

My Final Project

Linh Ngau

March 2020

1 Abstract

Write a quick summary (300 words or less) that includes major aspects of the paper 1) purpose of study and research problems investigated 2) basic design of study 3) major findings or trends found as a result of analysis and 4) brief summary of interpretations and conclusions.

Contents

1	Abstract	1
2	Introduction	3
3	Methods	3
3.1	Regex Function	5
3.2	Analysis of BMI and Age in Patients	6
4	Results	7
5	Discussion	7
6	Figures	7

List of Figures

2 Introduction

The dataset was originally retrieved from the National Institute of Diabetes and Digestive and Kidney Diseases. The dataset aims to diagnostically predict whether a patient has diabetes based on particular diagnostic measurements retrieved from patients who are of Pima Indian Heritage. Early diagnosis of diabetes is important because, if a patient is found to have symptoms of diabetes, then there are preventative measures that can be taken in order to avoid being diagnosed with type 2 diabetes. Preventative measures that can be taken and advised to patients are lifestyle changes including increasing exercise frequency in conjunction with changes in diet by opting for more healthy food options.

Outline:

- Overview of project:
 - Purpose: to aid in identifying potential health factors that could contribute to a diagnosis of diabetes
 - Methods: Obtain data on 8 health factors of females over the age of 21 that are of Pima Indian Heritage
 - Major findings: there were factors that looked like they had more of an impact on the amount of people who had diabetes vs. not
 - : Biological impact: early diagnosis of diseases like diabetes can be life- altering because there are preventative measures that can be taken to prevent these diseases

3 Methods

```
1
2 \subsection{Importing the data}
3
4 The manipulated dataset was originally from the National Institute
  of Diabetes and Digestive and Kidney Diseases. All patients
  that data was recorded from are all females of at least 21
  years old and of Pima Indian Heritage.
5
6 import numpy
7 numpy.loadtxt(fname='pima-indians-diabetes.csv',delimiter=',')
8
9 #768 rows and 9 columns where the last column shows prediction of 1
  = diabetes and 0 = no diabetes
10 #column 1: Number of times pregnant
11 #column 2: Plasma glucose concentration a 2 hours in an oral
    glucose tolerance test
12 #column 3: diastolic blood pressure (mmHG)
13 #column 4: triceps skin fold thickness
14 #column 5: 2-hour serum insulin (mu U/ml)
15 #column 6: body mass index (weight in kg/(heightin m)^2)
16 #column 7: diabetes pedigree function
17 #column 8: age (years)
```

```

18
19 Data = numpy.loadtxt(fname='pima-indians-diabetes.csv',delimiter=',
    ')
20 import numpy as np
21
22 \subsection{Analysis of Predictor Variables of Diabetes}
23
24 maxval, minval = np.max(Data, axis=0), np.min(Data,axis=0)
25 stdval = np.std(Data,axis=0)
26
27 #shows the max, min and stdval of all columns given by axis = 0
28 print ('max:', maxval, ', min:',minval, ', stdev:', stdval)
29
30 #gives mean # of times pregnant from all patients in the data
    collection
31 NumberofTimesPregnant = np.mean(Data[:,0], axis=0)
32 print (NumberofTimesPregnant)
33
34 #visualize data through a dynamic function that allows user to
    retrieve data they choose to compare
35 import matplotlib.pyplot as plt
36 import os
37 import sys
38 def dynamic_plot():
39     xx= input ("Which column do you want to plot: ")
40     x = Data[:,int(xx)]
41     plt.hist(x, bins = 50)
42     yy= input ("What's the x axis label: ")
43     plt.xlabel(str(yy))
44     zz= input ("What's the y axis label: ")
45     plt.ylabel(str(zz))
46     aa= input ("What's the graph title: ")
47     plt.title (str(aa))
48     plt.show()
49
50 x = Data[:,2]
51 plt.hist(x, bins = 50)
52 plt.xlabel('Blood pressure')
53 plt.ylabel('number of individuals')
54 plt.title ('Blood Pressure')
55 plt.show()
56
57 import os
58 #Read file and datapath
59 datapath= '/home/eebc177student/developer/repos/eeb-c177-project/
    analyses'
60 workingdir= '/home/eebc177student/developer/repos/eeb-c177-project'
61 os.chdir(datapath)
62
63 #Patient data collected with each row indicating each patient
64 with open ("pima-indians-diabetes.csv", "r") as ff:
65     for line in ff:
66         print (line)
67
68
69 #Function allows me to access individual patient's data
70 dd= []

