SCHOOL OF COMPUTING
Faculty of Engineering

SEMESTER 2

SESSION 2019/2020

SECI2143

PROJECT 2 :
PREDICTING DIABETES

PREPARED BY :

NUR ALEEYA SYAKILA BINTI MUHAMAD SUBIAN (A19EC0127)

LECTURER'S NAME: DR. CHAN WENG HOWE
SECTION : 02
SUBMITTED ON : 27/6/2020

Contents

INTRODUCTION

DETAILS ABOUT DATASET:

The dataset was published by **Faysal Islam**. This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The purpose of the dataset is to diagnostically predict whether a patient has diabetes, based on certain diagnostic measurements included in the dataset. The link to the dataset is

https://www.kaggle.com/faysalislam/diabetes/metadata .

There is 2 type of diabetes. **Type 1** diabetes, when known as adolescent diabetes or insulin-subordinate diabetes, is a ceaseless condition in which the pancreas delivers practically zero insulin. Insulin is a hormone expected to permit sugar (glucose) to enter cells to create vitality. Various components, including hereditary qualities and some infections, may add to type 1 diabetes. In spite of the fact that type 1 diabetes normally shows up during youth or puberty, it can create in grown-ups (Mayo, 2017). **Type 2** diabetes is an interminable, possibly weakening and frequently deadly ailment requiring standard checking of a person's glucose level and treatment. In type 2 diabetes, the body either doesn't appropriately deliver or utilize insulin, a hormone created by the pancreas that assists move with sugaring into cells. Consequently, the body gets impervious to insulin. This opposition causes high glucose levels (OAC, 2012).

From this dataset that I chose, I want to **predict diabetes** by considering a few variables/factors (such as pregnancies, glucose etc) that possibly has relationship or rely on each other that gives the result to have diabetes. The following variables have been provided to help I predict whether a person is diabetic or not:

- **Pregnancies:** Number of times pregnant
- **Glucose:** Plasma glucose concentration over 2 hours in an oral glucose tolerance test
- **BloodPressure:** Diastolic blood pressure (mm Hg)
- **SkinThickness:** Triceps skin fold thickness (mm)
- **Insulin:** 2-Hour serum insulin (mu U/ml)
- **BMI:** Body mass index (weight in kg/(height in m)2)
- **DiabetesPedigreeFunction:** Diabetes pedigree function (a function which scores likelihood of diabetes based on family history)
- **Age:** Age (years)
- **Outcome:** Class variable (0 if non-diabetic, 1 if diabetic)
- **BMI Range:** BMI in ranging from normal, obese, overweight, underweight

1-SAMPLE TEST ON MEAN

A one sample hypothesis test on mean **means** compares the **mean** of a **sample** to a pre-specified value and **tests** for a deviation from that value.

In this 1-sample test on mean, I wanted to test a claim saying that *Normal* blood sugar levels are *less than 140* mg/dL two hours after eating (Nazario, 2018). I will be using the variable, **Glucose** which is the data for Plasma glucose concentration over 2 hours in an oral glucose tolerance test.

Hypothesis statement:

$H_0: \mu = 140$

$H_1: \mu < 140$

Execution of test:

Rstudio:

```
> #1 sample test mean
> n=757
> glu=diabetes$Glucose
> sd=sd(glu)
> sd
[1] 32.06143
> xbar=mean(glu)
> xbar
[1] 121.1361
> mu=140
> alpha=0.05
> z=(xbar-mu)/(sd/sqrt(n))
> z.alpha=qnorm(1-alpha)
> z
[1] -16.18816
> c(-z.alpha)
[1] -1.644854
> pval=pnorm(z)
> pval
[1] 3.056619e-59
```

From the Rstudio execution:

