DiabetesRiskPrediction

Magdalena Stefanowicz

2021-01-07

Introduction

This project was done for a course: HarvardX: PH125.9x Data Science: Capstone in HarvardX Professional Certificate Program in Data Science. The project can be found at: https://github.com/MagdalenaStefanowicz/DiabetesRiskPrediction

World Health Organization (WHO) estimates that diabetes was the seventh leading cause of death in 2016. The number of people with diabetes rose from 108 million in 1980 to 422 million in 2014 [1]. Being able to predict diabetes among patients could help many to avoid severe consequences of the disease as well as help in diagnosing it in good time.

This project aims to predict a risk of diabetes based on early stage factors among patients. The prediction is based on various diagnostic measurements included in the dataset: Early stage diabetes risk prediction dataset [2]. The dataset contains symptoms data of newly diabetic or would be diabetic patients. The data was collected with direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh. Dataset size: 520 observations (patients) and 17 variables.

The following key steps are performed in order to achieve classification of diabetes diagnosis: 1. dataset is analyzed in order to find the appropriate machine learning model 2. train set and test set are created 3. diabetes diagnosis are predicted using models with logisting regression and decision tree 4. confusion matrix is used to evaluate the approaches

Linear regression model can achieve accuracy of 95%.

Analysis

Read data

The following packages and the dataset are downloaded and installed.

Let's analyze the available data: variables and their classes. The variable "class" represents diabetes diagnosis with result values: Negative or Positive.

```
## # A tibble: 6 x 17
##     Age Gender Polyuria Polydipsia `sudden weight ~ weakness Polyphagia
##   <dbl> <chr>  <chr>    <chr>      <chr>            <chr>    <chr>
## 1    40 Male   No       Yes        No               Yes      No
## 2    58 Male   No       No         No               Yes      No
## 3    41 Male   Yes      No         No               Yes      Yes
## 4    45 Male   No       No         Yes              Yes      Yes
## 5    60 Male   Yes      Yes        Yes              Yes      Yes
```

```
## 6    55 Male   Yes      Yes        No               Yes      Yes
## # ... with 10 more variables: 'Genital thrush' <chr>, 'visual blurring' <chr>,
## #   Itching <chr>, Irritability <chr>, 'delayed healing' <chr>, 'partial
## #   paresis' <chr>, 'muscle stiffness' <chr>, Alopecia <chr>, Obesity <chr>,
## #   class <chr>
```

Diabetes dataset has 520 observations and 17 variables.

```
## [1] 520  17
```

There are no NA's values in the dataset. No cleaning is needed for further analyzis.

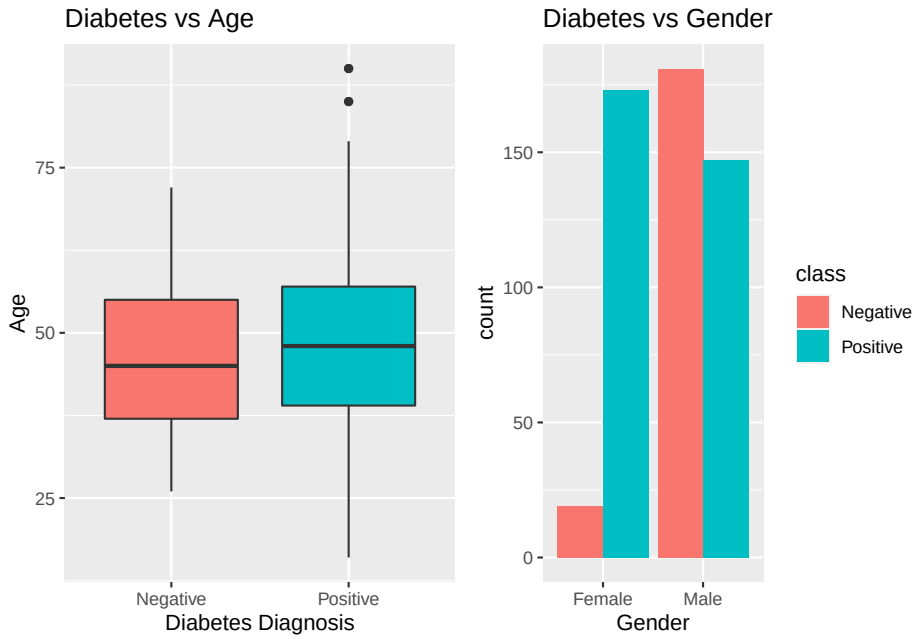
```
## [1] 0
```

Exploratory data analyzis

First, Let's start by checking distribution of positive diabetes diagnosis among the patients.

