kg/(height in m)^2)
- **DiabetesPedigreeFunction**: Diabetes pedigree function (a measure of genetic influence)
- **Age**: Age in years

Outcome: Diabetes outcome (0 for non-diabetic, 1 for diabetic)

Display

```
## 'data.frame':    768 obs. of  9 variables:
##  $ Pregnancies              : int  6 1 8 1 0 5 3 10 2 8 
##  $ Glucose                  : int  148 85 183 89 137 116 
##  $ BloodPressure            : int  72 66 64 66 40 74 50 0
##  $ SkinThickness            : int  35 29 0 23 35 0 32 0 4
##  $ Insulin                  : int  0 0 0 94 168 0 88 0 54
##  $ BMI                      : num  33.6 26.6 23.3 28.1 43
##  $ DiabetesPedigreeFunction: num  0.627 0.351 0.672 0.16
##  $ Age                      : int  50 31 32 21 33 30 26 2
##  $ Outcome                  : int  1 0 1 0 1 0 1 0 1 1 ..
```

Summary statistics

```
##   Pregnancies        Glucose      BloodPressure    SkinT
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Q
##  Median : 3.000   Median :117.0   Median : 72.00   Media
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean
##  3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Q
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.
##     Insulin          BMI        DiabetesPedigreeFunctio
##  Min.   :  0.0   Min.   : 0.00   Min.   :0.0780
##  1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437
##  Median : 30.5   Median :32.00   Median :0.3725
##  Mean   : 79.8   Mean   :31.99   Mean   :0.4719
##  3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262
##  Max.   :846.0   Max.   :67.10   Max.   :2.4200
##     Outcome
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
```

Diabetes

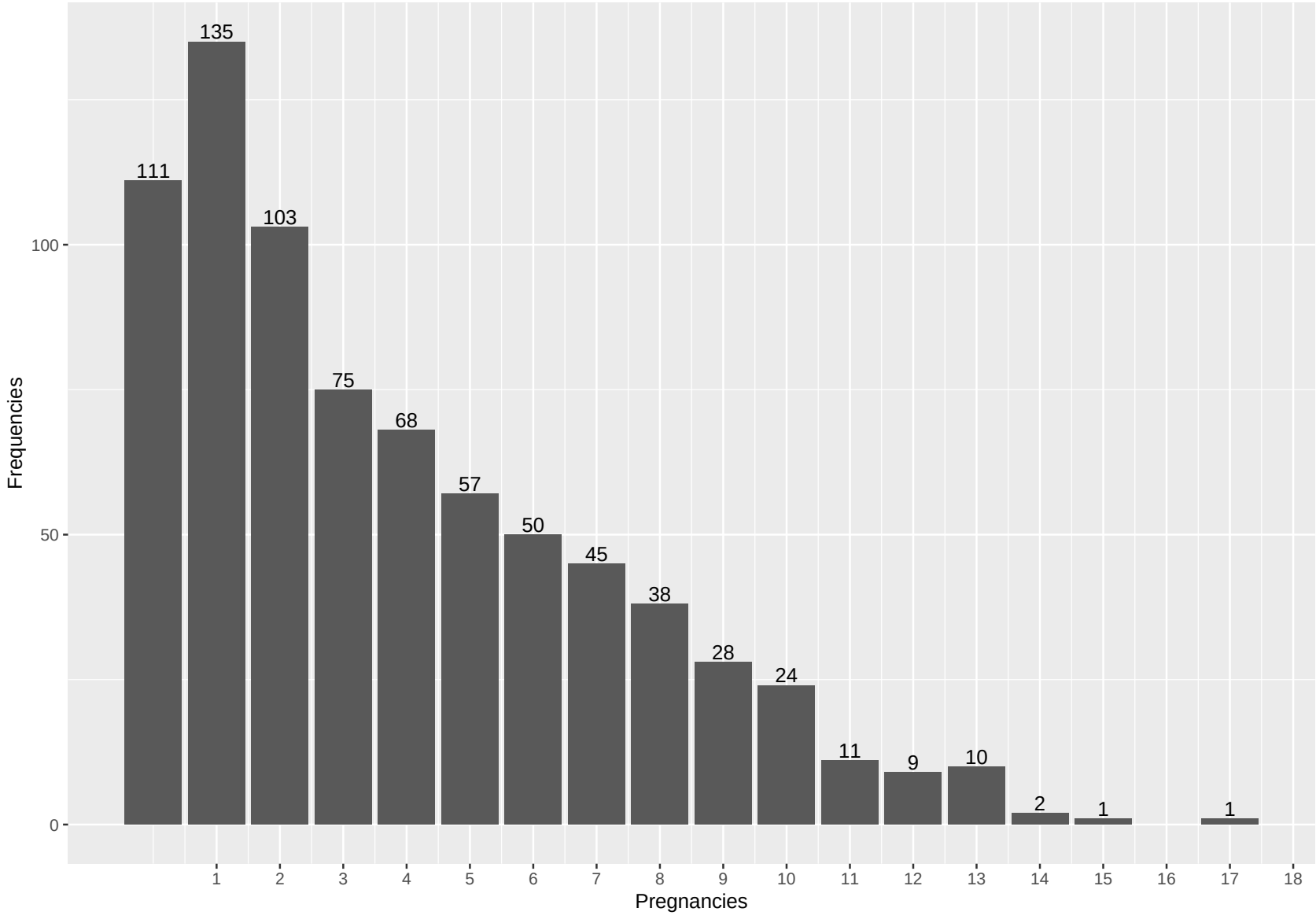
There are three main types of diabetes mellitus:

Type 1 diabetes: failure of the pancreas to *produce enough insulin* due to loss of beta cells

Type 2 diabetes: begins with insulin resistance, a condition in which cells *fail to respond to insulin* properly.

Gestational diabetes: occurs when pregnant women - without a previous history of diabetes develop high blood sugar levels

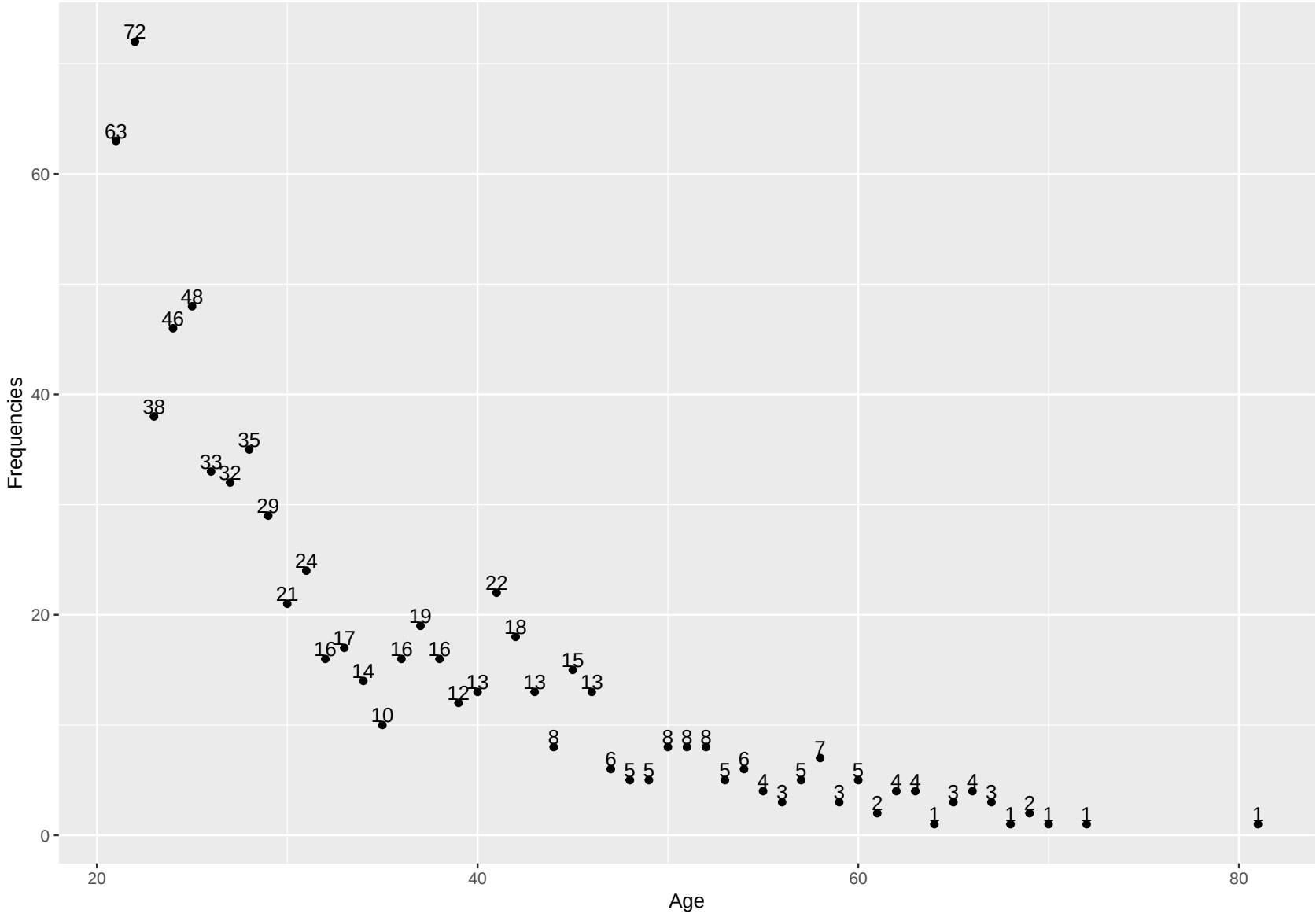
Pregnancy


Frequencies
100
50
0
111
135
103
75
68
57
50
45
38
28
24
11
9
10
2
1
1
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
Pregnancies

