Programming with R

Group Project

Patricija Milaseviciute - u746955
Anouk Heemskerk - u684364
Emily Wijker - u214079
Yuexia Tian - u976100

1. INTRODUCTION

According to the World Health Organization (2020), the number of people with diagnosed diabetes is rising. Diabetes is a condition that causes high blood sugar levels which could lead to several health problems. It is even estimated to be the seventh leading cause of death in 2016 (World Health Organization, 2020). However, diabetes often goes undiagnosed for years (HonorHealth, n.d.). Therefore, it is important to distinguish the most important symptoms of diabetes in order to be able to recognize diabetes earlier. This leads to the following research question:

Which symptoms contribute most in predicting diabetes?

To answer this question, a dataset about early stage diabetes risk predictors (Islam, Ferdousi, Rahman, & Bushra, 2020) will be used. This data is collected via direct questionnaires to patients from Sylhet Diabetes Hospital in Bangladesh and is approved by a doctor. Before answering the question, some preprocessing and exploratory data analysis will be done. Next, different models will be fitted using the classifiers k-nearest neighbors, logistic regression and random forest. Lastly, the results will be interpreted and conclusions will be drawn.

2. METHODOLOGY

The exploratory data analysis was done firstly by looking at the data summary. There are 192 females and 328 males in the dataset in terms of gender division. Out of the total number, 200 people belonged to the negative class (no diabetes) and 320 people belonged to the positive class (diabetes). The minimum age of the people is 16, maximum is 90 and the mean age is 48. There is no missing data in the dataset. All column names were changed into lower case for consistency.

When exploring the age distribution using “ggplot2” (Wickham, 2016), it becomes clear that most of the people are between the ages of 25 and 75 (Appendix A1). Additionally, a boxplot was used to inspect the item age for any outliers. Some outliers were identified above

80 (Appendix A2). These outliers were removed from the dataset. Next, the age distribution was separated between males and females, to see whether there are any visible differences that could impact the results. The distributions turned out to be similar (Appendix A3).

Moreover, a barplot was used to visualize the difference between males and females belonging to two diabetes classes - positive and negative (Appendix A4). Although the positive class seemed to be distributed equally, the negative class was heavily outweighed by males.

Finally, to have an indication about which symptoms are more often possessed by people with diabetes, the data was filtered using “dplyr” (Wickham, 2020) to look only at the positive class. This showed how many people in the positive class had specific symptoms, versus the amount of people that did not have those symptoms (Appendix A5). This revealed interesting insights. There were more people in the positive class who had polyuria, polydipsia, sudden weight loss, weakness, polyphagia, visual blurring and partial paresis. Visual blurring, itching and delayed healing were quite equally distributed, with around half of the people having these symptoms, and the other half not. Finally, there were significantly more people in the positive class without genital thrush, irritability, muscle stiffness, alopecia, and most surprisingly, obesity.

3. RESULTS

The first classifier used to analyze the data is the k-nearest neighbor algorithm using the package “caret” (Kuhn, 2020). This algorithm is a supervised machine learning algorithm that assumes that similar data points are close to each other. Thus, classification is done by predicting that every observation belongs to the same class as the ‘k’ number of closest neighbors. First, we have to split our data into a train and a test set. 70 percent was used in the train set and 30 percent of the data was used for the test set. After that we use KNN to predict diabetes (positive/negative) with the symptoms in the dataset. We have applied five fold cross validation and set k between three and ten with steps of one. The model shows that k = 3 is the best k with an accuracy of 0.91 (Appendix B1). That means that this model can accurately predict around 91 percent of the instances with three closest neighbors. After running this model, we applied it to the unseen test set and used a confusion matrix (Appendix B1). Out of 154 persons 94 had diabetes and the model predicted 84, so the accuracy of this model is 94 percent. Moreover, the sensitivity of the model is 98 percent and the specificity is 90 percent.

