Diabetes in Pima Indians based on various variables

Loizos Vlasiadis AM: 685

Department of Computer Science and Biomedical Informatics, University of Thessaly, Lamia, 35100, Greece

∗Correspondence to: lvlasiadis@uth.gr

***Abstract*— One of the most common diseases in Pima Indians is diabetes. Diabetes mellitus is a metabolic disorder which occurs when the blood glucose level is high. This brief study focuses on to determine whether there is a correlation between diabetes in Pima Indian women and various medical variables such as: Age, Number of Pregnacies, their Body Mass Index (BMI), Insulin level, Glucose level, Blood Pressure, Skin Thickness, Diabetes Pedigree Function and a target variable which is Outcome, whether they suffer from Diabetes or not. The dataset is taken from the National Institute of Diabetes and Digestive and Kidney Diseases.**

I. INTRODUCTION

The Pima Indians or Pima People are a tribe of Native Americans who locates at the central Arizona, USA. Their name translates as River People.[1]

According to [2] during the era of missionaries and the Spanish occupation a lot of new diseases appeared to the Native Americans. Besides the germs that the European conquerors brought with them, they changed the traditional way of life of the natives, with new crops but also new customs and traditions in order to access a more "white way of life".

Some of the biggest changes in the way of life of the Native Americans were the seizure of their lands and their restriction to camps known as Indian reservations[3], as well as the introduction of alcohol to their daily diet and the increasing levels of obesity among the Indigenous people. [4][5]

The cutting off of the Indigenous people from their traditional pastures but also the abandonment of their nomadic way of life with the inclusion in the camps had significant effects on them. Their quality of living has fallen and to this day unemployment rates are very high among Indigenous peoples. On some reservations the quality of life is comparable to countries of the so called developing world: low life-expectancy, poverty, bad nutrition and alcohol abuse.

Based on the above the Native Americans turned to gambling as an alternative way of income and based on certain federal laws to the creation of casinos inside their reservations.[6]

There are studies investigating whether it is a correlation between gambling and alcohol abuse among the Native Americans.[7] While the myth of the ''firewater'' has been debunked [8][9] the number of Indigenous people who have faced an alcohol abuse associated problem is higher compared to the Caucasian people.[10]

A. Diabetes Mellitus

Diabetes Mellitus, also known as Diabetes, is a disease which affects the human metabolism resulting in an increase in the percentage of blood sugar and a disorder of glucose metabolism. The main types of diabetes are type 1 and type2 and gestational diabetes which is about pregnant women who have high levels of blood glucose during their pregnancy although their blood glucose levels used to be normal before being pregnant.[11][12]

Insulin is a hormone which is produced by the pancreas and it is responsible for keeping stable the levels of blood sugar by metabolizing carbohydrates, lipids and proteins.[13]

By consuming food, insulin transports glucose to the cells for energy production. However, if you have diabetes, your body cannot break down glucose into energy.[12]

Diabetes occurs when the pancreas does not produce enough insulin or when the body does not respond to the insulin it produces.

The main symptoms of diabetes among the others are: unintended weight loss, night urination, increased thirst and hunger, the feeling of being constantly tired and wounds that take a long period to heal.

Diabetes is a chronic disease and can cause a number of serious complications such as: cardiovascular disease, chronic kidney failure, retinal damage, nerve damage, erectile dysfunction.

Diagnosis of diabetes is easy when there are classic symptoms and it is enough to confirm it by measuring blood sugar. A number of factors, such as smoking, high cholesterol levels, obesity, high blood pressure and sedentary lifestyle can accelerate complications.[14]

Once a person is diagnosed with diabetes should change and adjust its diet, start exercising and using the appropriate medication, more frequently subscribed insulin.

II. MATERIALS AND METHODS

The dataset on which this survey is based on was acquired from the Kaggle.com platform and originally from the National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK).[15]

The NIDDK is part of the National Institutes of Health (NIH), a United States of America medical research agency. Among its roles is to conduct and support biomedical research and health information to the public.[16]

All the participants in the dataset are female, twenty-one (21) years old at least and of Pima Indian heritage.
This specific dataset can be used to predict whether a female patient has diabetes based on certain diagnostic factors.