$\bar{x}$ = 121.1361

σ = 32.06143

μ = 140

n = 757

p-value = 3.056619e-59

$\boldsymbol{\alpha = 0.05}$

$$\sum x^2 = 11885296, \sum x = 91700, \bar{x} = \frac{91700}{757} = 121.1361, n = 757$$

$$\sigma = \sqrt{\frac{\sum x^2 - \frac{(\sum x)^2}{757}}{756}} = \sqrt{\frac{11885296 - \frac{8408890000}{757}}{756}} = 32.06143$$

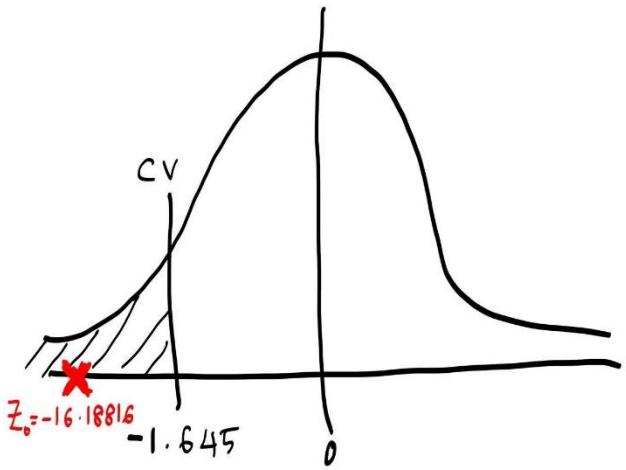
Finding the test statistic traditionally:

$$\boldsymbol{z_0} = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$= \frac{121.1361 - 140}{32.06143/\sqrt{757}}$$

$$= \mathbf{-16.18816}$$

CV : $\boldsymbol{z_{-0.05}}$ **= -1.645**



Interpretation of results

In this 1-sample hypothesis test, I used significance level of 0.05. I used left tailed test because from the claim it is said that normal blood sugar level is below than 140. From the result that I have obtained, the p-value is very small which is 3.056619e-59 compared to the alpha (0.05). In the other hand, the test statistic, $\boldsymbol{z_0}$= -16.18816 which is smaller than the critical value = -1.645. This causes the test statistic falls in the rejection region.

Conclusion and discussion

I **reject the null hypothesis**. This is because there is enough evidence to support the claim that Normal blood sugar levels are less than 140 mg/dL two hours after eating (Nazario, 2018). If an individual has blood sugar levels that are above than 140 mg/dL, the individual is predicted to have diabetes.

CORRELATION

The **correlation coefficient** is a statistical measure of the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0. A calculated number greater than 1.0 or less than -1.0 means that there was an error in the **correlation** measurement.

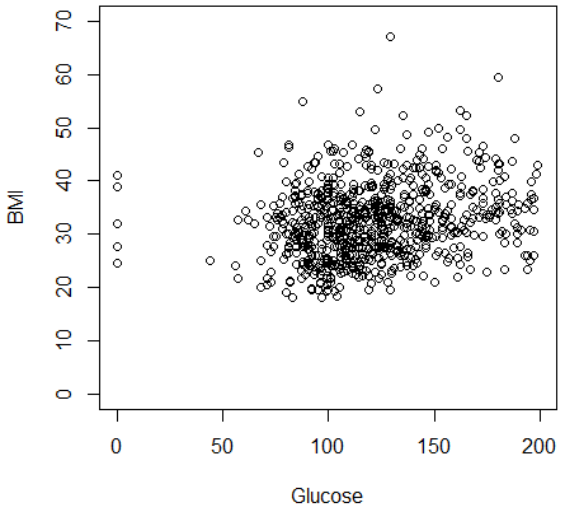
There is a claim where it is said that an increase in blood glucose level will result in increase in BMI causing increased lipid biosynthesis and hence body weight. (Neelam Agrawal, 2017). In this correlation test, I will measure the strength of the relationship between **Glucose** and **BMI.** The variable, **Glucose** refer to the Plasma glucose concentration over 2 hours in an oral glucose tolerance test. While the variable, **BMI** refer to the Body mass index.