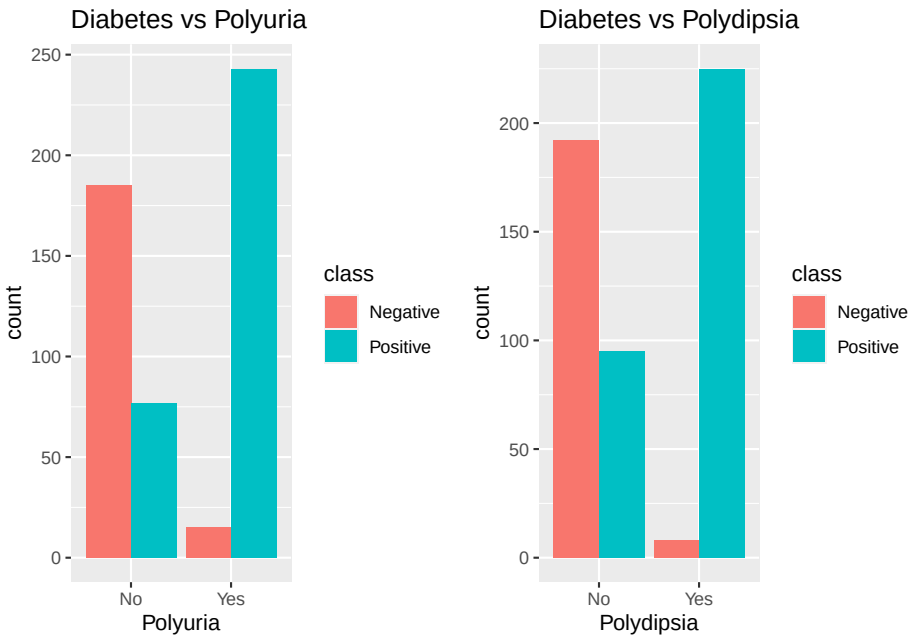
```
##           count
## class      200 320
##   Negative   1   0
##   Positive   0   1
```

Now we're going to analyze affect of each available parameters on the diabetes diagnosis - two by two. Starting from Age and Gender.



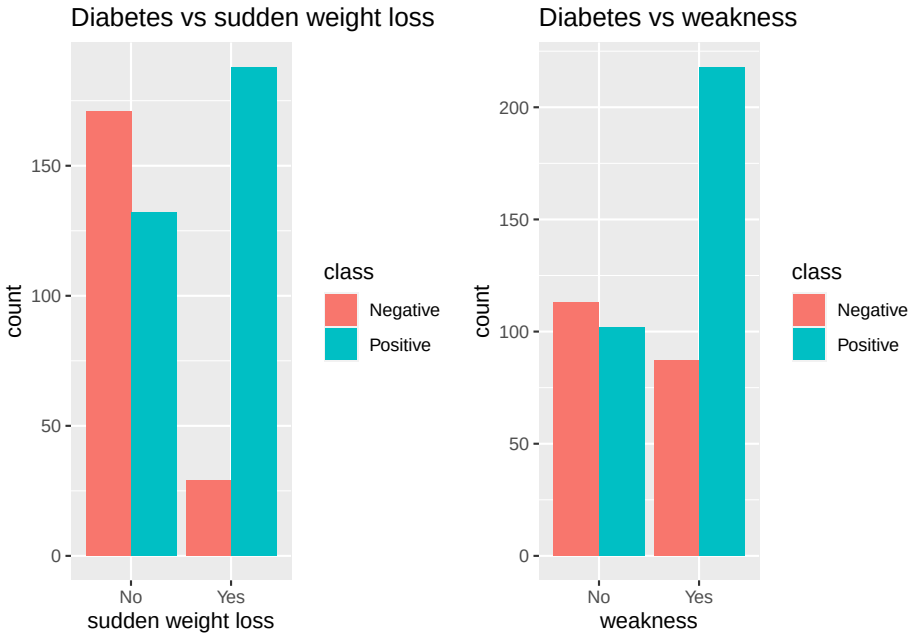
The boxplot for Age show little variation with diabetes. No clear correlation among the patients. The barplot for Gender shows that there are far more male participants then female. Female in the dataset are also far more likely to have diabetes.

