STATS/CSE 780
Project Proposal

Pao Zhu Vivian Hsu (Student Number: 400547994)

2023-10-24

Introduction

Diabetes is a chronic disease that occurs when the body cannot effectively produce or use insulin to regulate sugar levels in the blood. According to the World Health Organization, 422 million people have diabetes worldwide and 1.5 million deaths that occur every year are linked to the disease (World Health Organization, n.d.). In this paper, we propose a study that leverages machine learning techniques to predict and prevent diabetes.

The data in this proposal was originally collected by Islam et al. from patients in Sylhet Diabetes Hospital in Sylhet, Bangladesh (2020) and was later openly published on Kaggle (Larxel, 2023).

Islam et al.'s study also used machine learning techniques to predict diabetes. In particular, they used naive Bayes, logistic regression, and random forest techniques and found that their random forest model had the best accuracy (Islam et al., 2020). This study will also use machine learning techniques, but will focus on random forest and neural network methods instead.

Methods

The first method will be a decision tree with ensemble methods, specifically random forest. The tuning parameter is

The second method will be a neural network with regularization, specifically lasso. The tuning parameter is We will measure the computational cost of the neural network using .

-modifications, -tuning parameters, -feature engineering, -computational cost, interpretability of the results, reproducibility

The neural network is .

After the two models are built, they will each be assessed using mis-classification rate, accuracy, specificity and sensitivity. A comparison of these four measurements will reveal which model is a stronger fit for to predict diabetes.

Preliminary Analysis

This section provides preliminary insight into the data set that may be useful for the proposed study.

The data set consists of 520 observations and 17 attributes. Before any analysis was done, a transformation was applied to the data to ensure that all categorical variables were expressed with binary indicators so they can be easily used in the analysis. The response variable is a binary measure called class that indicates whether the patient has a positive or negative risk for diabetes. The remaining attributes describe the patient and whether they experience a selection of symptoms related to the disease, such as weakness, itching, and obesity. A full list of the attributes and their meanings are outlined in Figure 1 in the Supplementary Materials section.

There are no missing values in this data set since missing data was already addressed by Islam et al. after data collection (2020). To verify this, we performed a check for nulls and blanks as summarized in Figure 2 in the Supplementary Materials.

Next, each of the individual attributes were explored. Figure 3 shows a box plot of patient ages. The are a few outliers

Timelines

This project will include two main components, a presentation and a written report. The presentation will occur on November 30th, 2023. Slides for the presentation will be completed by November 21, 2023. The written report will be finalized by December 11, 2023.

Supplementary Materials

Figures

Attribute	Values
Age	In years
Gender	1 = Male, 0 = Female
Polyuria	1 = Yes, 0 = No
Polydipsia	1 = Yes, 0 = No
Sudden weight loss	1 = Yes, 0 = No
Weakness	1 = Yes, 0 = No
Polyphagia	1 = Yes, 0 = No
Genital thrush	1 = Yes, 0 = No
Visual blurring	1 = Yes, 0 = No
Itching	1 = Yes, 0 = No
Irritability	1 = Yes, 0 = No
Delayed healing	1 = Yes, 0 = No
Partial paresis	1 = Yes, 0 = No
Muscle stiffness	1 = Yes, 0 = No
Alopecia	1 = Yes, 0 = No
Obesity	1 = Yes, 0 = No
Class	1 = Positive risk, 0 = Negative risk