The dataset consists of total nine (9) variables from which eight (8) of them are used as measurements to diagnose whether there is a correlation between these variables and the ninth (9th) variable, the Outcome. The Outcome represents whether a female participant has diabetes or not.

The other eight variables are:

- Pregnancies, the number of times a female was pregnant
- Glucose, Plasma glucose concentration after 2 hours in an oral glucose tolerance test
- Blood Pressure, Diastolic blood pressure (mm Hg)
- Skin Thickness, Skin fold thickness (mm)
- Insulin, 2-Hour serum insulin (mu U/ml)
- BMI, Body Mass Index.
- DiabetesPedigreeFunction, is a function which scores likelihood of diabetes based on family history
- Age, age of the participants (years)

The data was analyzed for frequency distributions and during this brief survey the following software programs were used:

- Anaconda Navigator 1.9.12
- Jupyter Notebook 6.0.3
- Spyder 4.0.1

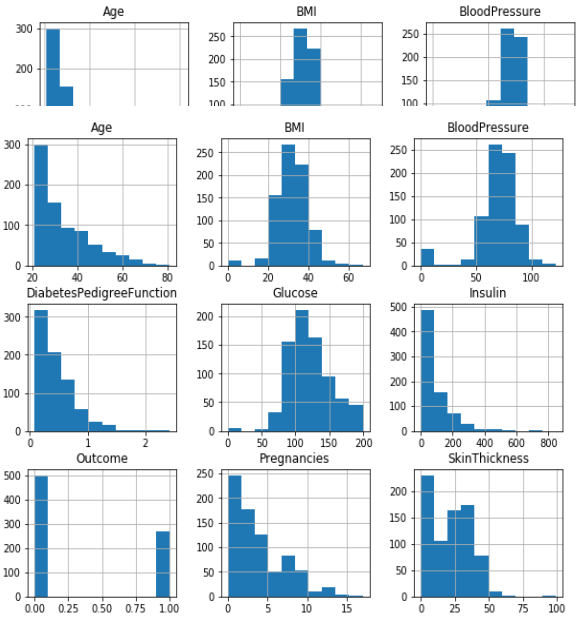

Figure 1 All nine (9) variables and features

III. RESULTS

There were seven hundred sixty eight (768) female Pima Indian people who participated in the research and a total of nine (9) variables and the basic characteristics can be shown above on Figure1.

The basic characteristics can be shown above on Figure1. Looking more thoroughly on the various histograms of the variables of our dataset we can make some observations.

We notice that the Age variable does not follow a normal

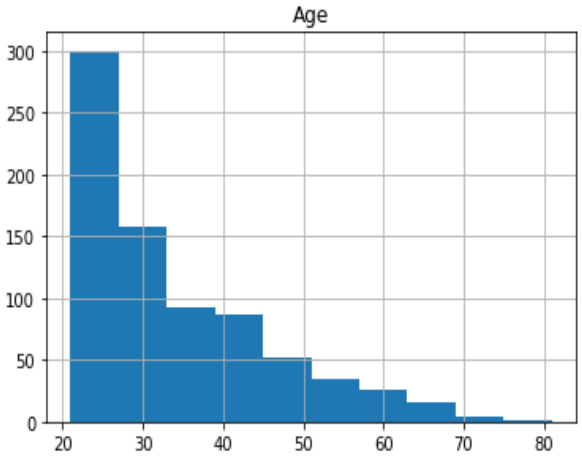

Figure 2 Age

distribution as we may expected. On the other hand, the values we would say that are quite normal as the minimum age is 21 and the maximum 81.

On the other hand, the variable BloodPressure seems to follow aquite normal distribution.

On some histograms we observe some interesting values i.e.