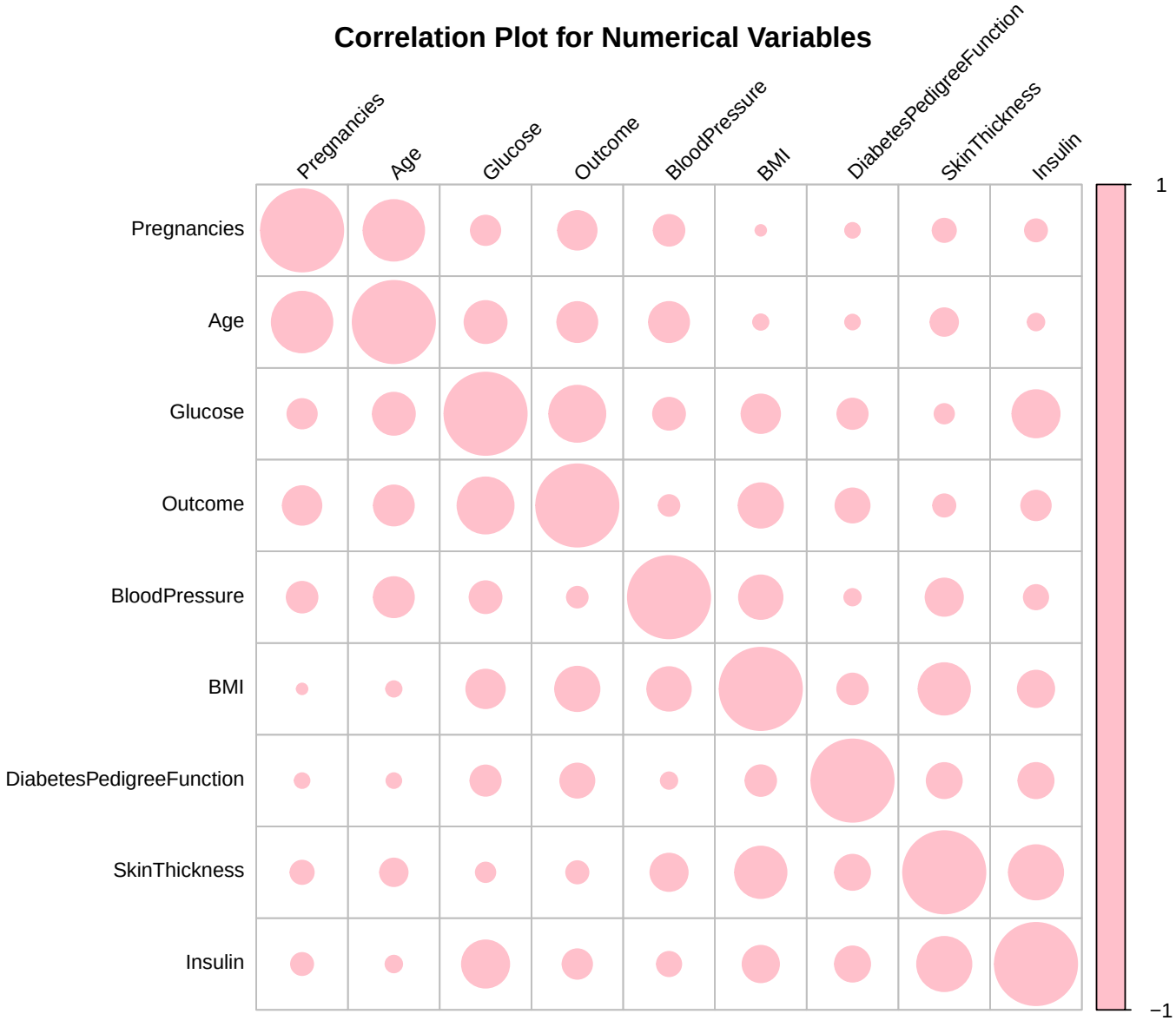
BMI

$$BMI = \frac{Weight_{(kg)}}{Height^2_{(m)}}$$

Ages


Frequencies
Age
0
20
40
60
20
40
60
80
72
63
48
46
38
35
33
32
29
24
21
22
19
18
17
16
16
16
15
14
13
13
13
12
10
8
8
8
8
7
6
6
5
5
5
5
4
4
4
4
3
3
3
3
2
2
1
1
1
1
1

Regression


Correlation Plot for Numerical Variables
Pregnancies
Age
Glucose
Outcome
BloodPressure
BMI
DiabetesPedigreeFunction
SkinThickness
Insulin
1
−1

Road map

1.We will collect the data.

2.Then , we'll import data set to read our data from a CSV file and manipulate it for further use.

3.We'll also use Knit to convert out data into a format suitable to feed our classification model.

4.We will use many packages for visualizations.

5.Then,we import logistic regression algorithm and cross validation.

6.Lastly, we have some conclusion about relation between diabetes and some factors.

Some references

1.https://towardsdatascience.com/end-to-end-data-science-example-predicting-diabetes-with-logistic-regression-db9bc88b4d16

2.https://www.cdc.gov/diabetes/basics/diabetes.html#:~:text=Diabetes to%20release%20insulin.

3.National Diabetes Data Group (US), National Institute of Diabetes and Kidney Diseases (US), 1995. Diabetes in America (No. 95). National Institutes of Health, National Institute of Diabetes and Digestive and Kidney Diseases.

4.Forouhi, N.G. and Wareham, N.J., 2010. Epidemiology of diabetes. Medicine, 38(11), pp.602-606.

5.https://r-graph-gallery.com/scatterplot.html

6.Disease Control, Centers for, and Prevention. 2022. "What Is Diabetes?" https://www.cdc.gov/diabetes/basics/diabetes.html#print.

THANKS FOR YOUR ATTENTION