Figure 1: Description of attributes

Nulls	Blanks
0	0

Figure 2: No missing data

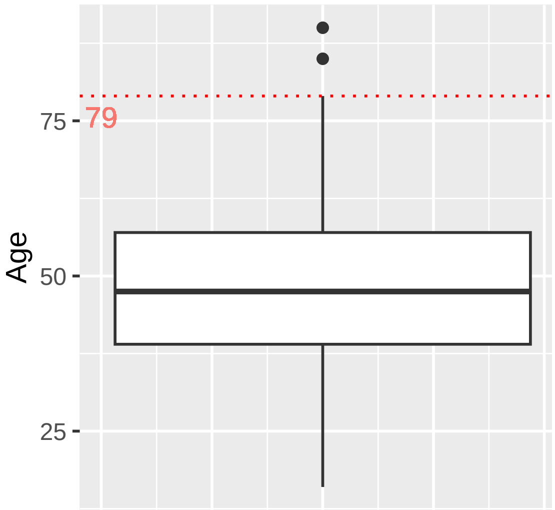

Figure 3: Boxplot of age

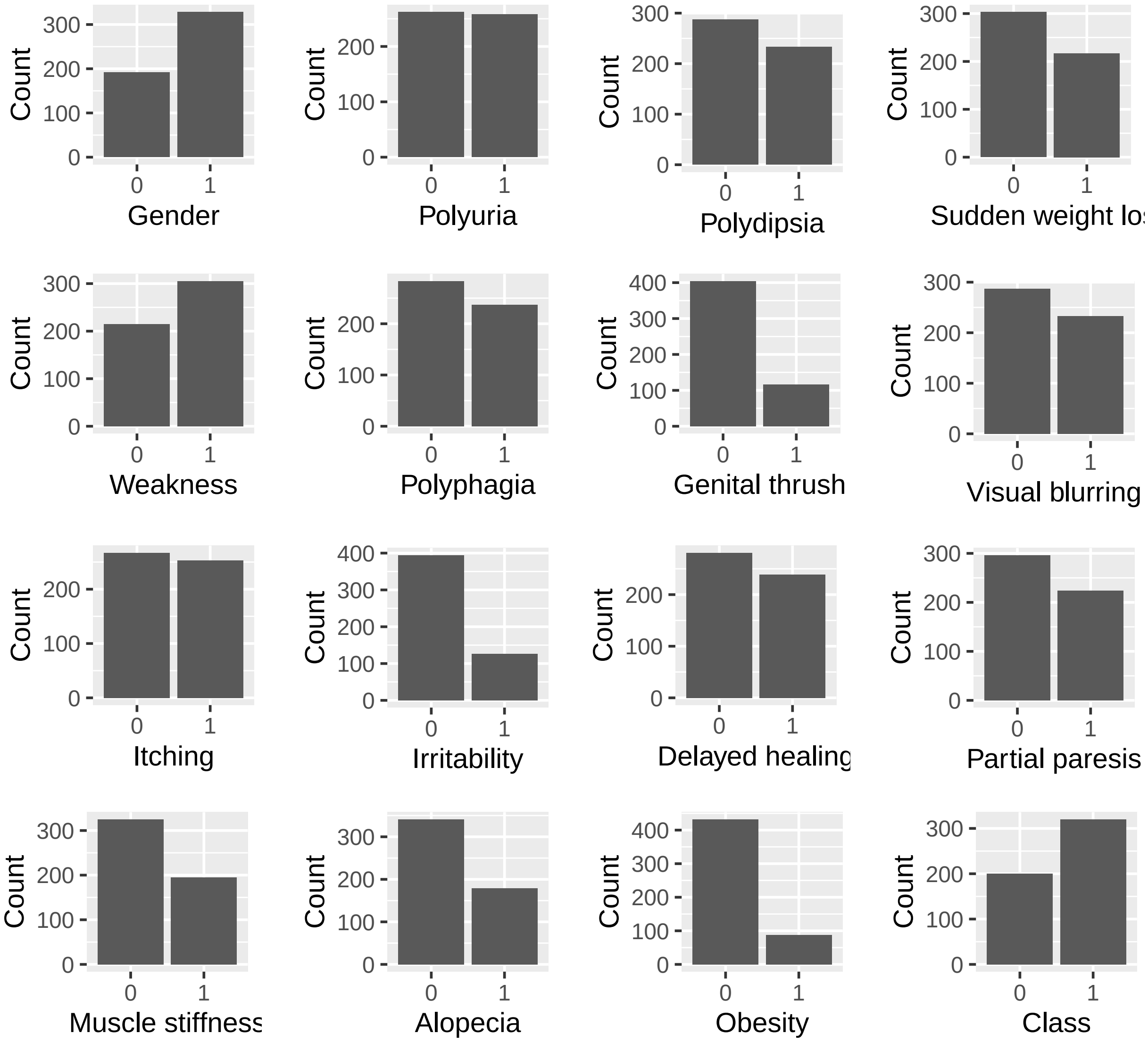

Figure 4: Bar charts of categorical variables

Code

```
library(knitr)
library(tidyverse)


# ----- DATA CLEANSING ----- #
diabetes_raw <- read.csv("diabetes_data.csv", sep=";")
diabetes <- diabetes_raw %>%
  mutate(gender = as.factor(ifelse(gender=="Male",1,
                            ifelse(gender=="Female",0,
                            NA))),
         polyuria = as.factor(polyuria),
         polydipsia = as.factor(polydipsia),
         sudden_weight_loss = as.factor(sudden_weight_loss),
         weakness = as.factor(weakness),
         polyphagia = as.factor(polyphagia),
         genital_thrush = as.factor(genital_thrush),
         visual_blurring = as.factor(visual_blurring),
         itching = as.factor(itching),
         irritability = as.factor(irritability),
         delayed_healing = as.factor(delayed_healing),
         partial_paresis = as.factor(partial_paresis),
         muscle_stiffness = as.factor(muscle_stiffness),
         alopecia = as.factor(alopecia),
         obesity = as.factor(obesity),
         class = as.factor(class))

# ----- DATA EXPLORATION ----- #
# Data description
attribute <- c("Age","Gender","Polyuria","Polydipsia",
               "Sudden weight loss","Weakness","Polyphagia",
               "Genital thrush","Visual blurring","Itching",
               "Irritability","Delayed healing","Partial paresis",
               "Muscle stiffness","Alopecia","Obesity","Class")
values <- c("In years","1 = Male, 0 = Female","1 = Yes, 0 = No","1 = Yes, 0 = No",
            "1 = Yes, 0 = No","1 = Yes, 0 = No","1 = Yes, 0 = No","1 = Yes, 0 = No",
            "1 = Yes, 0 = No","1 = Yes, 0 = No","1 = Yes, 0 = No","1 = Yes, 0 = No",
            "1 = Yes, 0 = No","1 = Yes, 0 = No","1 = Yes, 0 = No","1 = Yes, 0 = No",
            "1 = Positive risk, 0 = Negative risk")
```

```r
data_summary <- data.frame(Attribute=attribute, Values=values)
kable(data_summary)

# Check for missing data
nulls <- sapply(diabetes, function(col){ifelse(is.na(sum(col == "")), 0, sum(col == ""))})
blanks <- sapply(diabetes, function(col){ifelse(is.na(sum(col == "")), 0, sum(col == ""))})
kable(data.frame(Nulls=sum(nulls), Blanks=sum(blanks)))

# Boxplot of continuous variable
bplot <- ggplot(diabetes, aes(y = age)) + geom_boxplot() + labs(x="",y="Age") +