Hypothesis statement:

$H_0: \rho = 0$ (no linear correlation between Glucose and BMI)

$H_1: \rho \neq 0$ (there is linear correlation between Glucose and BMI)

Execution of test:



From the Rstudio execution:

n = 757, df = 755, t = 6.1758, p-value = 1.075e-09

Σx=91700, Σy=24570.3, Σxy=3013158

Σx^2=11885296, Σy^2=833743.9, $\boldsymbol{\alpha = 0.05}$

Rstudio:

```
> #correlation test
> x<-c(diabetes$Glucose)
> y<-c(diabetes$BMI)
> totalx=sum(x)
> totaly=sum(y)
> totalxy=sum(x*y)
> x2=sum(x^2)
> y2=sum(y^2)
> totalx
[1] 91700
> totaly
[1] 24570.3
> totalxy
[1] 3013158
> x2
[1] 11885296
> y2
[1] 833743.9
> cor.test(x,y)

        Pearson's product-moment
correlation

data:  x and y
t = 6.1758, df = 755, p-value = 1.075e-09
alternative hypothesis: true correlation
is not equal to 0
95 percent confidence interval:
 0.1503835 0.2860768
sample estimates:
      cor
0.2192903

> plot(x,y,xlim = c(0,200),ylim =c(0,70),
xlab = "Glucose", ylab = "BMI" )
```

Finding the correlation coefficient traditionally:

$$r = \frac{\Sigma xy - (\Sigma x \Sigma y)/n}{\sqrt{[(\Sigma x^2) - (\Sigma x)^2/n][(\Sigma y^2) - (\Sigma y)^2/n]}}$$

$$= \frac{3013158 - (2253096510)/757}{\sqrt{[(11885296) - (91700)^2/757][(833743.9) - (24570.3)^2/757]}}$$

=0.21929

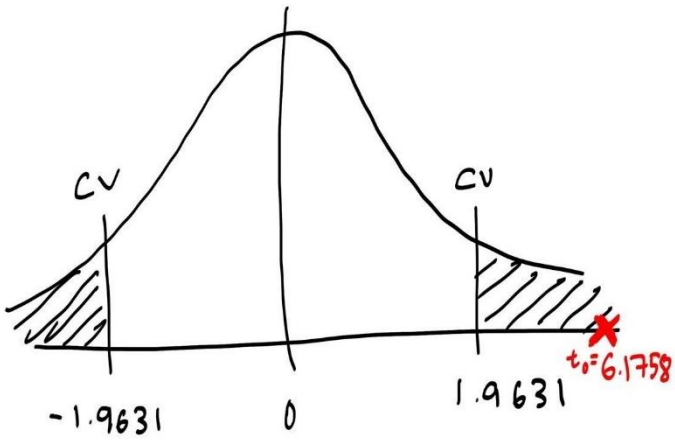
Finding the test statistic traditionally:

$$\boldsymbol{t_0} = \frac{\mathrm{r}}{\sqrt{\frac{1-r^2}{n-2}}}$$

$$= \frac{0.2192903}{\sqrt{\frac{1-0.2192903^2}{757-2}}}$$

$$= \mathbf{6.1758}$$

cv : $\boldsymbol{t_{0.05/2,755}}$ **= +/- 1.963111**



Interpretation of results

In this correlation test, I used significance level (alpha) of 0.05. This is a two-tailed test as the test statistic as if the $\rho = 0$ there will be no linear correlation between BMI and Glucose and if otherwise, it has linear correlation between BMI and Glucose.

From the result that I have obtained, the correlation coefficient is 0.21929. This resulting to **it has linear relationship** between *BMI* and *Glucose* and it is **weak linear relationship**. The result is significance at p-value<0.05. The p-value from the test is 1.075e-09 which is a smaller value than the significance level of 0.05. The test statistic, $t_0 = 6.1758$ is a larger value compare to the critical value, $\boldsymbol{t_{0.05/2,755}}$ =1.9631 which makes the test statistic falls in the rejection region.