Secondly, logistic regression is used to fit a model with “caret” as well (Kuhn, 2020). Logistic regression is a form of regression specific for classification. The model assumes linear relationships between the explanatory variables and the logarithm of the odds of one of the

categories of the binary dependent variable. In our case, we try to predict diabetes (positive/negative) with the symptoms as explanatory variables. With our model, we want to maximize the number of true positives so we focus on recall. It is decided to apply five fold cross validation to train and test the classifier. The logistic regression model shows a recall of 81 percent (Appendix B1), which means the model will correctly find 81 percent of the true positives. According to the summary of the model, the strongest predictors for diabetes are polyuria and polydipsia with estimated coefficients of 3.6 and 3.1 respectively (Appendix B2). Both predictors are also highly significant ($p = 0.000$). To evaluate the performance of this model, we apply it to the unseen test set and investigate the confusion matrix (Appendix B2). Out of 154 people in our test set, 95 were predicted to have diabetes, and of those, 84 actually did. In conclusion, the output shows that our model has an accuracy of 86 percent. Moreover, the sensitivity and thus the recall of the model is almost 82 percent.

Lastly, the package “randomForest” (Liaw & Wiener, 2002) is used for analyzing the data applying random forest. In our first model, the number of trees is 500 and the number of variables tried at each split is 4 by default. The error rate is 2.75 percent (Appendix B3). In the second model, the number of variables tested at each split is increased to 6 in an attempt to optimize tuning parameters, which results in a decreased error rate of 2.47 percent and sensibility of 100 percent (Appendix B4). Furthermore, for this optimized model, the accuracy of prediction on train dataset is 100 percent with no misclassification while in the case of test dataset 3 data points are misclassified and accuracy is 98.08 percent. In general, the random forest model outperforms the KNN model and the logistic regression model above in terms of prediction accuracy and recall, which shows a strong power of random forest model for doing prediction. Lastly, the function importance is used to check important variables (Appendix B5). Among all variables, polyuria has a highest MeanDecreaseGini which means that polyuria plays the biggest role in reducing node impurity and in splitting the data into the defined classes. In other words, polyuria is the strongest predictor of diabetes, followed by polydipsia.

4. CONCLUSION

After performing some preprocessing and exploring the data, different models have been fitted to find out which symptoms contribute most in predicting diabetes. First, a KNN model was used with an accuracy of 94 percent, sensitivity of 98 percent and specificity of 90 percent. Thus, we can conclude that the symptoms are indeed a good predictor of diabetes in general. To find out which symptoms contribute most, a logistic regression model and random forest were used. The logistic regression model showed an accuracy of 86 and a recall of 82 percent and returned

polyuria and polydipsia as strongest predictors for diabetes. The random forest model had the highest accuracy and recall, namely 98 and 100 percent respectively. This model also showed that polyuria and polydipsia are the strongest predictors of diabetes. Therefore, the conclusion and answer to the research question is that having these two symptoms could potentially serve as a warning that someone might have diabetes. It is important to note however, that the sample size used was small and therefore more research is needed in order to be able to draw more accurate conclusions in predicting diabetes.

REFERENCES

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Hadley Wickham (2020). tidyr: Tidy Messy Data. R package version 1.1.2. https://CRAN. R-project.org/package=tidyr

Hadley Wickham, Romain François, Lionel Henry and Kirill Müller (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package= dplyr

HonorHealth. (n.d.). *Signs, symptoms and diagnosis of diabetes*. Retrieved December 22, 2020, from https://www.honorhealth.com/medical-services/diabetes/signs- symptoms-diagnosis

Islam, M. M., Ferdousi, R., Rahman, S., & Bushra, H. Y. (2020). Likelihood prediction of diabetes at early stage using data mining techniques. *Computer Vision and Machine Intelligence in Medical Image Analysis*, *992*, 113–125. https://doi.org/10.1007/978-981-13-8798-2_12

Liaw, A and Wiener, M (2002). Classification and Regression by randomForest. R News 2(3), 18--22.