  theme(axis.text.x=element_blank(), axis.ticks.x=element_blank())
bplot_a1 <- as.integer(unlist(ggplot_build(bplot)$data)["ymax"])
bplot + geom_hline(yintercept = bplot_a1, linetype="dotted", color="red") +
  geom_text(aes(-0.4,bplot_a1,label = bplot_a1, vjust = 1.5, color="red"),
            show.legend = FALSE)
# Barplots for each categorical variable
ggplot(diabetes, aes(x = gender)) + geom_bar() + labs(y = "Count", x = "Gender")
ggplot(diabetes, aes(x = polyuria)) + geom_bar() + labs(y = "Count", x = "Polyuria")
ggplot(diabetes, aes(x = polydipsia)) + geom_bar() + labs(y = "Count", x = "Polydipsia")
ggplot(diabetes, aes(x = sudden_weight_loss)) + geom_bar() +
  labs(y = "Count", x = "Sudden weight loss")
ggplot(diabetes, aes(x = weakness)) + geom_bar() + labs(y = "Count", x = "Weakness")
ggplot(diabetes, aes(x = polyphagia)) + geom_bar() + labs(y = "Count", x = "Polyphagia")
ggplot(diabetes, aes(x = genital_thrush)) + geom_bar() +
  labs(y = "Count", x = "Genital thrush")
ggplot(diabetes, aes(x = visual_blurring)) + geom_bar() +
  labs(y = "Count", x = "Visual blurring")
ggplot(diabetes, aes(x = itching)) + geom_bar() + labs(y = "Count", x = "Itching")
ggplot(diabetes, aes(x = irritability)) + geom_bar() +
  labs(y = "Count", x = "Irritability")
ggplot(diabetes, aes(x = delayed_healing)) + geom_bar() +
  labs(y = "Count", x = "Delayed healing")
ggplot(diabetes, aes(x = partial_paresis)) + geom_bar() +
  labs(y = "Count", x = "Partial paresis")
ggplot(diabetes, aes(x = muscle_stiffness)) + geom_bar() +
  labs(y = "Count", x = "Muscle stiffness")
ggplot(diabetes, aes(x = alopecia)) + geom_bar() + labs(y = "Count", x = "Alopecia")
ggplot(diabetes, aes(x = obesity)) + geom_bar() + labs(y = "Count", x = "Obesity")
ggplot(diabetes, aes(x = class)) + geom_bar() + labs(y = "Count", x = "Class")
```

References

Islam, M. M. F., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2020). Likelihood prediction of diabetes at early stage using data mining techniques. In M. Gupta, D. Konar, S. Bhattacharyya, & S. Biswas (Eds.), *Computer vision and machine intelligence in medical image analysis* (pp. 113–125). Springer Singapore. https://doi.org/10.1007/978-981-13-8798-2_12

Larxel. (2023). *Early classification of diabetes*. https://www.kaggle.com/datasets/andrewmvd/heart-failure-clinical-data/data

R Core Team. (2023). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

World Health Organization. (n.d.). *Diabetes*. https://www.who.int/health-topics/diabetes#tab=tab_1