Conclusion and discussion

I **reject the null hypothesis**. There is sufficient evidence to support that BMI and Glucose has a linear correlation, but they shared a weak relationship. As BMI builds, insulin obstruction additionally expands which brings about expanded blood glucose level in body. Since body weight is related with BMI, it might be normal that BMI should associate with blood glucose levels (Neelam Agrawal, 2017). As a conclusion, we can predict an individual to have diabetes by considering their BMI and their Glucose tolerance test result.

REGRESSION

Regression testing (rarely non-regression testing) is re-running functional and non-functional tests to ensure that previously developed and tested software still performs after a change. If not, that would be called a regression.

It is thought that when the body creates an excessive amount of insulin and leptin because of a higher-carb diet, it causes blood pressure to increment. Hyperinsulinemia raises circulatory strain, to some degree, by diminishing sodium and water discharge in the kidneys, and straightforwardly vasoconstricting veins. (High blood pressure: Why excess sugar in the diet may be the culprit, 2017). In this regression test, I will test the linear relationship between **Insulin** and **Blood Pressure.** The variable, **Insulin** refer to the result of 2-Hour serum insulin (mu U/ml). While the variable **Blood Pressure** which is Diastolic blood pressure (mm Hg)

Hypothesis statement:

H_0: β1 = 0 (no linear relationship between Insulin and Blood Pressure)

H_1: β1 ≠ 0 (there is linear relationship between Insulin and Blood Pressure)

Execution of test:

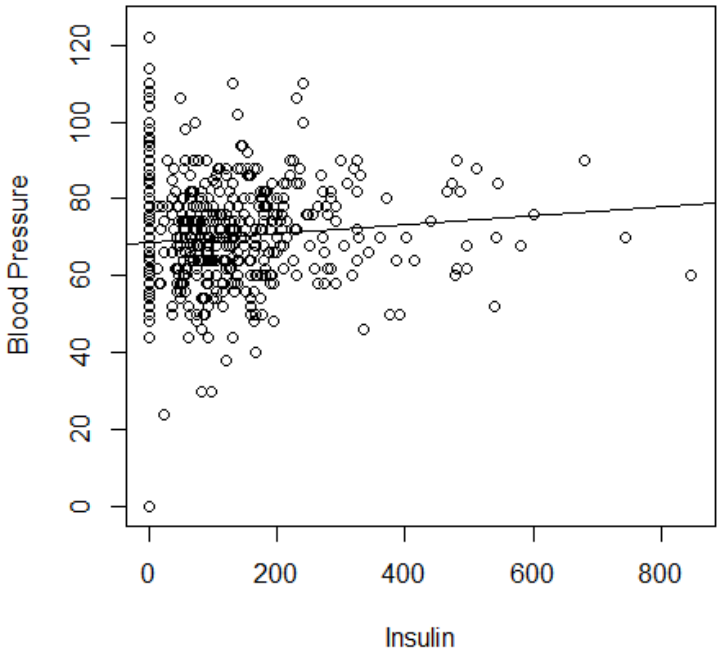
Rstudio:

```
> #regression test
> ins<-c(diabetes$Insulin)
> bp<-c(diabetes$BloodPressure)
> model<-lm(bp~ins)
> summary(model)

Call:
lm(formula = bp ~ ins)

Residuals:
    Min      1Q  Median      3Q     Max
-68.758  -6.758   1.488   9.742  53.242

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 68.758165   0.809182  84.972   <2e-16 ***
ins          0.011538   0.005735   2.012   0.0446 *
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.25 on 755 degrees of freedom
Multiple R-squared:  0.005332, Adjusted R-squared:  0.004014
F-statistic: 4.047 on 1 and 755 DF,  p-value: 0.0446

> plot(ins,bp, xlim = c(0,850), ylim = c(0,125), xlab =
"Insulin", ylab = "Blood Pressure")
> abline(model)
```

From the Rstudio execution:

$b_0 = 68.758165$, $b_1 = 0.011538$, **p-value** = 0.0446

$\alpha = 0.05$, $R^2 = 0.005332$



$\hat{y} = \mathbf{68.758165 + 0.011538}x$

SST= 252715.7, **SSE** = 251368.2, **SSR**= 1347.506

n = 757, x^2=15069335, $\sum x = 61197$

$$R^2 = \frac{SSR}{SST} = \frac{1347.506}{252715.7} = 0.005332$$

$$S_\varepsilon = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{251368.2}{757-2}} = 18.2466$$

$$S_{b_1} = \frac{S_\varepsilon}{\sqrt{\sum x^2 - \frac{(\sum x)^2}{n}}} = \frac{18.2466}{\sqrt{15069335 - \frac{61197^2}{757}}} = 0.0057352$$

Df= 757-2=755

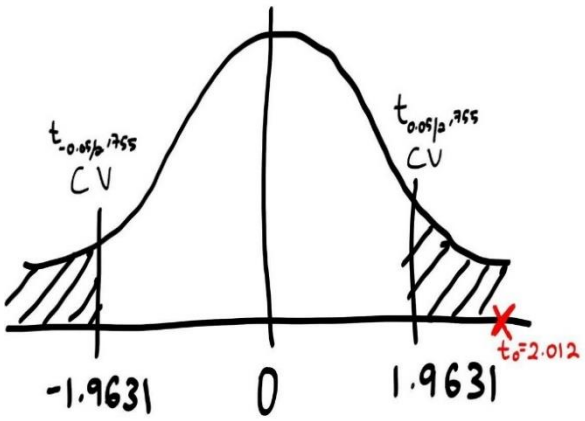
Finding the test statistic traditionally:

$$\boldsymbol{t_0} = \frac{b_1 - \beta_1}{S_{b_1}}$$

$$= \frac{0.011538 - 0}{0.0057352}$$

$$= \mathbf{2.01179}$$

CV : $\boldsymbol{t_{0.05/2,755}}$ **= +/- 1.963111**



Interpretation of results

From the results obtained, the plot graph above shows a line that intercept with the y-axis. The intercept, $\boldsymbol{b_0} = \mathbf{68.758165}$ is the estimated average value of Y (Blood Pressure) when the value of X(Insulin) is zero. $\boldsymbol{b_1} = \mathbf{0.011538}$ is the estimated change in the average value of Y (Blood Pressure) as a result of a one-unit change in X(Insulin). This will make the estimated regression model :

$\hat{y} = \mathbf{68.758165 + 0.011538}x$.

As example, if the insulin has 10 as a value, this will make the change of value of the blood pressure become 68.873545. The strongest **linear relationship** occurs when the slope is 1. This **means** that when one variable increases by one, the other variable also increases by the same amount. In this case, the $R^2 = 0.005332$ is located between $0 < R^2 < 1$. This shows that it has **weaker linear relationship** between Insulin and Blood Pressure. Some but not all of the variation in Blood Pressure is explained by variation in Insulin. About 0.53% of the variation in Blood Pressure is explained by variation in Insulin.

The result is significance at p-value<0.05. The p-value of the test is 0.0446 which is less than the alpha (0.05). The test statistic, $\boldsymbol{t_0 = 2.01179}$ is a larger value than the critical value, $\boldsymbol{t_{0.05/2,755}} = +1.963111$ which resulting the test statistic to falls in the rejection region.

Conclusion and discussion

I **reject the null hypothesis**. There is enough evidence to support the claim that Insulin and Blood Pressure has a linear relationship. Insulin can expand blood pressure through a few components: expanded renal sodium reabsorption, initiation of the thoughtful sensory system, modification of transmembrane particle transport, and hypertrophy of obstruction vessels (Salvetti A, 2012). Even though the results show a weak linear relationship between Insulin and Blood Pressure, it didn't change the fact that we can predict someone to have diabetes by looking at these two variables.