Let's check how Polyuria and Polydipsia affect diabetes. Polyuria is defined as a urine output exceeding 3 L/day in adults and 2 L/m2 in children [3]. Polydipsia is defined as excessive thirst or excess drinking [4].



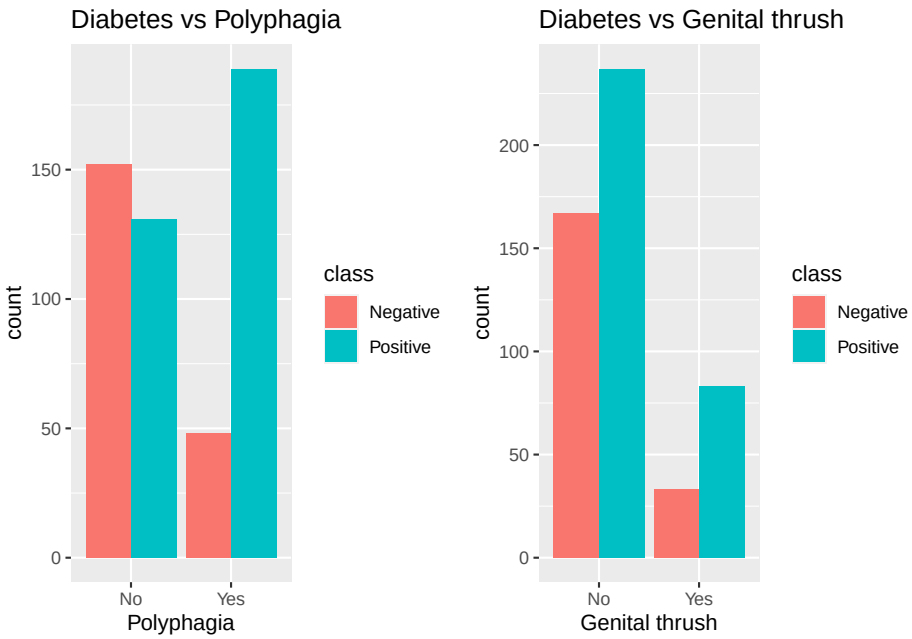
Very clear corelation between Polyuria and Diabetes. Almost all patients with Polydipsia does suffer from Diabetes. Very strong correlation as well.

Let's check how `sudden weight loss` and weakness affect diabetes.



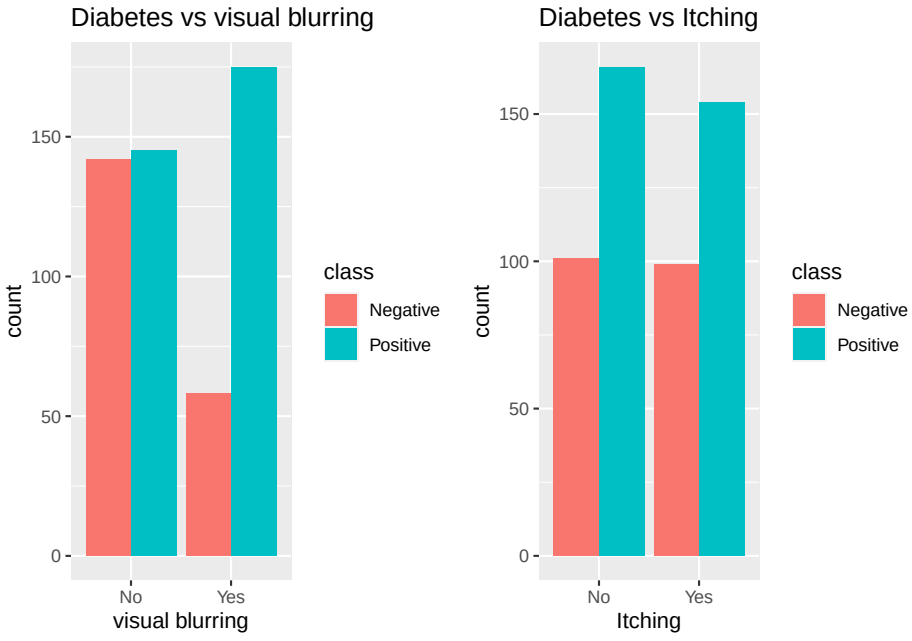
Patients with sudden weight loss do suffer from Diabetes more often. Clear correlation. No strong correlation between weakness and Diabetes.

