Diabetes Data Analysis

Sandra ROCHE

December 5, 2023

1 Introduction

From the World Health Organization (WHO) description, Diabetes is a chronic, metabolic disease characterized by elevated levels of blood glucose (or blood sugar), which leads over time to serious damage to the heart, blood vessels, eyes, kidneys and nerves. The most common is type 2 diabetes, usually in adults, which occurs when the body becomes resistant to insulin or doesn't make enough insulin. In the past 3 decades the prevalence of type 2 diabetes has risen dramatically in countries of all income levels. Type 1 diabetes, once known as juvenile diabetes or insulin-dependent diabetes, is a chronic condition in which the pancreas produces little or no insulin by itself. For people living with diabetes, access to affordable treatment, including insulin, is critical to their survival. There is a globally agreed target to halt the rise in diabetes and obesity by 2025.

About 422 million people worldwide have diabetes, the majority living in low-and middle-income countries, and 1.5 million deaths are directly attributed to diabetes each year. Both the number of cases and the prevalence of diabetes have been steadily increasing over the past few decades.

Symptoms of type 1 diabetes include the need to urinate often, thirst, constant hunger, weight loss, vision changes and fatigue. These symptoms may occur suddenly. Symptoms for type 2 diabetes are generally similar to those of type 1 diabetes but are often less marked. As a result, the disease may be diagnosed several years after onset, after complications have already arisen. For this reason, it is important to be aware of risk factors.

Type 1 diabetes cannot currently be prevented. Effective approaches are available to prevent type 2 diabetes and to prevent the complications and premature death that can result from all types of diabetes. These include policies and practices across whole populations and within specific settings (school, home, workplace) that contribute to good health for everyone, regardless of whether they have diabetes, such as exercising regularly, eating healthily, avoiding smoking, and controlling blood pressure and lipids.

Does the data set contain parameters directly linked to the Diabetes diagnostic ?

2 Methodology

2.1 Data set

The data were collected from the Iraqi society, as they data were acquired from the laboratory of Medical City Hospital and (the Specializes Center for Endocrinology and Diabetes-Al-Kindy Teaching Hospital).

Patients' files were taken and data extracted from them and entered in to the database to construct the diabetes data set. The data consist of medical information and laboratory analysis.

The data set is provided by the Kaggle website https://www.kaggle.com/datasets/aravindpcoder/diabetes-dataset/data. It contains 14 variables for 1000 patients like age, gender, cholesterol rate and so on.

2.2 Variables

Several parameters (N=14) are found in this data set : ID, No_Pation, Gender, AGE, Urea, Cr, HbA1c, Chol, TG, HDL, LDL, VLDL, BMI, CLASS; and some of them are explained in the Table 1.

Variable	Explanation
BMI	Body Mass Index
Chol	Cholesterol
Cr	Creatinine rate
TG	Triglycerides
HbA1c	Haemoglobin A1c (blood marker)
LDL	”bad” cholesterol
HDL	”good” cholesterol
CLASS	Diabetes diagnostic

Table 1: Variables explanation

2.3 Method

Data set is imported from the Kaggle website which contains a lot of data on multiple thematic. At the beginning, data are analyzed from a quality point of view including search of missing value, duplicate and typography issue. Then, some variable distributions are observed.

Next, an analyse is performed to find linear correlation between all numerical variables and some statistical tests are also executed to find which variables are linked to the Diabetes diagnostic.

Finally, limitations and discussion are proposed at the end of the report and a comparison is performed with what is written in the literature .

3 Results

The data set contains no missing value (N=1000) and no duplicate data. Some typography issues have been identified and corrected in gender and Diabetes diagnostic variables.

Data distribution for the age, the gender and the Diabetes diagnostic is done to observe bias in the data set.

A total of 1000 patients aged from 20 to 79 years are involved in the study. Figure 1 shows that 74% of the data are coming from patients between 50 and 63 years old. Thus, a bias could be introduced by the age.