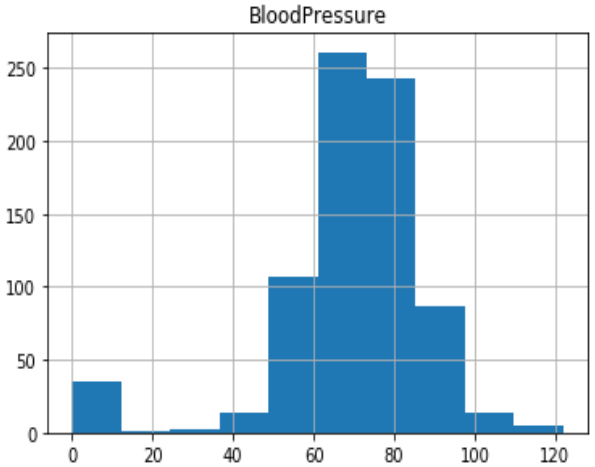

Figure 3 BloodPressure

Body mass index (BMI) “Fig. 4”, BloodPressure, Glucose concentration “Fig. 5”, Insulin and skinfold Thickness which are zero (0) and this measure does not make sense.

The count of zeros in the various variables is being presented on Figure 6.

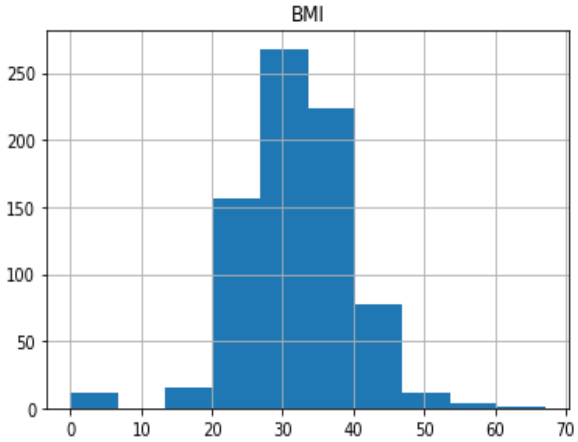

Figure 4 BMI

Zero (0) values for Glucose concentration are quite unusual for a living creature.

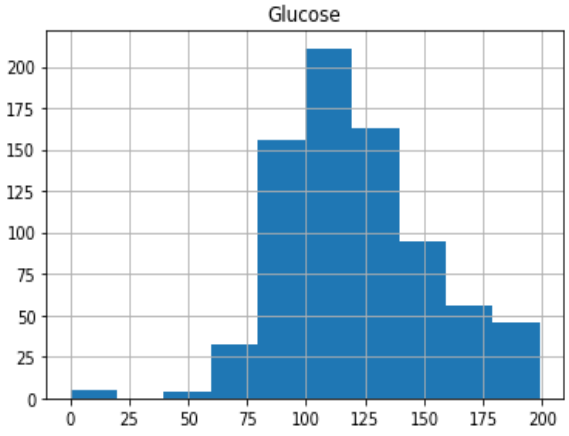

Figure 5 Glucose

Below on Fig. 6 we observe that only on the variable Pregnancies we expect to find zero (0) values as it is quite possible for a woman not to give birth to a child.

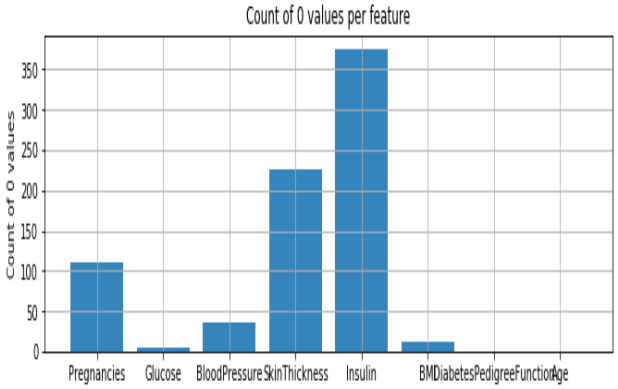

Figure 6 Count of Zeros

Based on the measures of Fig. 6, we assumed that there were data missing, perhaps the participants did not answer, and those missing values were replaces by zeros.