Let's check how Polyphagia and 'Genital thrush' affect diabetes. Polyphagia is the medical term for excessive or extreme hunger [5]. 'Genital thrush'is a common vagina/penis condition caused by a type of yeast called Candida [6].



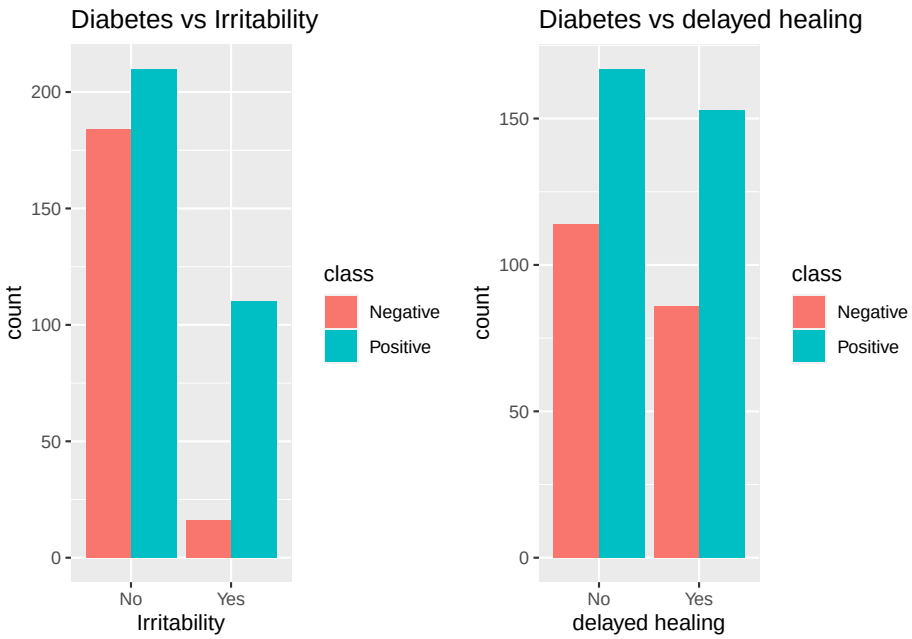
Most of the patients with Polyphagia does have Diabetes. Clear correlation. No correlaton between Genital thrush and Diabetes.

Let's check how `visual blurring` and Itching affect diabetes.



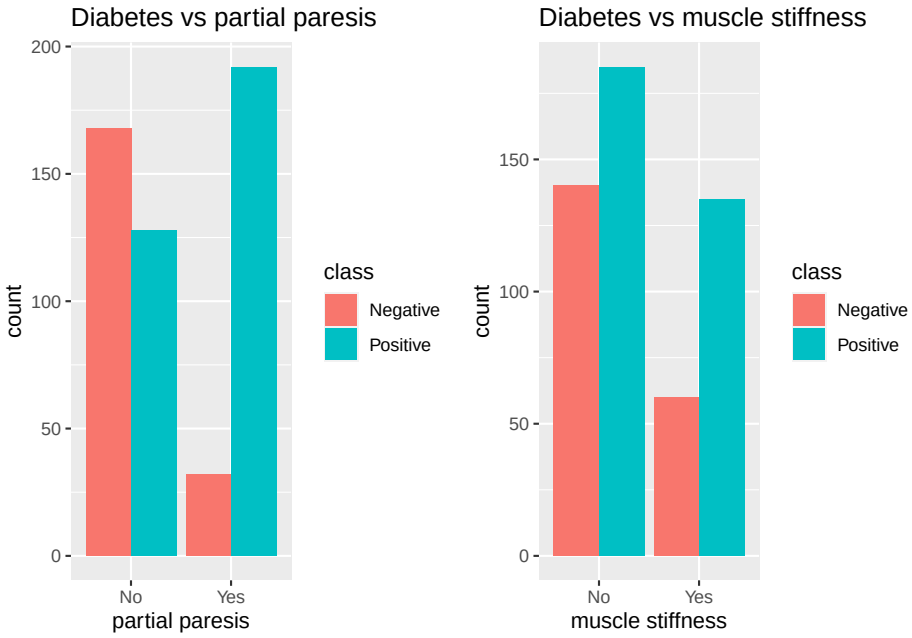
Patients with visual blurring does suffer from Diabetes somewhat more often. No correlation between Itching and Diabetes.