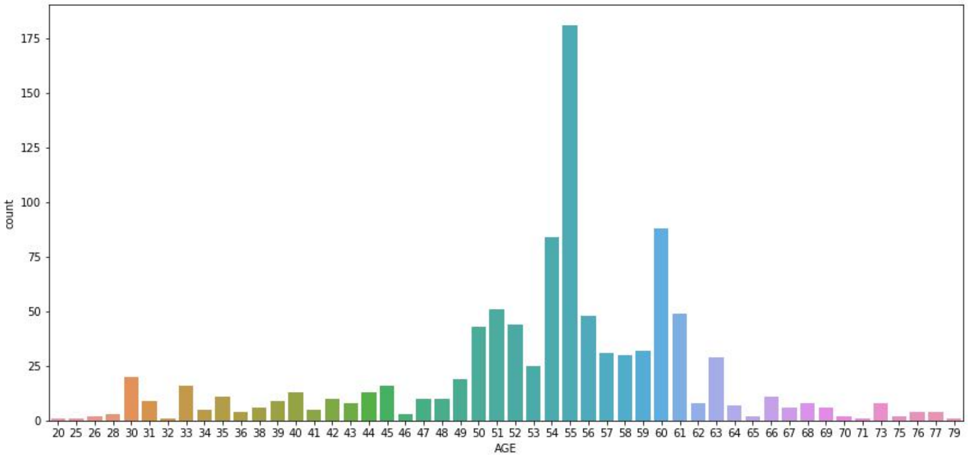

Figure 1: Data distribution for the age

The Diabetes diagnostic (CLASS variable) is shared out in ”Non-Diabetic” (N), ”Predicted Diabetic” (P) and ”May be Diabetic” (Y) categories. Figure 2 shows that majority of the Diabetes diagnostic data are classified as ”May be Diabetic”.

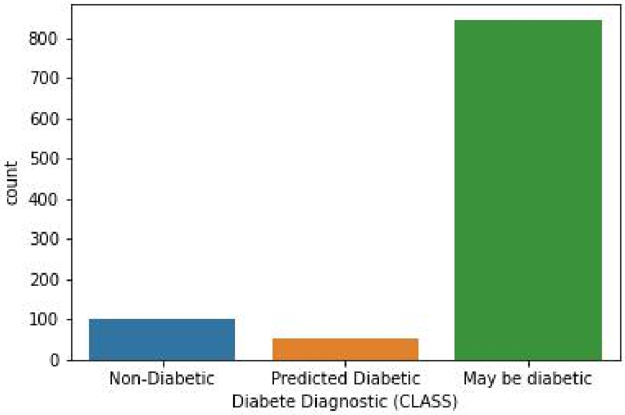

Figure 2: Data distribution for the Diabetes diagnostic

Moreover, both genders are well and fair represented: 43.5% from women and 56.5% from men.

Then, linear correlation are evaluated between numerical variables and Figure 3 summarizes results. Urea and creatinine ratio (Cr) appear to be linked and that's confirmed by a Chi2 test (chi2(12208) = 33906 (p-value $\leq$ 0.01)). A smaller correlation is observed between age, BMI and the HbA1c blood marker. Furthermore, HbA1c seems to be linked to age, cholesterol (Chol), triglycerides (TG) and BMI. Moreover, the cholesterol (Chol) seems to be linked to triglycerides (TG) and "good" cholesterol (HDL).