On the next chart is a boxplot which give us an indication of how much the values in our dataset are spread as well as the skewness of the data.[17]

Above on Fig.7 we observe that the boxplots on Age, Glucose

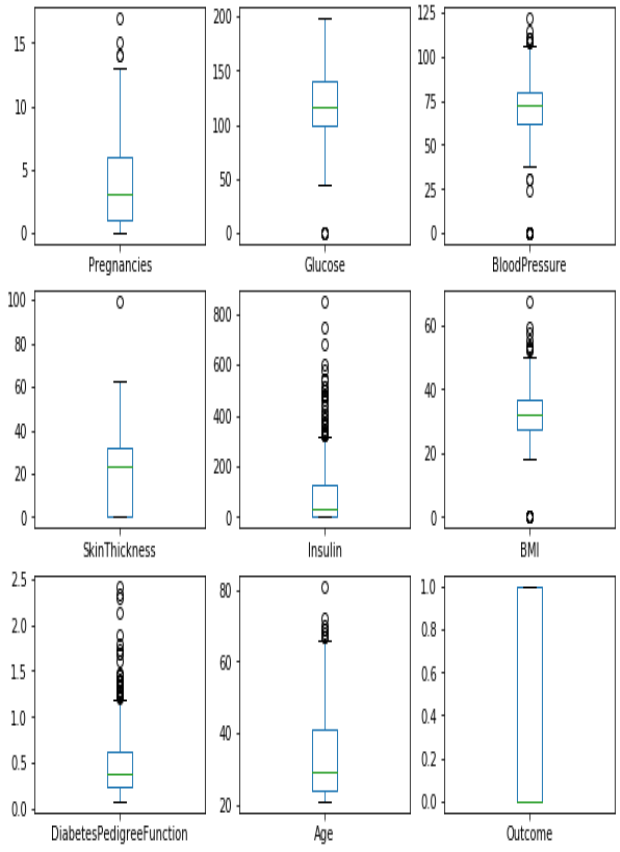

Figure 7 Count of zeros

and on Diabetes Pedigree Function are skewed to smaller values.
We estimate that the zeros that we saw on Fig. 8 may have a crucial role to the distribution of our dataset.

On Fig. 8 (Correlation plot) we can make an assumption whether a variable is correlated with another one. Each square of the correlation matrix gives us a number which represents the correlation between two variables. If the number is 1.0 we have a strong correlation between these two variables.

The diagonal is 1.0 which is being interpreted as each variable is fully correlated with itself. The -1.0 means we have a strong negative correlation.

We also have a heat map, where the darker purple colours stand for negative correlation or no correlation and as we move to brighter colours we have a strong correlation on white coloured squares.

From Fig. 8 we observe that the variable Outcome, whether a Pima Indian woman has diabetes or no, has 0.5 correlation with the variable Glucose. The next closest variable is BMI with 0.3.

We can see variable such as Age and Pregnacies score a correlation 0.5 which is something we expected because woman's age has a crucial role to their pregnancy.

IV. DISCUSSION

As can been seen from this brief review in order to find the characteristics and the various factors that affect Diabetes Mellitus we must take in mind a lot of different parameters and variables.

Our dataset was consisted of eight (8) variables which were used as measurements to diagnose whether there is a correlation between these variables and the ninth (9th) variable, the Outcome. The Outcome represents whether a female participant has diabetes or not. The participants were 768 female women all of Pima-Indian heritage.

Besides the expected correlation between the high levels of Glucose in the blood of the participants and the Outcome, those participants who in fact have Diabetes Mellitus we did not find another strong correlation between the variables in our dataset.

Another study of our dataset is recommended in which the zeros that take place in some variables, such as Glucose, BlodPressure, Insulin et, Fig 6, and they should not been there, must be carried out.

Perhaps in the new study those zeros could be replaced by the median of the variables.

Beside the variables that we examined in this brief survey we must acknowledge how the modern lifestyle affects the onset of Diabetes Mellitus as long as the dietary and the everyday habits of a person. [18]

Finally we must take in consideration the benefits of physical activity[19] and conduct new studies to examine those factors as well.

