Project 1

Tien Le, Jackson Kelly

2024-10-17

1 Abstract

Diabetes is a chronic disease affecting millions worldwide, often leading to severe complications like heart disease, stroke, and kidney failure. Our study focuses on identifying significant factors related to diabetes and developing predictive models to aid in diagnosis. Using a dataset from the National Institute of Diabetes and Digestive and Kidney Diseases, we examine the relationships between various attributes—such as glucose levels, BMI, and age—and the likelihood of developing diabetes. Through correlation analysis and statistical modeling, we found that glucose levels and BMI have the strongest correlations with diabetes. A predictive model was constructed using these factors, though its accuracy was limited by the dataset size. Our findings suggest that glucose and BMI are strong indicators of diabetes risk, and further research with larger datasets and more features, such as family history, could enhance diagnostic precision.

2 Introduction

Our study aims to better understand diabetes and predict it based on various individual characteristics. Diabetes is one of the most chronic, costly, and consequential diseases. By gaining a deeper understanding of diabetes and making accurate predictions, we can help in treatment and care for those affected. By examining different contributing factors to diabetes, such as BMI, age, and plasma glucose concentration, we can identify patterns in those affected and and make predictions about others based on these observations. We hope that the findings of our study can improve diagnosis by more accurately identifying diabetes in at risk individuals.

Diabetes is a serious issue that affects an estimated 8.5% of adults worldwide. It can cause heart attack, strokes, kidney failure, and death. For type 2 diabetes, which is the majority of cases, it is often preventable, and the worst effects can be avoided by an early diagnosis and treatment (“Diabetes — Who.int”). More information about diabetes can be found here (https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes) (“What Is Diabetes? - NIDDK — Niddk.nih.gov”). Therefore, being able to effectively diagnose diabetes is critical, and we hope our findings can help contribute to doing so.

3 Data

Our data set comes from the National Institute of Diabetes and Digestive and Kidney Diseases. This data set examines 768 individuals, with measurements for their number of pregnancies, glucose level, blood pressure, skin thickness, insulin levels, BMI, age, and diabetes pedigree (a genetic estimate of an individual’s likelihood to develop diabetes). 268 of these individuals do have diabetes, and are labeled accordingly. This label is found in the outcome column and is binary, with a 1 signifying an individual with diabetes, and a 0 an individual without.

This data set has a lot of key information with regards to diabetes. For example, according the WHO, 95% of diabetes cases are type 2, which is often attributed to being overweight, not getting enough exercise, and genetics. Thus a measurement like BMI is very relevant to examining diabetes. Additionally, features like insulin and glucose level relate directly to the disease, and thus their measurement is important (“Diabetes — Who.int”). The limitation of the data set is that we are only looking at 8 features, and there are more factors that can potentially cause diabetes. The data set is also relatively small, with only 268 diabetic individuals potentially not properly illustrating true patterns. (“Diabetes Healthcare: Comprehensive Dataset-AI — Kaggle.com”)

4 Analysis

```
data <- read.csv("health care diabetes.csv")
data <- na.omit(data)

library(corrplot)
```

```
## corrplot 0.95 loaded
```

```
matrix <- cor(data)

corrplot(matrix, method = "color", col = colorRampPalette(c("white", "blue"))(500), addCoef.col
= "black", tl.srt = 45, tl.col = "black")
```

```
## Warning in ind1:ind2: numerical expression has 2 elements: only the first used
```