CHI SQUARE TEST-TEST OF INDEPENDENCE

A **chi-square test**, also written as χ^2 **test**, is a statistical hypothesis test that is valid to perform when the test statistic is chi-square distributed under the null hypothesis, specifically Pearson's chi-square test and variants thereof.

A survey showed that an increase in BMI is generally associated with a significant increase in prevalence of diabetes mellitus (Bays, 2007). In this chi-square test, I want to determine whether the **outcome** is independent of the **BMI Range** or not. The variable, **outcome** represents the status of diabetes result of an individual whether if 0=non-diabetic, 1=diabetic. While the variable, **BMI Range** represents the Body Mass Index (BMI) in its range from "underweight", " normal" , "overweight" and "obese".

Hypothesis statement:

H_0: Outcome is independent of the BMI Range

H_1: Outcome is dependent of the BMI Range

Execution of test:

Rstudio:

```
> #chisq test
> tbl=table(diabetes$`BMI
range`,diabetes$Outcome)
> tbl

                0   1
  normal       95   7
  obese       254 220
  overweight  138  39
  underweight   4   0
> chisq.test(tbl,correct = FALSE)

        Pearson's Chi-squared test

data:  tbl
X-squared = 77.724, df = 3, p-value < 2.2e-16

> alpha<-0.05
> x2.alpha<-
qchisq(alpha,df=3,lower.tail=FALSE)
> x2.alpha
[1] 7.814728
```

From the Rstudio execution:

$x^2 = 77.724$, $df = 3$, **p-value** = < 2.2e-16

$\alpha = 0.05$

Frequency table

OUTCOME			
BMI RANGE	0	1	TOTAL
NORMAL	95	7	102
OBESE	254	220	474
OVERWEIGHT	138	39	177
UNDERWEIGHT	4	0	4
TOTAL	491	266	757

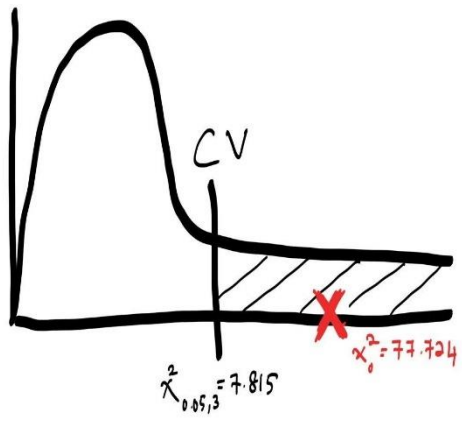
cell	Observed (o)	Expected (E)	$\frac{(o_{ij}-e_{ij})^2}{e_{ij}}$
1,1	95	$\frac{491x102}{757}=66.159$	12.57330024
1,2	7	$\frac{266x102}{757}=35.8415$	23.20861059
2,1	254	$\frac{491x474}{757}=307.4425$	9.289881365
2,2	220	$\frac{266x474}{757}=166.5575$	17.14786372
3,1	138	$\frac{491x177}{757}=114.8045$	4.686503219
3,2	39	$\frac{266x177}{757}=62.1955$	8.650650679
4,1	4	$\frac{491x4}{757}=2.5945$	0.761457893
4,2	0	$\frac{266x4}{757}=1.4055$	1.405548217
			$\sum x^2 = 77.72381592$

$\sum x^2 = 77.72381592$

CV : $\alpha = 0.05$

DF= (4-1)(2-1)=3

$x^2_{0.05,3} = 7.815$



Interpration of results

In this chi square test, I used significance level of 0.05 to test. From the data, we can see the percentage of the BMI range of the sample of 757 people with normal (13.47%) , obese (62.62%), overweight (23.38%) and underweight (0.53%). While the percentage of the outcome of the diabetes result which is 0 and 1 are 64.86% and 35.14% respectively. The result is significance at p-value<0.05. From the result that I have obtained, the p-value is < 2.2e-16, which means it is very small compared to the significance level, 0.05. In the other hand, the test statistic,$x^2 = 77.72381592$ is a larger value than the critical value, $x^2_{\mathbf{0.05,3}} = \mathbf{7.815}$**.** This resulting the test statistic falls in the rejection region.