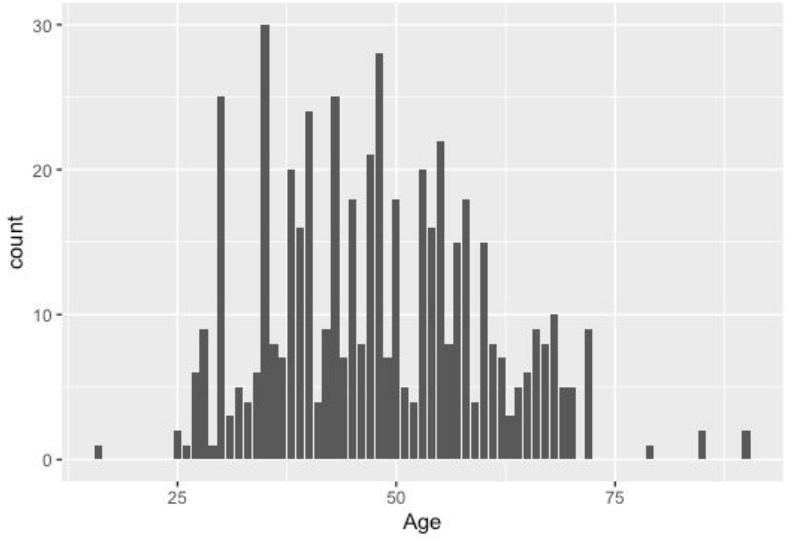
Max Kuhn (2020). caret: Classification and Regression Training. R package version 6.0-86. https://CRAN.R-project.org/package=caret

World Health Organization. (2020, June 8). *Diabetes*. Retrieved December 22, 2020, from https://www.who.int/news-room/fact-sheets/detail/diabetes