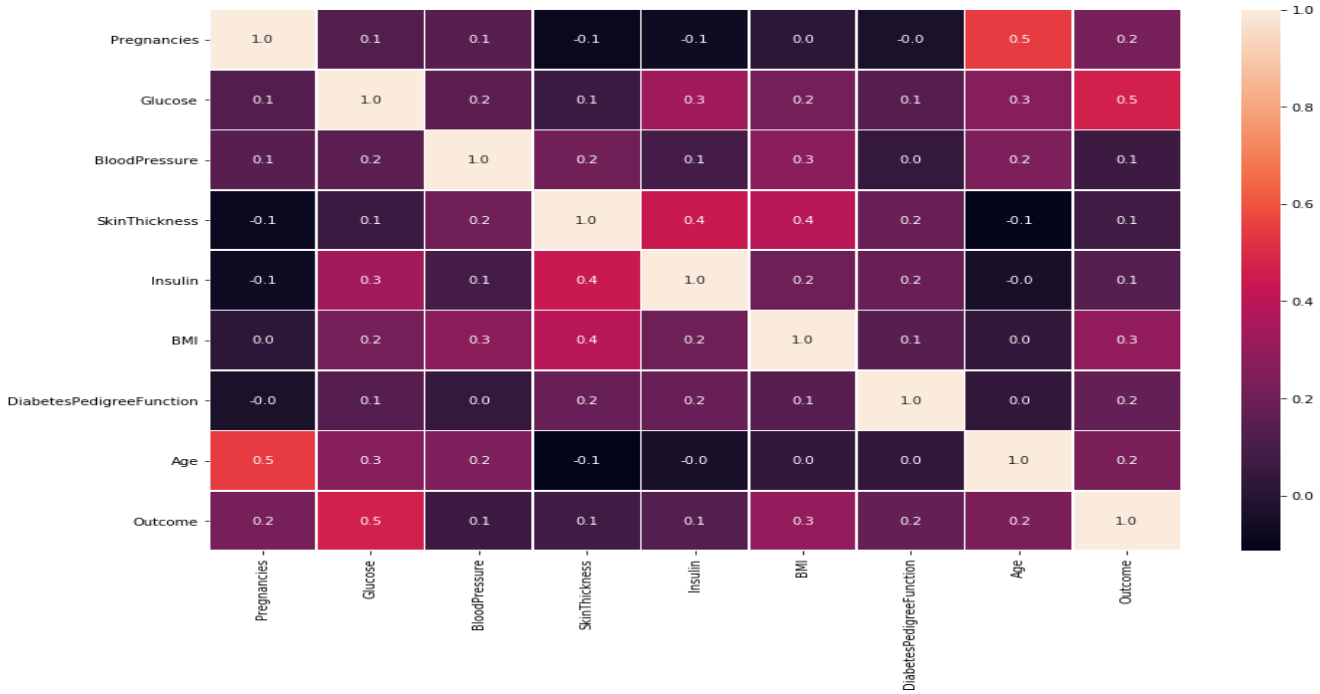

Figure 8 Correlation matrix

V. REFERENCES

[1] Pima people - Wikipedia n.d. https://en.wikipedia.org/wiki/Pima_people (accessed April 18, 2020).

[2] Diamond JM. Guns, germs, and steel : the fates of human societies. W.W. Norton & Co; 1997.

[3] Indian_Reservation n.d.:wikipedia.org. https://en.wikipedia.org/wiki/Indian_reservation (accessed May 2, 2020).

[4] Ethnicity and Health in America Series: Obesity in the Native American Community n.d. https://www.apa.org/pi/oema/resources/ethnicity-health/native-american/obesity (accessed May 2, 2020).

[5] Schell LM, Gallo M V. Overweight and obesity among North American Indian infants, children, and youth. Am J Hum Biol 2012;24:302–13. https://doi.org/10.1002/ajhb.22257.

[6] United States. Department of the Interior. Office of the Solicitor., Seaton FA (Fred A, Bennett EF. Federal Indian law. The Lawbook Exchange; 2008.

[7] Patterson-Silver Wolf DA, Welte JW, Barnes GM, Tidwell MCO, Spicer P. Sociocultural influences on gambling and alcohol use among native Americans in the United States. J Gambl Stud 2014;31:1387–404. https://doi.org/10.1007/s10899-014-9512-z.