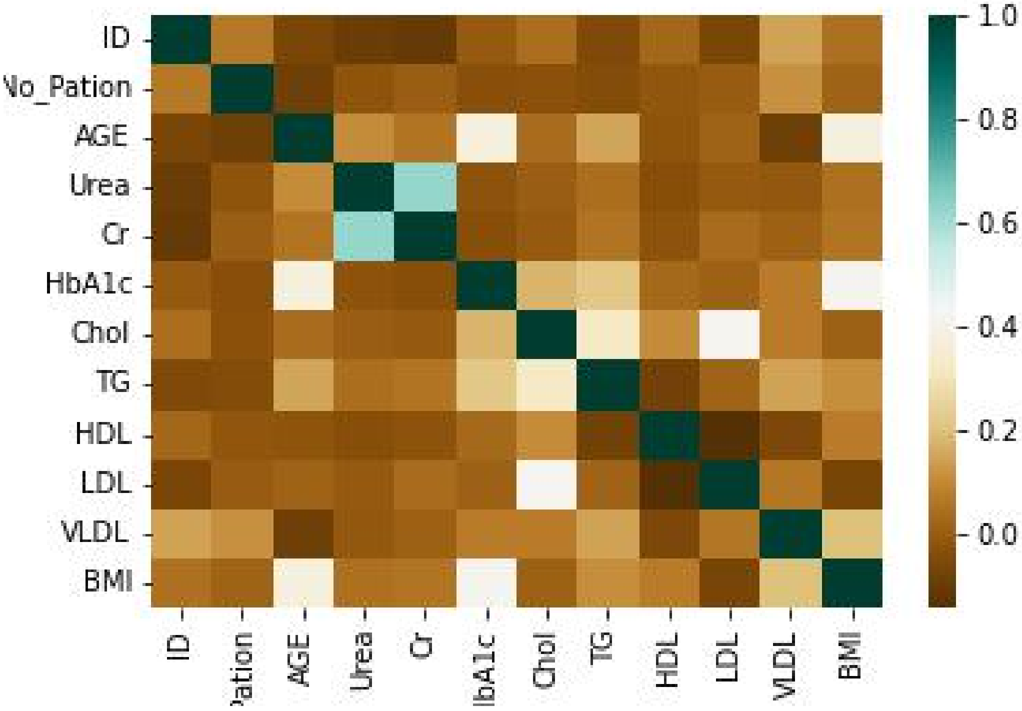

Figure 3: Linear correlation between variables

Correlation between these numerical variables is also studied from a statistical point of view and Figure 4 summarized results. Variables BMI, age, cholesterol (Chol) and triglycerides (TG) are positively and significantly linked to the HbA1 blood marker. Variables VLDL, Creatinine ratio (Cr), Urea, LDL and HDL are positively linked to the HbA1 but it's not statistically significant.