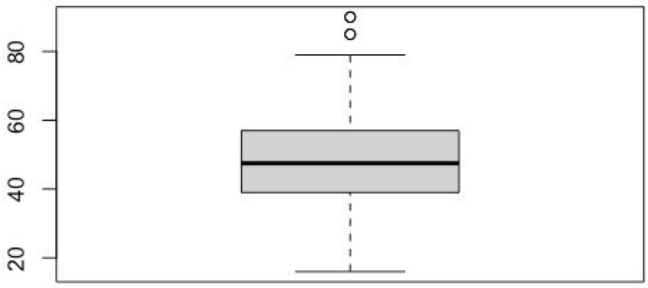
APPENDICES

Appendix A: Visualizations

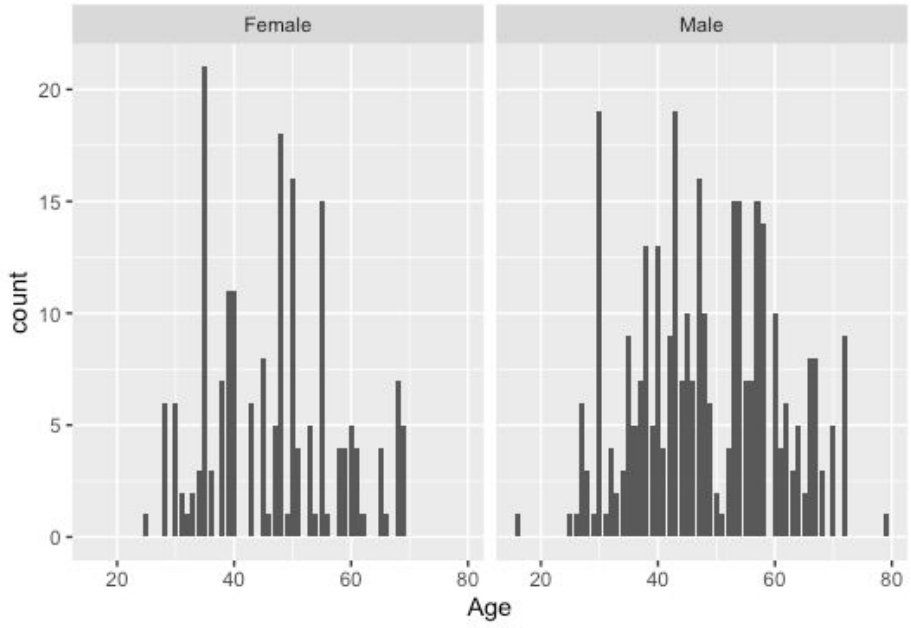
A1: Age Distribution



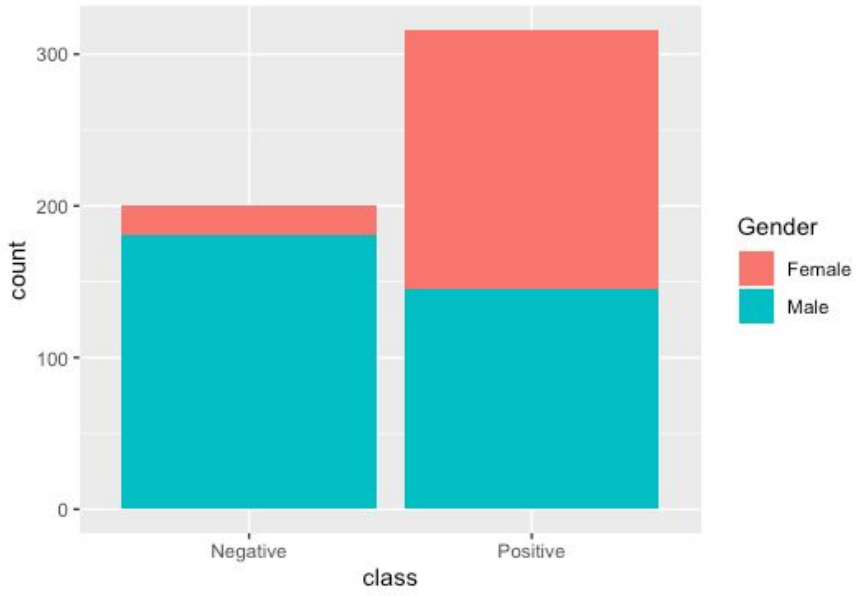
A2: Age Boxplot for Outlier Detection



A3: Age Distributions by Gender



A4: Gender Distribution by Class



A5: Summary of the Positive Class

```
      age           gender    polyuria  polydipsia sudden.weight.loss weakness  polyphagia
 Min.   :16.00   Female:171   No : 75   No : 95    No :132            No :100   No :129
 1st Qu.:39.00   Male  :145   Yes:241   Yes:221    Yes:184            Yes:216   Yes:187
 Median :48.00
 Mean   :48.59
 3rd Qu.:57.00
 Max.   :79.00
 genital.thrush visual.blurring itching   irritability delayed.healing partial.paresis
 No :237        No :145         No :166   No :206      No :165         No :126
 Yes: 79        Yes:171         Yes:150   Yes:110      Yes:151         Yes:190



 muscle.stiffness alopecia  obesity        class
 No :185          No :240   No :255   Negative:  0
 Yes:131          Yes: 76   Yes: 61   Positive:316
```

Appendix B: Relevant output

B1: K-Nearest Neighbor Model

```
k-Nearest Neighbors

362 samples
 16 predictor
  2 classes: 'Negative', 'Positive'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 290, 290, 289, 289, 290
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   3  0.9061644  0.8074976
   4  0.8784627  0.7494765
   5  0.8536149  0.7013670
   6  0.8536149  0.7011003
   7  0.8425419  0.6802949
   8  0.8454718  0.6908568
   9  0.8426560  0.6862932
  10  0.8537291  0.7086507

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 3.
```

Confusion Matrix KNN

```
Confusion Matrix and Statistics

          Reference
Prediction Negative Positive
  Negative       59        9
  Positive        1       85

               Accuracy : 0.9351
                 95% CI : (0.8838, 0.9684)
    No Information Rate : 0.6104
    P-Value [Acc > NIR] : < 2e-16

                  Kappa : 0.8667

 Mcnemar's Test P-Value : 0.02686

            Sensitivity : 0.9833
            Specificity : 0.9043
         Pos Pred Value : 0.8676
         Neg Pred Value : 0.9884
             Prevalence : 0.3896
         Detection Rate : 0.3831
   Detection Prevalence : 0.4416
      Balanced Accuracy : 0.9438

       'Positive' Class : Negative
```