[8] wikipedia. Alcohol_and_Native_Americans n.d. https://en.wikipedia.org/wiki/Alcohol_and_Native_Americans (accessed May 2, 2020).

[9] Quintero G. Making the Indian: Colonial Knowledge, Alcohol, and Native Americans. Am Indian Cult Res J 2001;25:57–71. https://doi.org/10.17953/aicr.25.4.d7703736566686m4.

[10] Beauvais F. American Indians and alcohol. Alcohol Res Heal 1998;22:253–9.

[11] wikipedia.org. Diabetes Mellitus n.d. https://en.wikipedia.org/wiki/Diabetes (accessed May 2, 2020).

[12] NHS. Diabetes n.d. https://www.nhs.uk/conditions/diabetes/ (accessed May 2, 2020).

[13] Kharroubi AT. Diabetes mellitus: The epidemic of the century. World J Diabetes 2015;6:850. https://doi.org/10.4239/wjd.v6.i6.850.

[14] Διαβήτης (ασθένεια) - Βικιπαίδεια n.d. https://el.wikipedia.org/wiki/Διαβήτης_(ασθένεια) (accessed May 2, 2020).

[15] Pima Indians Diabetes Database | Kaggle n.d. https://www.kaggle.com/uciml/pima-indians-diabetes-database (accessed May 2, 2020).

[16] NIDDK Frequently Asked Questions | NIDDK n.d. https://www.niddk.nih.gov/about-niddk/faqs#general-information (accessed May 2, 2020).

[17] Box plot - Wikipedia n.d. https://en.wikipedia.org/wiki/Box_plot (accessed June 27, 2020).

[18] (20) (PDF) The prevention and control the type-2 diabetes by changing lifestyle and dietary pattern n.d. https://www.researchgate.net/publication/261754246_The_prevention_and_control_the_type-2_diabetes_by_changing_lifestyle_and_dietary_pattern (accessed June 28, 2020).

[19] Colberg SR, Sigal RJ, Fernhall B, Regensteiner JG, Blissmer BJ, Rubin RR, et al. Exercise and type 2 diabetes: The American College of Sports Medicine and the American Diabetes Association: Joint position statement. Diabetes Care 2010;33:e147. https://doi.org/10.2337/dc10-9990.

.

APPENDIX

```
# -*- coding: utf-8 -*-
"""
Created on Sun Jun 28 12:31:57 2020

@author: Pc User
"""


import pandas as pd
import numpy as np
import seaborn as sn
import matplotlib.pyplot as plt
import matplotlib.mlab as mlab

#--- Εισαγωγή του dataset---
data = pd.read_csv("D:\MsC\Μεθοδολογία της Έρευνας\Project\pima-indians-diabetes-database\diabetes.csv")
data.info()
data.head()
data.shape
data.describe()
data.hist(figsize=(15,15))
data.hist(figsize=(10,8))
data.hist('Glucose')
data.hist('Age')

#---Εκτύπωση count_of_zero---
featurelist = []
count_of_zero_list = []
for col in data:
    cnt = 0
    for i in data[col]:
        if i==0:
            cnt = cnt + 1
    if col!='Outcome':
        #print (col, "-", cnt)
        featurelist.append(col)
        count_of_zero_list.append(cnt)

objects = tuple(featurelist)
y_pos = np.arange(len(featurelist))
performance = count_of_zero_list

fig_size = plt.rcParams["figure.figsize"]
fig_size[0] = 11
fig_size[1] = 3

plt.bar(y_pos, performance, align='center', alpha=0.9)
plt.xticks(y_pos, objects)

plt.ylabel('Count of 0 values')
plt.title('Count of 0 values per feature')
plt.grid(True)
plt.show()

#---Εκτύπωση box_plot---
data.plot(kind= 'box' , subplots=True, layout=(3,3), figsize=(10,8))

#--- Εκτύπωση correlation matrix---
import seaborn as sns
f,ax = plt.subplots(figsize=(15, 10))
```

```
61. sns.heatmap(data.corr(), annot=True, linewidths=.5, fmt= '.1f',ax=ax)
62. plt.show()
```