```

```

71 with open("pima-indians-diabetes.csv", "r") as ff:
72     for line in ff:
73         dd.append(line)
74 #to call for user's data: dd[row aka patient number here]
75
76 import csv
77
78 with open('pima-indians-diabetes.csv', newline='') as csvfile:
79     reader = csv.DictReader(csvfile)
80     for row in reader:
81         print(row)
82
83 with open('pima-indians-diabetes.csv', mode = 'r', encoding = 'utf
84 -8-sig') as csvfile:
85     reader = csv.DictReader(csvfile)
86     occurrences = 0
87     for row in reader:
88         occurrences = occurrences + 1
89     print('There are {} occurrences of values to a key.\n'.format(
90     occurrences))
91     print('There are {} rows, with the first row being keys '.
92     format(occurrences+1))
93     print('and all the subsequent rows being values to the keys.\n'
94     )
95
96 #Useful to look into number of occurrences of values within columns
97 #still working on debugging this

```

3.1 Regex Function

This implements the re.findall function that will analyze all of the integers within the list. In order to have this function work, I had to convert each column into its own text file. It will then output the maximum number within the column and the code can be manipulated to find the smallest value (aka minimum) as well to see the range of data.

```

1
2 colnames = ['number_of_times_pregnant', 'glucose_conc', '
3     diastolic_bp', 'skin_fold_thickness', '2hr_serum_insulin', '
4     bmi', 'diabetes_ped_fxn', 'age']
5 data = pandas.read_csv("pima-indians-diabets.csv", names=colnames)
6 number_of_times_pregnant.txt= data.number_of_times_pregnant.tolist
7     ()
8 glucose_conc.txt= data.glucose_conc.tolist()
9 diastolic_bp.txt= data.diastolic_bp.tolist()
10 skin_fold_thickness.txt= data.skin_fold_thickness.tolist()
11 2hr_serum_insulin.txt= data.2hr_serum_insulin.tolist()
12 bmi.txt= data.bmi.tolist()
13 diabetes_ped_fxn.txt= data.diabetes_ped_fxn.tolist()
14 age.txt=data.age.tolist()
15
16 import csv
17 import re
18
19 def extractMax("bmi.txt"):
20     numbers = re.findall ('\\d+', input)
21     numbers = map(int,numbers)

```

```

19 print max(numbers)
20
21 if max == "__main__":
22     extractMax("bmi.txt")
23 #found the last part online which allows the max number to be
    extracted from the text file

```

3.2 Analysis of BMI and Age in Patients

Using RStudio, the dataset was imported and two predictor variables were focused on, namely BMI and age of patients from data recorded.

```

1 library(dplyr)
2 library(ggplot2)
3
4 #Must import libraries in order to use them
5 library(ggplot2)
6 library(dplyr)
7 #used the import dataset button to import data through txt file
8 # pima.indians.diabetes <- read.csv("C:/Users/ngaul/Downloads/pima-
    indians-diabetes.csv", header=FALSE)
9
10 #assigning the dataset to a variable called df
11 df <- pima.indians.diabetes
12
13 #wanted to check how the dataset was imported and to see column
    names
14 head(df)
15
16 #define p1 as a ggplot composed of data from variable df and
    including all rows and column V6 which is BMI data
17 z <- df$V6
18 p1 <- ggplot(data=df, aes(x=1:768, y=V6)) +
19     geom_bar(stat="identity") +
20     geom_hline(yintercept = mean(z), color="blue")
21 #added a line that shows the average BMI given all patient data,
    which shows that most patients are below average
22 #add axis and graph titles
23 p1 + labs(title="Patient BMI Data", x = "Patient (Number)", y = "
    BMI (weight in kg/(height in m)^2)")
24
25
26 #did the same as the above in order to get a bar graph of age of
    all patients
27 a <- df$V8
28 p2 <- ggplot(data=df, aes(x=1:768, y=V8)) +
29     geom_bar(stat="identity") +
30     geom_hline(yintercept = mean(a), color="blue")
31 p2 + labs(title="Patient Age Data", x = "Patient (Number)", y = "
    Age (Years)")

```

4 Results

5 Discussion

It is vital to note that these predictions are based on very minimal health factors and to understand that there are a myriad of other factors that may affect one's chances of being diagnosed with diabetes.

6 Figures