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

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

```

\subsection{Importing the data}

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.

import numpy
numpy.loadtxt(fname='pima-indians-diabetes.csv',delimiter=',')

#768 rows and 9 columns where the last column shows prediction of 1
     = diabetes and 0 = no diabetes
#column 1: Number of times pregnant
#column 2: Plasma glucose concentration a 2 hours in an oral
    glucose tolerance test
#column 3: diastolic blood pressure (mmHG)
#column 4: triceps skin fold thickness
#column 5: 2-hour serum insulin (mu U/ml)
#column 6: body mass index (weight in kg/(heightin m)^2)
#column 7: diabetes pedigree function
#column 8: age (years)
```

```

Data = numpy.loadtxt(fname='pima-indians-diabetes.csv',delimiter=',')
import numpy as np

\subsection{Analysis of Predictor Variables of Diabetes}

maxval, minval = np.max(Data, axis=0), np.min(Data,axis=0)
stdval = np.std(Data,axis=0)

#shows the max, min and stdval of all columns given by axis = 0
print ('max:', maxval, ', min:',minval, ', stdev:', stdval)

#gives mean # of times pregnant from all patients in the data collection
NumberofTimesPregnant = np.mean(Data[:,0], axis=0)
print (NumberofTimesPregnant)

#visualize data through a dynamic function that allows user to retrieve data they choose to compare
import matplotlib.pyplot as plt
import os
import sys
def dynamic_plot():
    xx= input ("Which column do you want to plot: ")
    x = Data[:,int(xx)]
    plt.hist(x, bins = 50)
    yy= input ("What's the x axis label: ")
    plt.xlabel(str(yy))
    zz= input ("What's the y axis label: ")
    plt.ylabel(str(zz))
    aa= input ("What's the graph title: ")
    plt.title (str(aa))
    plt.show()

x = Data[:,2]
plt.hist(x, bins = 50)
plt.xlabel('Blood pressure')
plt.ylabel('number of individuals')
plt.title ('Blood Pressure')
plt.show()

import os
#Read file and datapath
datapath= '/home/eebc177student/developer/repos/eeb-c177-project/analyses'
workingdir= '/home/eebc177student/developer/repos/eeb-c177-project'
os.chdir(datapath)

#Patient data collected with each row indicating each patient
with open ("pima-indians-diabetes.csv", "r") as ff:
    for line in ff:
        print (line)


#Function allows me to access individual patient's data
dd= []
```

```
with open ("pima-indians-diabetes.csv", "r") as ff:
    for line in ff:
        dd.append(line)
#to call for user’s data: dd[row aka patient number here]

import csv

with open(’pima-indians-diabetes.csv’, newline=’’) as csvfile:
    reader = csv.DictReader(csvfile)
    for row in reader:
        print(row)

with open(’pima-indians-diabetes.csv’, mode = ’r’, encoding = ’utf-8-sig’) as csvfile:
    reader = csv.DictReader(csvfile)
    occurances = 0
    for row in reader:
        occurances = occurances + 1
    print(’There are {} occurances of values to a key.\n’.format(occurances))
    print(’There are  {} rows, with the first row being keys ’.format(occurances+1))
    print(’and all the subsequent rows being values to the keys.\n’)

#Useful to look into number of occurances of values within columns
#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.

```

colnames = [’number_of_times_pregnant’,  ’glucose_conc’,  ’diastolic_bp’,  ’skin_fold_thickness’, ’2hr_serum_insulin’, ’bmi’, ’diabetes_ped_fxn’, ’age’]
data = pandas.read_csv("pima-indians-diabets.csv", names=colnames)
number_of_times_pregnant.txt= data.number_of_times_pregnant.tolist()
glucose_conc.txt= data.glucose_conc.tolist()
diastolic_bp.txt= data.diastolic_bp.tolist()
skin_fold_thickness.txt= data.skin_fold_thickness.tolist()
2hr_serum_insulin.txt= data.2hr_serum_insulin.tolist()
bmi.txt= data.bmi.tolist()
diabetes_ped_fxn.txt= data.diabetes_ped_fxn.tolist()
age.txt=data.age.tolist()

import csv
import re

def extractMax("bmi.txt"):
  numbers = re.findall (’\d+’, input)
  numbers = map(int,numbers)
  print max(numbers)

if  max == "__main__":
  extractMax("bmi.txt")
#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.

```
library(dplyr)
library(ggplot2)

#Must import libraries in order to use them
library(ggplot2)
library(dplyr)
#used the import dataset button to import data through txt file
# pima.indians.diabetes <- read.csv("C:/Users/ngaul/Downloads/pima-indians-diabetes.csv", header=FALSE)

#assigning the dataset to a variable called df
df <- pima.indians.diabetes

#wanted to check how the dataset was imported and to see column names
head(df)

#define p1 as a ggplot composed of data from variable df and including all rows and column V6 which is BMI data
z <- df$V6
p1 <- ggplot(data=df, aes(x=1:768, y=V6)) +
  geom_bar(stat="identity")
#add axis and graph titles
p1 + labs(title="Patient BMI Data", x = "Patient (Number)", y = "BMI (weight in kg/(height in m)^2)")

#did the same as the above in order to get a bar graph of age of all patients
p2 <- ggplot(data=df, aes(x=1:768, y=V8)) +
  geom_bar(stat="identity")
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