B2: Logistic Regression Model trained by Recall

```
Generalized Linear Model

362 samples
 16 predictor
  2 classes: 'Negative', 'Positive'

No pre-processing
Resampling: Cross-Validated (5 fold)
Summary of sample sizes: 290, 290, 289, 289, 290
Resampling results:

  AUC        Precision  Recall     F
  0.8140105  0.7605842  0.8142857  0.7858628
```

Summary of the model showing the intercept and coefficients

```
Call:
NULL

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.87836  -0.40815   0.03021   0.19971   2.13537

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.51637    0.33154  -4.574 4.79e-06 ***
polyuriaYes           3.60084    0.68494   5.257 1.46e-07 ***
polydipsiaYes         3.09804    0.64220   4.824 1.41e-06 ***
sudden.weight.lossYes 0.28844    0.48379   0.596 0.551038
weaknessYes           1.16043    0.49642   2.338 0.019410 *
polyphagiaYes         0.43614    0.44822   0.973 0.330533
genital.thrushYes     1.22993    0.51552   2.386 0.017042 *
visual.blurringYes    1.04680    0.60121   1.741 0.081656 .
itchingYes           -2.34461    0.61773  -3.796 0.000147 ***
irritabilityYes       0.98286    0.52218   1.882 0.059803 .
delayed.healingYes   -0.04844    0.54324  -0.089 0.928945
partial.paresisYes    1.12632    0.46483   2.423 0.015390 *
muscle.stiffnessYes  -1.28287    0.54158  -2.369 0.017848 *
alopeciaYes          -1.02084    0.56945  -1.793 0.073027 .
obesityYes           -0.21733    0.52081  -0.417 0.676466
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 483.10  on 361  degrees of freedom
Residual deviance: 190.65  on 347  degrees of freedom
AIC: 220.65

Number of Fisher Scoring iterations: 7
```

Confusion matrix

```
Confusion Matrix and Statistics

          Reference
Prediction Negative Positive
  Negative       49       10
  Positive       11       84

               Accuracy : 0.8636
                 95% CI : (0.7991, 0.9136)
    No Information Rate : 0.6104
    P-Value [Acc > NIR] : 4.151e-12

                  Kappa : 0.7124

 Mcnemar's Test P-Value : 1

            Sensitivity : 0.8167
            Specificity : 0.8936
         Pos Pred Value : 0.8305
         Neg Pred Value : 0.8842
             Prevalence : 0.3896
         Detection Rate : 0.3182
   Detection Prevalence : 0.3831
      Balanced Accuracy : 0.8551

       'Positive' Class : Negative
```

B3: Random Forest Model-1

```
Call:
 randomForest(formula = class ~ ., data = trn_diabetes, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 4

        OOB estimate of  error rate: 2.75%
Confusion matrix:
          Negative Positive class.error
Negative      135        5  0.03571429
Positive        5      219  0.02232143
```

B4: Random Forest Model-2

```
Call:
 randomForest(formula = class ~ ., data = trn_diabetes, ntree = 500,      mtry = 6, importa
nce = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 6

        OOB estimate of  error rate: 2.47%
Confusion matrix:
          Negative Positive class.error
Negative      136        4  0.02857143
Positive        5      219  0.02232143
```

B5: Random Forest Model-2 important variables

```
> importance(diabetes_rf2)
                  Negative  Positive MeanDecreaseAccuracy MeanDecreaseGini
Age               35.41448 16.511230             37.30391        18.180631
Gender            42.74521 39.494781             52.68423        19.881335
Polyuria          59.31965 37.201761             62.40795        41.850825
Polydipsia        60.62848 23.019866             53.47944        35.236613
sudden.weight.loss 16.18835 14.843531             18.48507         7.524994
weakness          12.56239 13.282083             17.08380         2.756491
Polyphagia        12.55475  7.491902             13.73507         2.724710
Genital.thrush    14.22034 11.663422             16.83935         3.877471
visual.blurring   14.23749 13.716054             18.24263         4.446113
Itching           18.37366  5.224225             18.37659         4.214959
Irritability      19.02855 16.605586             23.23854         6.572539
delayed.healing   22.74445 12.338748             24.59876         5.234314
partial.paresis   12.83995 13.814209             16.76839         4.439343
muscle.stiffness  16.02002  9.809892             17.62157         3.550763
```