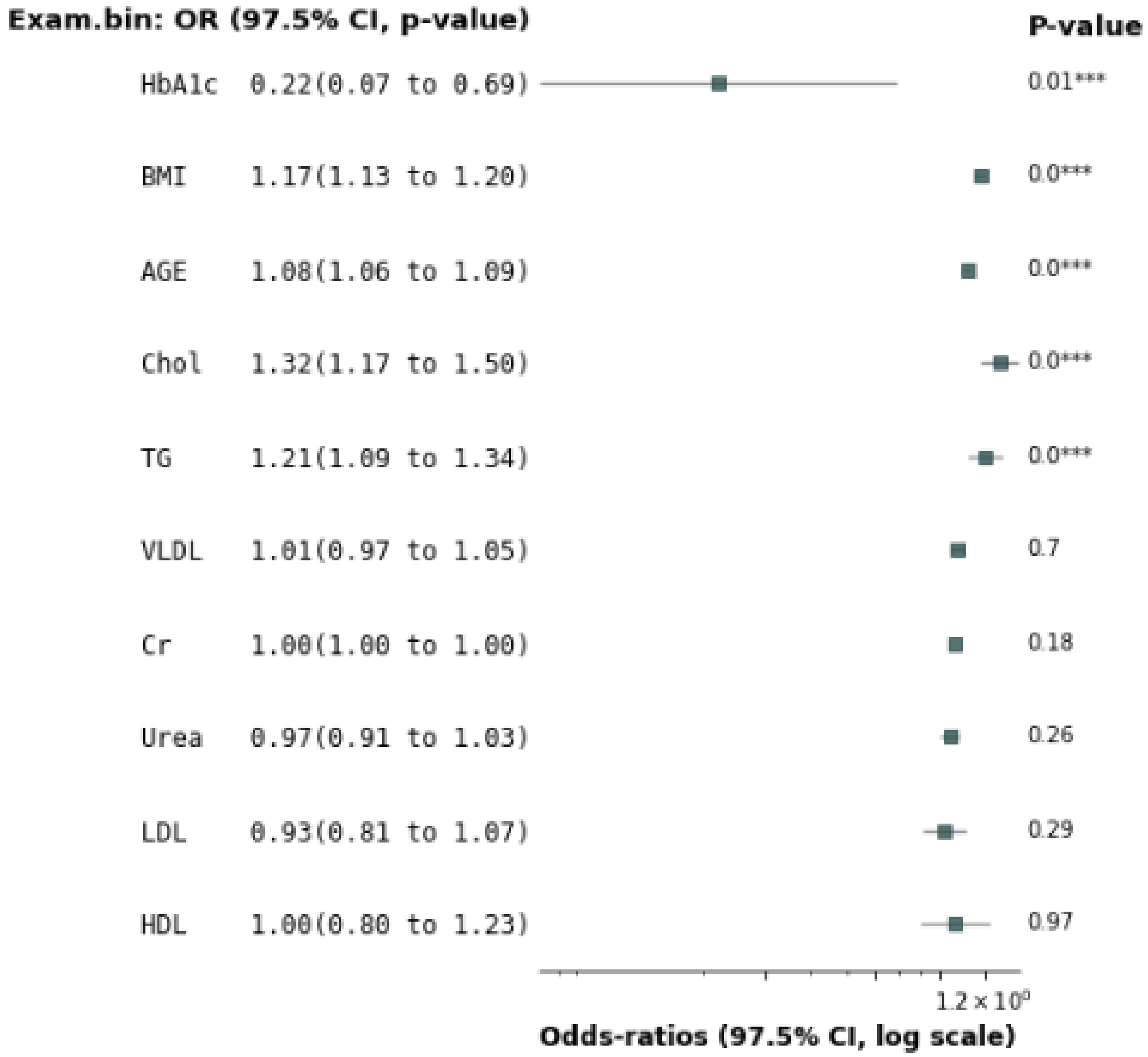

Figure 4: Forestplot representing correlation between numerical variables

To continue the analyse, variables linked to the Diabetes diagnostic are searched. As illustrated in the Figure 5, Urea and Creatinine ratio (Cr) are linked together but they are not linked to the Diabetes diagnostic.

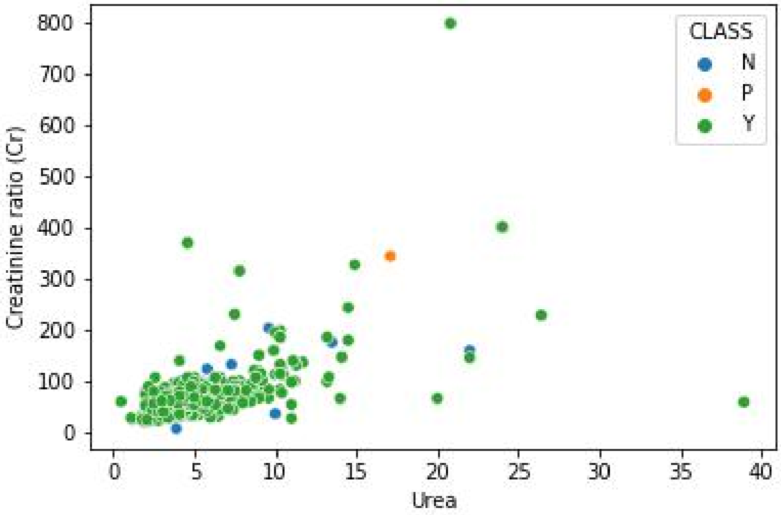

Figure 5: Correlation between urea, creatinine ratio and Diabetes diagnostic

On the other side, Figure 6 shows that the HbA1c blood marker increase when the Diabetes diagnostic is "May be diagonstic" compared to "Non-Diabetc" and "Predicted Diabetic" is in the middle (chi2(220) = 1360 (p-value ≤ 0.01)). Moreover, no link is observed between the gender and the Diabetes diagnostic.

