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 the risk and prevent the onset of diabetes.

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

In Islam et al.'s study (2020), machine learning techniques were used to predict diabetes risk. In particular, naive Bayes, logistic regression, and random forest techniques were applied. The study found that the random forest model had the best accuracy (Islam et al., 2020). Other studies have also applied machine learning to predict diabetes and varying methods were found to be most accurate. For example, Kumar and Velide compared seven different techniques and found that their J48(C4.5) model performed most accurately (2014) while Rabina and Chopra compared a decision tree with neural networks and found the decision tree to be more accurate (2016).

The study we propose will also utilize machine learning techniques to predict diabetes. It will focus on decision tree and neural network methods.

Methods

The first machine learning method will be a decision tree. This method was selected because Islam et al.'s study found that a decision tree model was most accurate (2020) and so this study will aim to reproduce such results. The index on the subtrees α will be tuned using cross-validation. Pruning will be applied to reduce the cost complexity of the tree, and random forest ensembling will be used to improve model performance.

The second method will be a neural network. This method was selected since Islam et al.'s study has not explored a model using neural networks (2020) and so there is a possibility that a neural network may be more accurate than a tree-based model. The number of hidden layers and the units per layer will be tuned through trial and error. Based on James et al.'s book on statistical

learning (2021), the units per layer will be set to some large value and overfitting will be controlled with ridge regularization. The strength of the regularization λ will be tuned at each layer.

After the two models are built, they will each be assessed using misclassification rate, accuracy, specificity, and sensitivity. A comparison of these four measurements will reveal which model is a stronger fit to predict diabetes. By nature, decision trees are easier to interpret and reproducible compared to neural networks. Thus, these qualities will also be considered in its comparison.

Preliminary Analysis

This section involves an exploratory analysis to provide insight on the data prior to machine learning.

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 to ease the analysis. The response variable is a binary attribute called class that indicates whether the patient has a positive or negative risk for diabetes. The remaining attributes describe the patient and if they experience common 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.

A correlation plot was created to check if there are any strong correlations between the attributes. We define a strong correlation as those with a correlation coefficient of 0.7 or larger. Based on Figure 3, all values are lower than 0.7 so there is no evidence of strong correlations.

Next, each of the individual attributes were explored. Figure 4 shows a box plot of patient ages. The median age in the population is 47. There are a few outliers that exist outside of the interquartile range. These values will be capped at 79, the upper bound of the range.

Figure 5 shows a bar chart for each of the 16 categorical variables in the data set. There appears to be a somewhat even split between the predictor values for polyuria and itching, while the remaining predictors are not evenly split. This is especially important for the class variable because there are

about 200 observations with a negative diabetes risk and about 300 observations with a positive risk. This suggests that if the model does not perform well, a resampling method such as cross-validation or bootstrapping may help improve the model.