Let's check how Irritability and `delayed healing` affect diabetes.



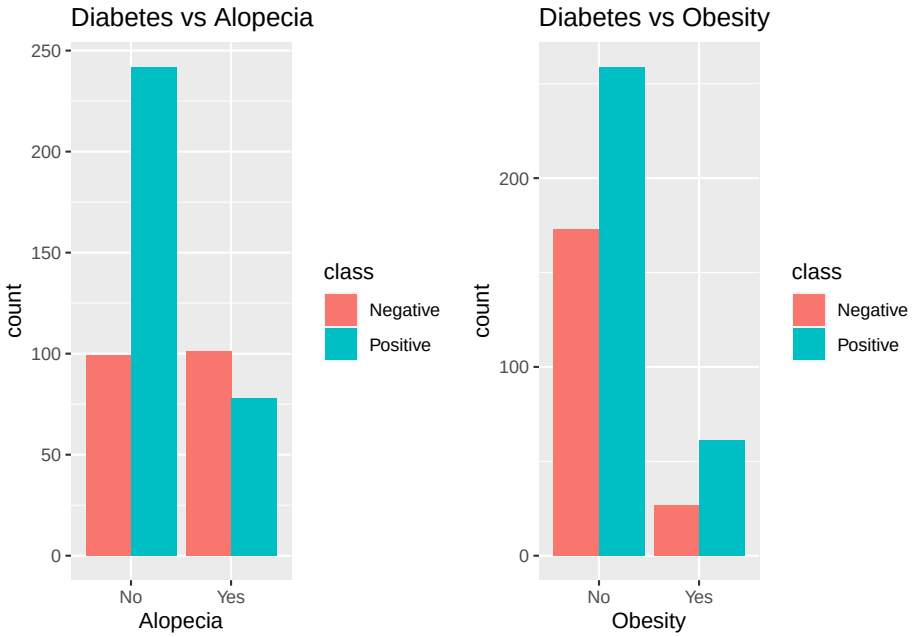
There seems to be a strong correlation between Irritability and Diabetes. No correlation between delayed healing and Diabetes.

Let's check how `partial paresis` and `muscle stiffness` affect diabetes.



Partial paresis has strong correlation with Diabetes. No significant correlation between muscle stiffness and Diabetes.

Let's check how Alopecia and Obesity affect diabetes. Alopecia is an autoimmune disorder that causes your hair to come out [7].

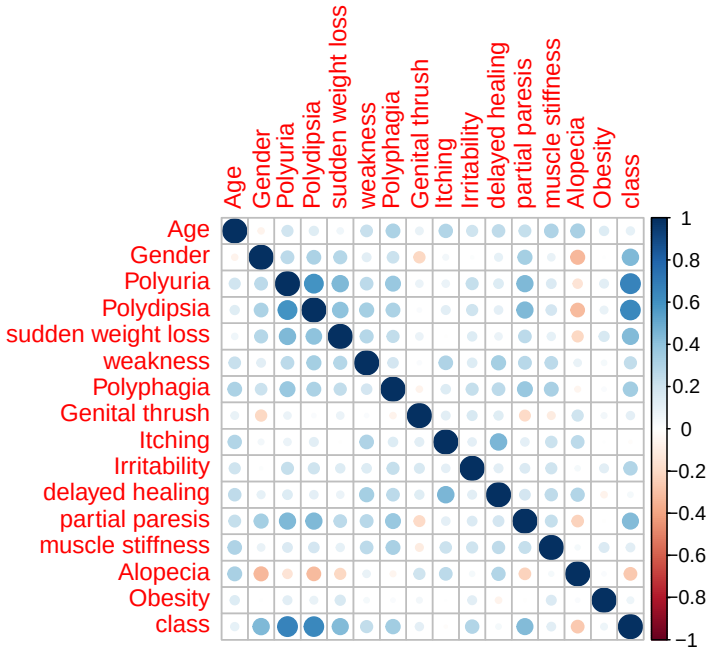


No clear correlation between Alopecia and Diabetes. What is very surprising, there is also no significant correlation between obesity and Diabetes among the patients in the dataset.