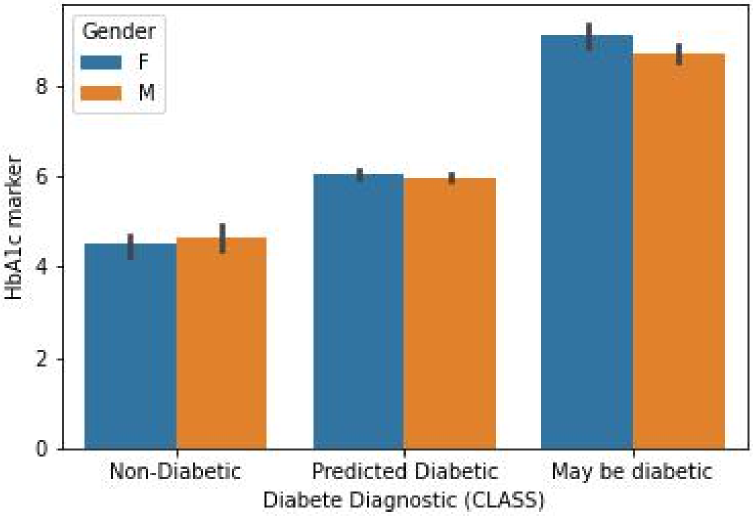

Figure 6: Diabetes diagnostic and HbA1c marker per gender

Another analyze illustrates in the Figure 7 shows that BMI ≥ 25 and HbA1c ≥ 6 seem to lead to the "May be Diabetic". Then BMI and Diabetes diagnostic seem to be linked (chi2(126) = 819 (p-value ≤ 0.01)).

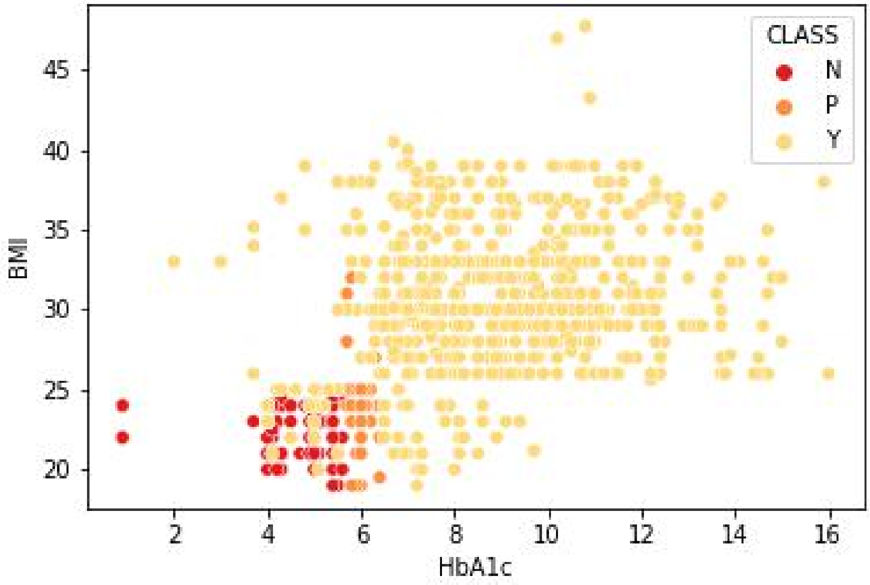

Figure 7: HbA1c and BMI per Diabetes diagnostic

4 Discussion

Data set contains heterogeneous data for age and Diabetes diagnostic introducing a bias in the analysis. More data from young people with and without Diabetes diagnostic and more data from people without Diabetes diagnostic aged more than 50 years may help to compare parameters link to Diabetes diagnostic, thus, conclusion might be easier and more secure.

In all cases, high level of HbA1c appears to be linked to ”May be Diabetic”. In the literature, this blood marker is effectively described as a link to the Diabetes diagnostic. In 2011, the WHO decided to accept the use of HbA1c testing in diagnosing diabetes and an HbA1c of 48mmol/mol (6.5%) is recommended as the cut off point for diagnosing diabetes. This is confirmed by the present study (Figures 6 and 7). However, HbA1c is also described to be not appropriated for diagnosis of diabetes for children and young people.