Diabetes Project Report

Manuel Alexis Mena Nieves

6/23/2020

1. Introduction

This project report is about creating a model prediction system for the HarvardX Data Science professional certificate program, using the Indians Pima Diabetes Dataset, originally from the National Institute of Diabetes and Digestive and Kidney Diseases.

This dataset consists of eight medical predictor variables and one target variable, which shows if the patient has diabetes or not. Predictor variables includes the number of pregnancies the patient has had, their BMI, insulin level, age, skin thickness, glucose level, blood pressure and computed value called Diabetes Pedrigree Function.

The goal of the project is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset using all the tools shown throughout the courses in this series. To accomplish this an exploratory analysis was done, in order to understand the data and summarize their main characteristics with tables and visual methods. After this, a machine learning model and an ensemble model was created to predict whether or not the patients in the dataset have diabetes.

2. Getting the data

The following code will be used to download the dataset. We begin loading the tidyverse, caret, skimr and some useful machine learning libraries:

The dataset was uploaded into GitHub, thus we can access the data.

```
# Read the file
url <- paste0("https://raw.githubusercontent.com/alexismenanieves/",
              "Diabetes_Project/master/dataset.txt")
dataset <- read.csv(url)
```

In order to analyze the dataset, we see the dimensions, variable types and a summary.

```
# A first view of the data, dimensions and variables
dim(dataset)
```

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

```
as_tibble(dataset)
```

```
## # A tibble: 768 x 9
##    Pregnancies Glucose BloodPressure SkinThickness Insulin   BMI
##          <int>   <int>         <int>         <int>   <int> <dbl>
```

```
##  1           6     148            72            NA       0  NA
##  2           1      85            NA            29       0  26.6
##  3           8     183            64             0       0  23.3
##  4           1      89            66            23      94  28.1
##  5           0     137            40            35     168  43.1
##  6           5     116            74             0       0  25.6
##  7           3      NA            50            32      88  31
##  8          10     115             0             0       0  35.3
##  9           2     197            70            45     543  30.5
## 10           8     125            96             0       0   0
## # ... with 758 more rows, and 3 more variables: DiabetesPedigreeFunction <dbl>,
## #   Age <int>, Outcome <int>
```

```
summary(dataset)
```

```
##   Pregnancies        Glucose      BloodPressure    SkinThickness
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
##  Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.52
##  3rd Qu.: 6.000   3rd Qu.:140.0   3rd Qu.: 80.00   3rd Qu.:32.00
##  Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00
##                   NA's   :2       NA's   :1        NA's   :1
##     Insulin           BMI        DiabetesPedigreeFunction      Age
##  Min.   :  0.0   Min.   : 0.00   Min.   :0.0780           Min.   :21.00
##  1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437           1st Qu.:24.00
##  Median : 30.5   Median :32.00   Median :0.3725           Median :29.00
##  Mean   : 79.8   Mean   :31.99   Mean   :0.4719           Mean   :33.24
##  3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262           3rd Qu.:41.00
##  Max.   :846.0   Max.   :67.10   Max.   :2.4200           Max.   :81.00
##                  NA's   :1
##     Outcome
##  Min.   :0.000
##  1st Qu.:0.000
##  Median :0.000
##  Mean   :0.349
##  3rd Qu.:1.000
##  Max.   :1.000
##
```

```
dataset %>% group_by(Outcome) %>%
  summarise(count = n()) %>% mutate(freq = count/sum(count))
```

```
## `summarise()` ungrouping output (override with `.groups` argument)
```

```
## # A tibble: 2 x 3
##   Outcome count  freq
##     <int> <int> <dbl>
## 1       0   500 0.651
## 2       1   268 0.349
```

We can see that the dataset has 768 observations, 8 variables and 1 outcome. The outcome however is an integer, it's better to transform it into a factor. What draws our attention is that the outcome is imbalanced

```r
# Let's apply some changes on the outcome name and encoding
dataset$Outcome <- as.factor(ifelse(dataset$Outcome == 1,"Yes","No"))
names(dataset)[9]<- "Diabetes"
```

```r
# Let's divide the dataset into a train and test set
set.seed(1979)
tt_index <- createDataPartition(dataset$Age, times = 1, p = 0.9, list = FALSE)
train_set <- dataset[tt_index,]
test_set <- dataset[-tt_index,]
```

```r
# See how many observations and variables are available
str(train_set)
```

```
## 'data.frame':    693 obs. of  9 variables:
##  $ Pregnancies             : int  6 1 8 1 0 5 3 10 2 8 ...
##  $ Glucose                 : int  148 85 183 89 137 116 NA 115 197 125 ...
##  $ BloodPressure           : int  72 NA 64 66 40 74 50 0 70 96 ...
##  $ SkinThickness           : int  NA 29 0 23 35 0 32 0 45 0 ...
##  $ Insulin                 : int  0 0 0 94 168 0 88 0 543 0 ...
##  $ BMI                     : num  NA 26.6 23.3 28.1 43.1 25.6 31 35.3 30.5 0 ...
##  $ DiabetesPedigreeFunction: num  0.627 0.351 0.672 0.167 2.288 ...
##  $ Age                     : int  50 31 32 21 33 30 26 29 53 54 ...
##  $ Diabetes                : Factor w/ 2 levels "No","Yes": 2 1 2 1 2 1 2 1 2 2 ...
```

```r
# Glimpse of mean, median and NA's
summary(train_set)
```

```
##   Pregnancies        Glucose      BloodPressure    SkinThickness
##  Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00
##  1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00
##  Median : 3.000   Median :117.0   Median : 72.00   Median :23.00
##  Mean   : 3.857   Mean   :120.4   Mean   : 69.18   Mean   :20.23
##  3rd Qu.: 6.000   3rd Qu.:139.0   3rd Qu.: 80.00   3rd Qu.:32.00
##  Max.   :17.000   Max.   :198.0   Max.   :122.00   Max.   :63.00
##                   NA's   :2       NA's   :1        NA's   :1
##     Insulin           BMI        DiabetesPedigreeFunction      Age
##  Min.   :  0.00   Min.   : 0.00   Min.   :0.0780           Min.   :21.00
##  1st Qu.:  0.00   1st Qu.:27.10   1st Qu.:0.2400           1st Qu.:24.00
##  Median : 36.00   Median :32.00   Median :0.3660           Median :29.00
##  Mean   : 79.29   Mean   :31.82   Mean   :0.4686           Mean   :33.17
##  3rd Qu.:128.00   3rd Qu.:36.33   3rd Qu.:0.6260           3rd Qu.:41.00
##  Max.   :846.00   Max.   :67.10   Max.   :2.4200           Max.   :81.00
##                   NA's   :1
##  Diabetes
##  No :457
##  Yes:236
##
##
##
##
##
```