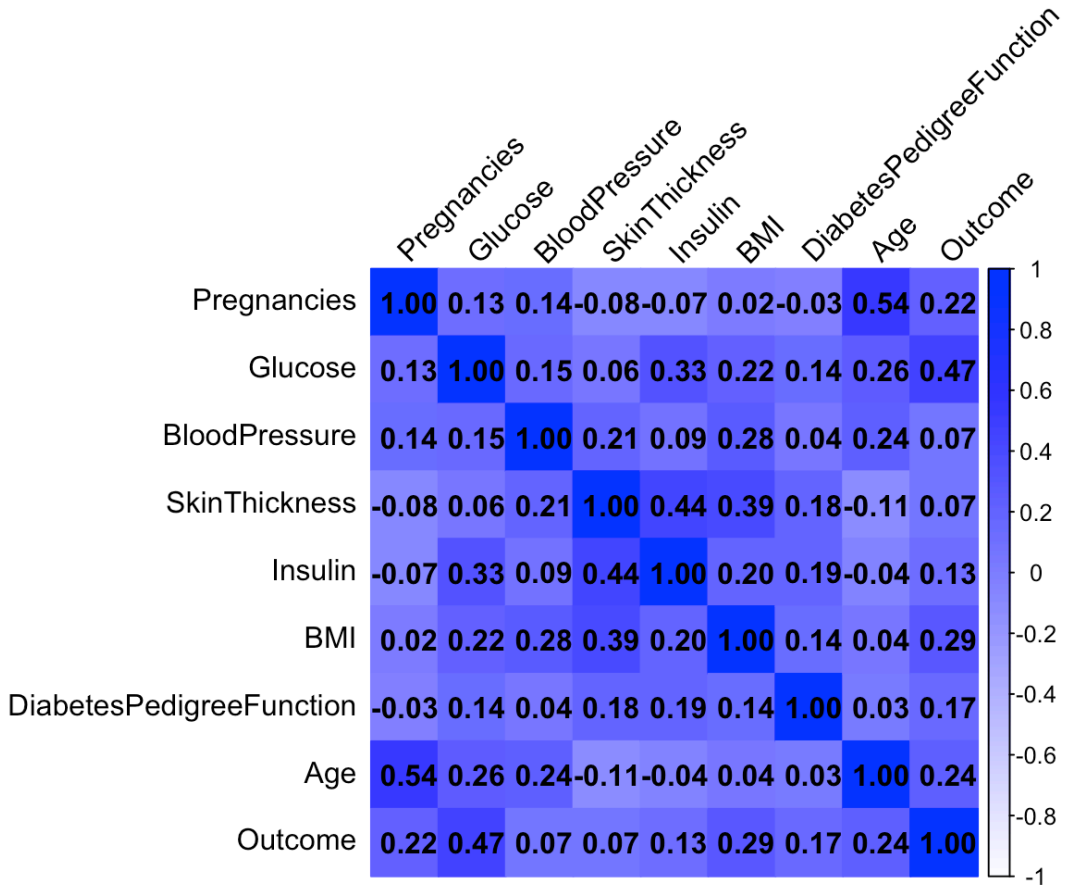

Figure 4.1: correlation plot

In this figure 4.1, we see the relationship of different attributes with each other, and ultimately with diabetes as a whole. This is done by calculating the correlation coefficient for each combination of attributes. Based on the bottom row, we can see that Glucose, BMI, and Age have the largest correlation with diabetes (at 0.47, 0.29, and 0.24 respectively). We also observe logical correlations in some of the attributes, such as number of pregnancies and age.

```
par(mfrow = c(1, 2))
hist(data[data$Outcome == 1, ]$Glucose, xlab = 'Glucose level', ylab = '# Observations', main='G
lucose Levels w/ Diabetes', xlim = c(min(data$Glucose), max(data$Glucose)), col = "red")
hist(data[data$Outcome == 0, ]$Glucose, xlab = 'Glucose level', ylab = '# Observations', main='G
lucose Levels w/out Diabetes', xlim = c(min(data$Glucose), max(data$Glucose)), col = "blue")
```