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

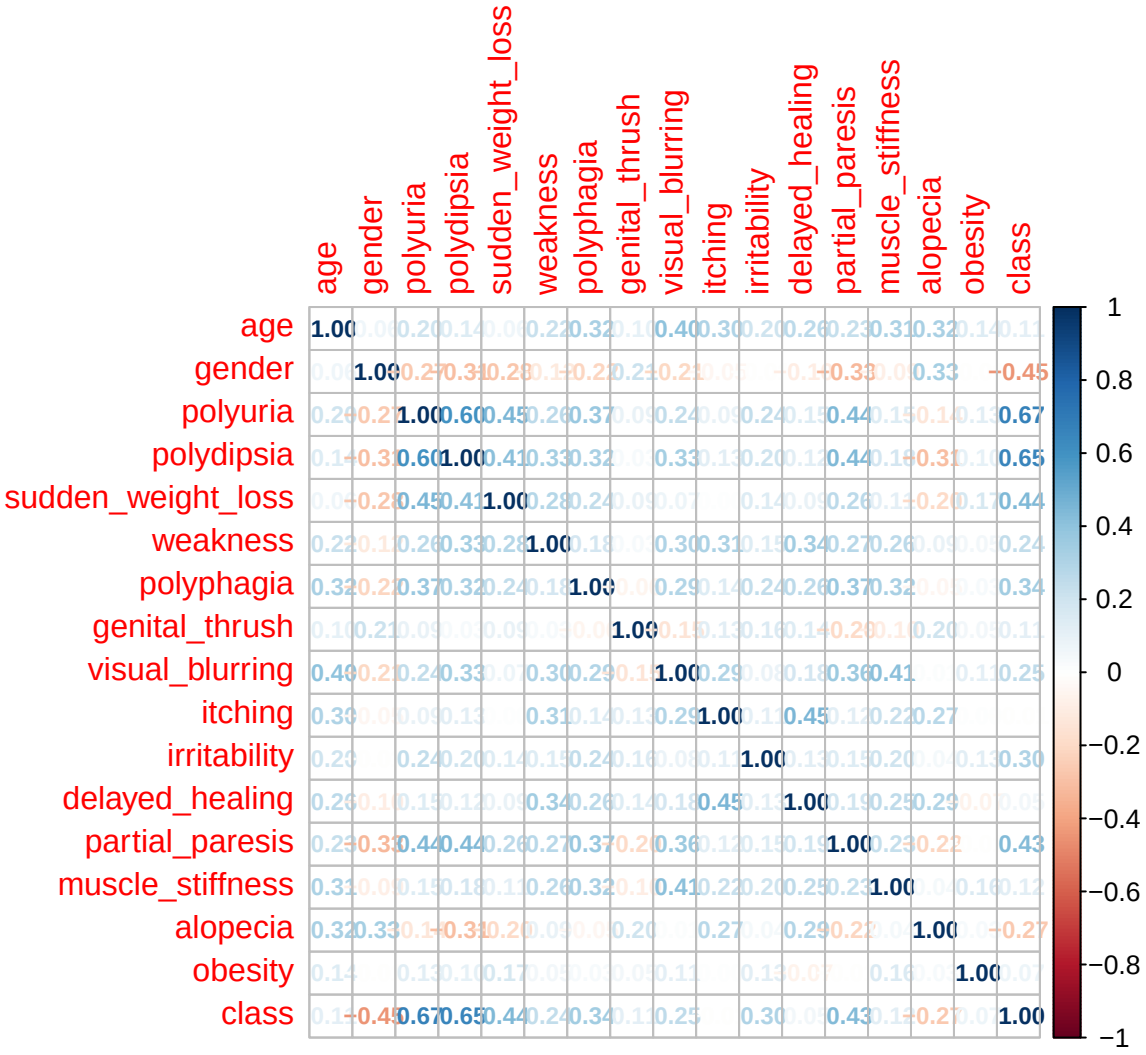

Figure 3: Correlation plot

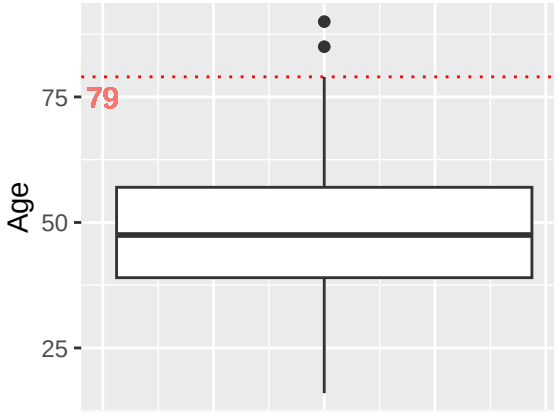

Figure 4: Boxplot of age

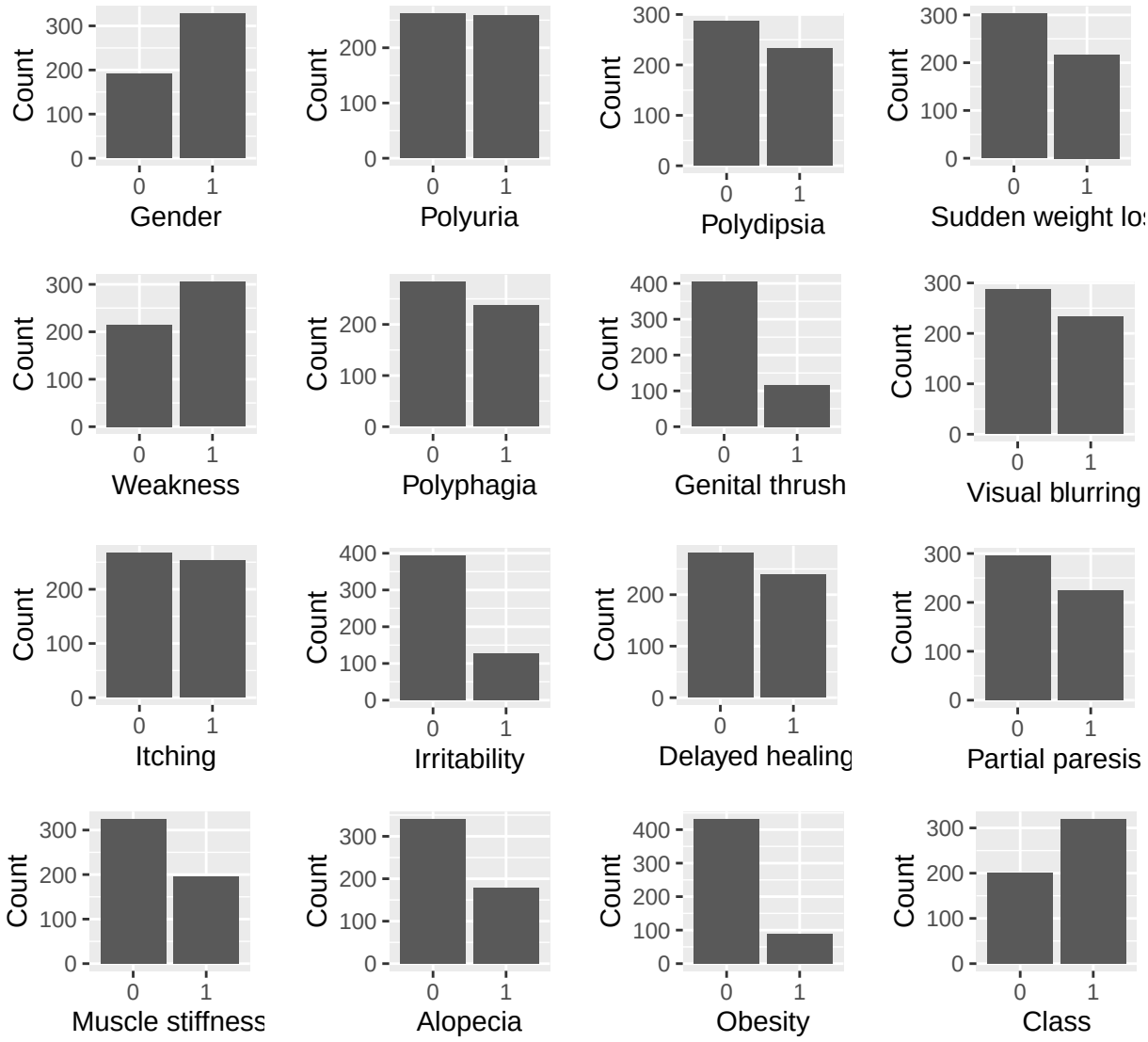

Figure 5: Bar charts of categorical variables

Code

```
library(knitr)
library(tidyverse)
library(corrplot)


# ----- DATA CLEANSING ----- #
diabetes_raw <- read.csv("diabetes_data.csv", sep=";")
```

```r
diabetes_int <- diabetes_raw %>%
  mutate(gender = as.integer(ifelse(gender=="Male",1,
                             ifelse(gender=="Female",0,
                             NA))))
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")
```

```r
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")
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)

# Cap outliers
diabetes_no_outliers <- diabetes %>% mutate(age = ifelse(age > bplot_a1, bplot_a1, age))

# 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")
```

```r
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")

# Correlation plot
corr_matrix <- cor(diabetes_int)
corrplot(round(corr_matrix,2), method = "number", number.cex=0.75)
```

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

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). Deep learning. In *An introduction to statistical learning with applications in r 2nd edition* (pp. 403–458). Springer.

Kumar, V., & Velide, L. (2014). *A data mining approach for prediction and treatment of diabetes disease.*

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/

Rabina, & Chopra, Er. A. (2016). *Diabetes prediction by supervised and unsupervised learning with feature selection.*

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