Conclusion and discussion

I **reject the null hypothesis**. There is enough sufficient evidence to support that the outcome of the diabetes test is dependent to the BMI. Being overweight (BMI of 25-29.9) or influenced by obesity (BMI of 30-39.9) or severe obesity (BMI of 40 or more noteworthy), greatly increases your risk of developing type 2 diabetes. The more overabundance weight you have, the more resistant your muscle and tissue cells become to your own insulin hormone. In excess of 90 percent of individuals with type 2 diabetes are overweight or influenced by a level of weight (OAC, 2012). In conclusion for this test, we can predict someone to have diabetes by referring their BMI Range which most of the case, obese people will turn out to be diabetic but things shouldn't be biased to obese people only as there a lots of other factors to get diabetes without having to be an obese person.

DISCUSSION

From all of the test that I have done, I was expecting some of the results such as the linear relationship between Insulin and Blood Pressure to have strong linear relationship by having the Coefficient of determination, R^2 to be exact value of 1 so that it will be a perfect linear relationship but indeed coefficient usually not "perfect".

CONCLUSION

In a nutshell, what I can say is that we can predict someone to have diabetes by looking at these variables such as the blood sugar level, BMI, the level of Insulin etc but we can't solely use them to diagnose people with diabetes. Definitely diabetes should be diagnosed by the professionals and doctors. What we can do is that to be concern if we have such condition in our body such as high blood pressure, a high level of blood sugar level or even a big value of BMI as it can be diabetes factors and we should bring more awareness to these diabetes factors to the public.

References

Bays, H. E. (2007, April 10). *The relationship of body mass index to diabetes mellitus, hypertension and dyslipidaemia: comparison of data from two national surveys*. Retrieved from onlinelibrary.wiley: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1742-1241.2007.01336.x

Dhandhania, K. (2018, May 25). *End-to-End Data Science Example: Predicting Diabetes with Logistic Regression*. Retrieved from towardsdatascience: https://towardsdatascience.com/end-to-end-data-science-example-predicting-diabetes-with-logistic-regression-db9bc88b4d16#:~:text=DiabetesPedigreeFunction%3A%20Diabetes%20pedigree%20function%20(a,%2Ddiabetic%2C%201%20if%20diabetic)

High blood pressure: Why excess sugar in the diet may be the culprit. (2017, May 05). Retrieved from diabetes.co.uk: https://www.diabetes.co.uk/in-depth/high-blood-pressure-excess-sugar-diet-may-culprit/

Mayo. (2017, August 07). *Type 1 diabetes*. Retrieved from mayoclinic: https://www.mayoclinic.org/diseases-conditions/type-1-diabetes/symptoms-causes/syc-20353011

Nazario, B. (2018, October 12). *webmd*. Retrieved from What are normal blood sugar levels?: https://www.webmd.com/diabetes/qa/what-are-normal-blood-sugar-levels

Neelam Agrawal, M. K. (2017, August 18). *Correlation between Body Mass Index and Blood Glucose Levels in*. Retrieved from ijcmr: https://www.ijcmr.com/uploads/7/7/4/6/77464738/ijcmr_1592.pdf

OAC. (2012, March 19). *Understanding Excess Weight and its Role in Type 2 Diabetes brochure*. Retrieved from obesityaction: https://www.obesityaction.org/get-educated/public-resources/brochures-guides/understanding-excess-weight-and-its-role-in-type-2-diabetes-brochure/#:~:text=Being%20overweight%20(BMI%20of%2025,to%20your%20own%20insulin%20hormone.

Salvetti A, B. G. (2012, October 22). *The inter-relationship between insulin resistance and hypertension*. Retrieved from pubmed: https://link.springer.com/article/10.2165/00003495-199300462-00024