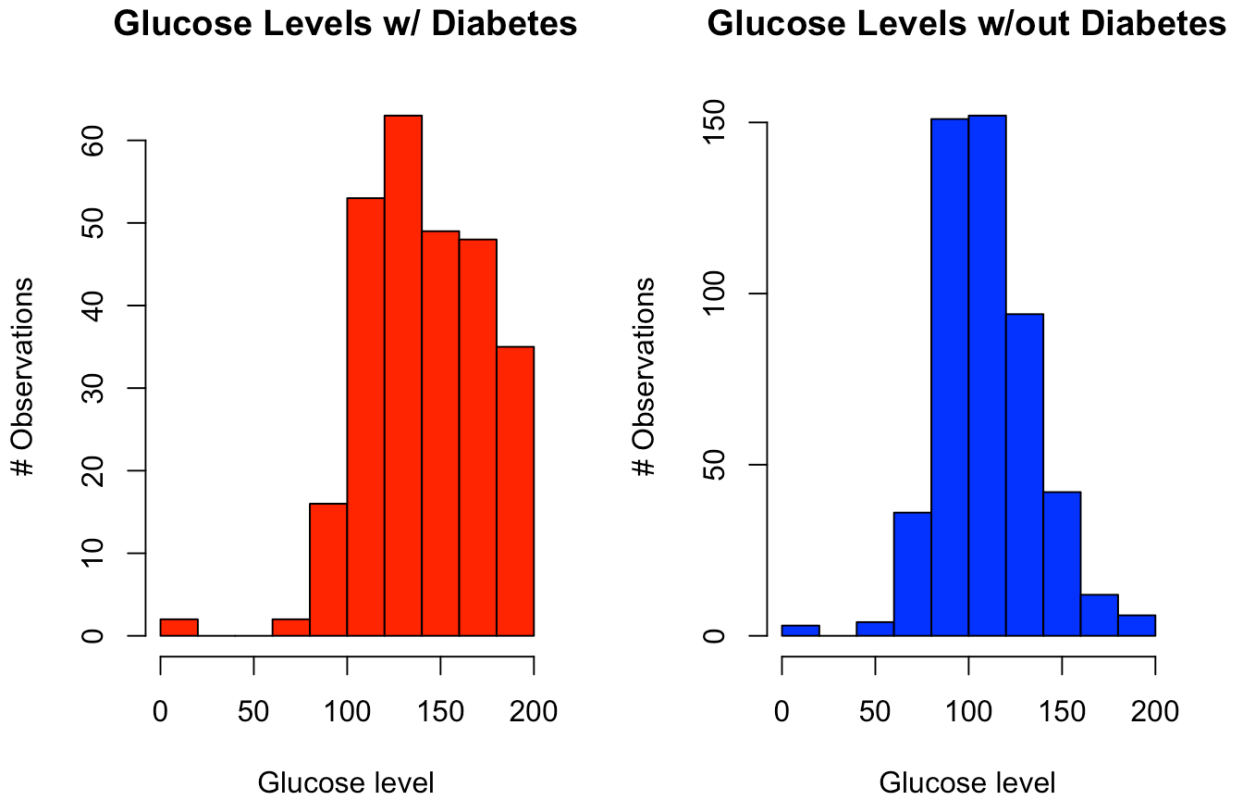

Figure 4.2: Comparing glucose levels with and without diabetes

```
par(mfrow = c(1, 1))
```

Since glucose was previously identified as the attribute with the largest correlation for our outcome, we will examine the relationship closer. These two histograms 4.2 demonstrate the difference in glucose levels between those with diabetes, and those without. For those with the disease, the glucose levels are left-skewed, being centered well above 100. In contrast, the histogram for those without diabetes shows no skew, and is centered much closer to 100. This illustrates a pattern of higher glucose levels in those with diabetes. Additionally, almost everyone with very high glucose values (>150) did have diabetes.

5 Significance of Glucose P-value

The p-value for Glucose is 2.6911924^{-28}. This small p-value, <0.05, suggests that Glucose has a statistically significant relationship with diabetes. We have high confidence that glucose level can predict the diagnostic of diabetes. Therefore, to support the above's histogram, high glucose correlates to the outcome of diabetes.

```
data$BMI_group <- cut(data$BMI, breaks = 7)
group_avg <- aggregate(Outcome ~ BMI_group, data=data, FUN=mean)
group_avg$group_num <- as.numeric(group_avg$BMI_group)

plot(group_avg$group_num, group_avg$Outcome, xlab = "BMI Grouping (low to high)", ylab = "Diabet
es", main = "BMI and Diabetes", col = "blue", pch = 16)
model <- lm(Outcome ~ group_num, data = group_avg)
abline(model, col = "blue", lwd = 1)
```

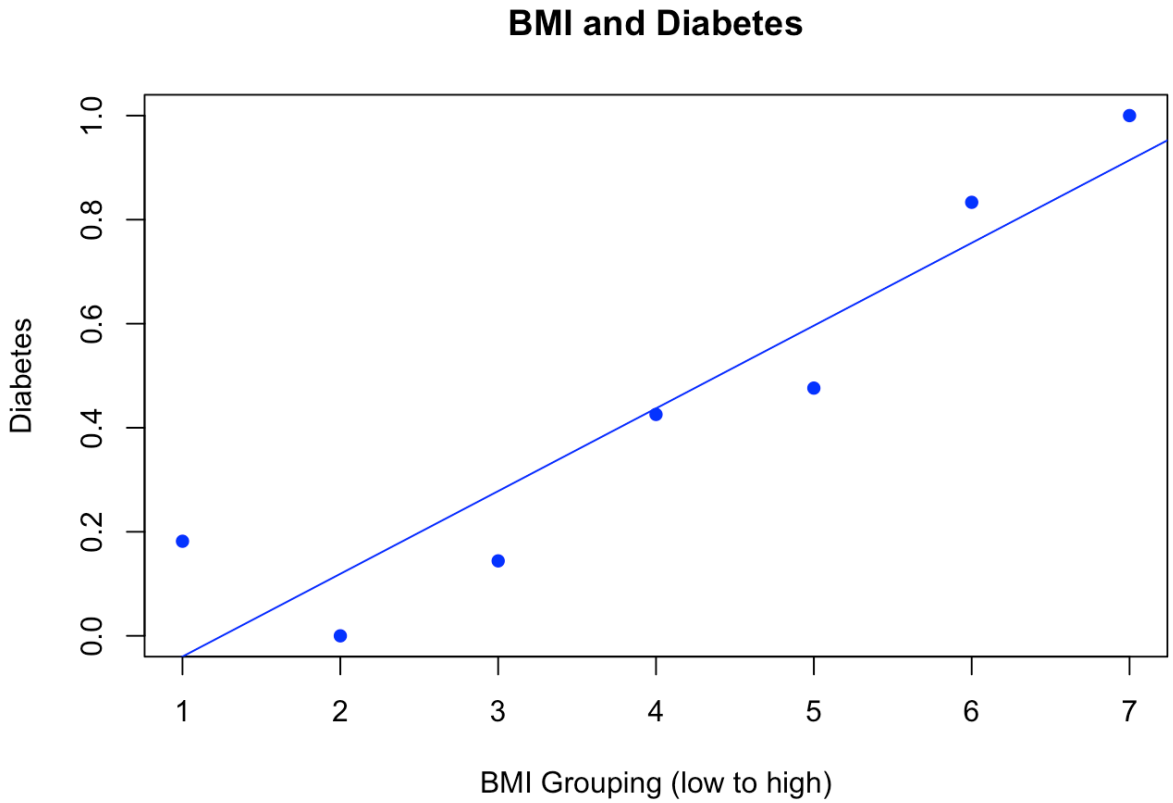

Figure 5.1: Demonstrates a trend of individuals with higher BMIs having higher rates of diabetes.
This plot 5.1 demonstrates the correlation between BMI and diabetes. We started by grouping the individuals in the dataset into 7 even groups based on their BMI. From there, we took the average of each group, with a 1 representing an individual with diabetes and a 0 representing one without. Based on these averages, we observed a trend of groups with larger BMIs having higher averages, and thus being more inclined to diabetes. Therefore, we demonstrate the strong correlation between BMI and diabetes in our dataset.

6 Significance of BMI P-value

The p-value for BMI is 3.8534837^{-10}. While being below 0.05, the BMI p-value demonstrates that BMI levels are siginificant enough or us to predict the outcome of diabetes. Therefore, to support the above's plot, we demonstrate the strong correlation between BMI and diabetes in our dataset.

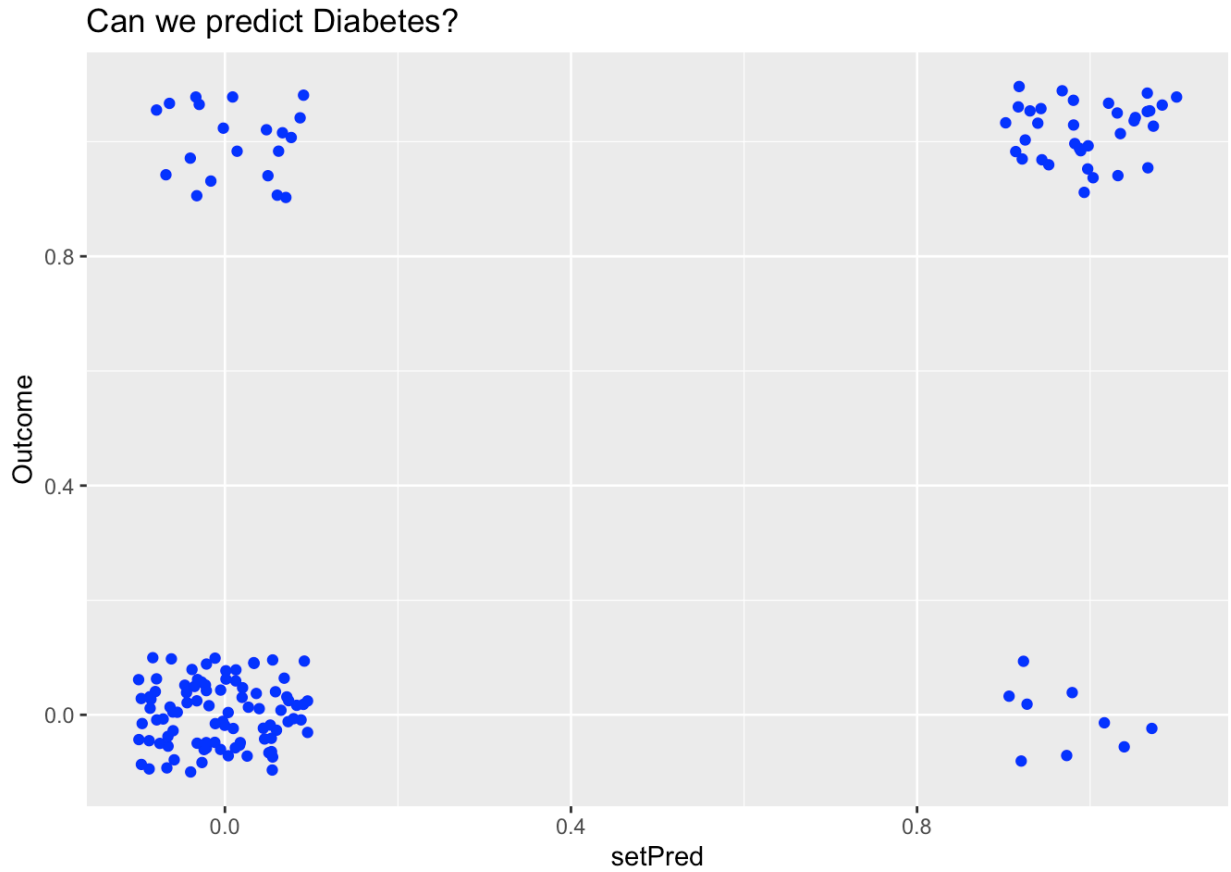

Figure 6.1: Predictions scatterplot

7 Diabetes Algorithm

The graph 6.1 above illustrates a predictive algorithm for diabetes. We started by splitting our data into training and test sets. Additionally, we chose to only look at the three most relevant features, being Glucose, BMI, and Age. The training set was used to give the model a reference to how the parameters affected the outcome, i.e. whether or not an individual has diabetes. We then apply the model to the rest of the data in the test set, and then compare to the known outcome given in the data. The scatterplot shows 4 potential outcomes of our prediction, a true negative at (0, 0), a false negative at (0, 1), a false positive at (1, 0), and a true positive at (1, 1). We can calculate the accuracy to be 0.8116883. While our predictions appear to be correct more often than not, the amount of errors is large enough to conclude that we cannot consistently make accurate predictions for diabetes just based on the features examined. (R Core Team 2023; RStudio Team 2020; “Learn Linear Regression in R Cheatsheet — Codecademy.com”)

8 Conclusion

Our study aimed to be able to predict diabetes effectively, as well as examine the underlying causes and symptoms. By examining trends in our data set, we were able to identify key features of diabetes, such as BMI, glucose levels, and age. Additionally, we were able to set up a model to make predictions about an individuals diagnosis based on their measurements in the data. These prediction models demonstrates significance where we can look at BMI, glucose and age to predict if the patient has diabetes. If future steps were taken, a bigger sample of individuals would be beneficial to get a better idea of the population and improve the accuracy of our predictions. Another idea is taking the frequency of how often the patients to examine, to account for skewness where BMI and glucose are inconsistent due to individual differences. Family history, Pedigree function, should also be taken in account, since genetics were not accounted in this study.

9 Contribution Statement

Tien Le: Wrote abstract and helped write conclusion, worked on analysis and results section, including performing linear model to describe graphs and its significance.
Jackson Kelly: Wrote intro and data set paragraphs. Worked on various plots and their analysis. Helped write conclusion statement.

Works Cited

"Diabetes — Who.int." https://www.who.int/news-room/fact-sheets/detail/diabetes#:~:text=People%20with%20diabetes%20have%20a,damage%20and%20poor%20blood%20flow. (https://www.who.int/news-room/fact-sheets/detail/diabetes#:~:text=People%20with%20diabetes%20have%20a,damage%20and%20poor%20blood%20flow.)

"Diabetes Healthcare: Comprehensive Dataset-AI — Kaggle.com." https://www.kaggle.com/datasets/deependraverma13/diabetes-healthcare-comprehensive-dataset (https://www.kaggle.com/datasets/deependraverma13/diabetes-healthcare-comprehensive-dataset).

"Learn Linear Regression in R Cheatsheet — Codecademy.com." https://www.codecademy.com/learn/learn-linear-regression-in-r/modules/linear-regression-in-r/cheatsheet (https://www.codecademy.com/learn/learn-linear-regression-in-r/modules/linear-regression-in-r/cheatsheet).

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/ (https://www.R-project.org/).

RStudio Team. 2020. *RStudio: Integrated Development Environment for r*. Boston, MA: RStudio, PBC. http://www.rstudio.com/ (http://www.rstudio.com/).

"What Is Diabetes? - NIDDK — Niddk.nih.gov." https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes (https://www.niddk.nih.gov/health-information/diabetes/overview/what-is-diabetes).