From the analyzis above it can be concluded that Polyuria and Polydipsia have biggest impact on 'class´ variable.

Let's encode the parameters from categorical to numerical values. By doing so we'll be able to identify and plot the correlation between all parameters.

Let's plot correlation between all the measurments.



Method

In this chapter we're going to fit two different machine learning models for the dataset: - logistic regression - decision tree.

Model setup

Let's start with dividing dataset into train and test sets. Test set will correspond to 20% of the dataset.

```
## [1] 416
```

```
## [1] 104
```

Model 0: Random sampling

Let's start with random sampling for which we assume equal probability for positive and negativ outcome of diabetes diagnosis. Probability would simply be 0.5.

Now we can make our first prediction using random sampling. As expected the accuracy is rather poor, only 45%.

```
## Confusion Matrix and Statistics
##
##
## rs_prediction  0  1
##             0 20 38
##             1 15 31
##
##                Accuracy : 0.4904
##                  95% CI : (0.391, 0.5903)
##     No Information Rate : 0.6635
##     P-Value [Acc > NIR] : 0.999906
##
##                   Kappa : 0.0178
##
##  Mcnemar's Test P-Value : 0.002512
##
##             Sensitivity : 0.5714
##             Specificity : 0.4493
##          Pos Pred Value : 0.3448
##          Neg Pred Value : 0.6739
##              Prevalence : 0.3365
##          Detection Rate : 0.1923
##    Detection Prevalence : 0.5577
##       Balanced Accuracy : 0.5104
##
##        'Positive' Class : 0
##
```

Model 1: Logistic Regression

Logistic regression model is used with consideration to all measurments. Logistic regression is considered an appropriate method for binary classification problems - problems with two class values.

```
##
## Call:
## glm(formula = class ~ Age + Gender + Polyuria + Polydipsia +
##     `sudden weight loss` + weakness + Polyphagia + `Genital thrush` +
##     `visual blurring` + Itching + Irritability + `delayed healing` +
##     `partial paresis` + `muscle stiffness` + Alopecia + Obesity,
##     family = binomial(link = "logit"), data = training)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1736  -0.2202   0.0008   0.0340   2.4630
##
## Coefficients:
##                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -0.95565    1.15567  -0.827 0.408281
## Age                  -0.07143    0.03088  -2.313 0.020700 *
## Gender                5.35821    0.81204   6.598 4.15e-11 ***
## Polyuria              6.03276    1.08242   5.573 2.50e-08 ***
## Polydipsia            6.41181    1.19991   5.344 9.11e-08 ***
## `sudden weight loss` -0.52320    0.70185  -0.745 0.455996
## weakness              0.88938    0.64070   1.388 0.165097
## Polyphagia            1.51915    0.63567   2.390 0.016855 *
## `Genital thrush`      1.57955    0.68689   2.300 0.021473 *
## `visual blurring`     0.93711    0.79162   1.184 0.236493
## Itching              -3.79135    0.98727  -3.840 0.000123 ***
## Irritability          2.30885    0.67256   3.433 0.000597 ***
## `delayed healing`    -0.68303    0.65995  -1.035 0.300678
## `partial paresis`     1.46162    0.65163   2.243 0.024896 *
## `muscle stiffness`   -1.13660    0.69959  -1.625 0.104234
## Alopecia              0.40830    0.78199   0.522 0.601582
## Obesity              -0.25550    0.60343  -0.423 0.671988
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 558.79  on 415  degrees of freedom
## Residual deviance: 125.85  on 399  degrees of freedom
## AIC: 159.85
##
## Number of Fisher Scoring iterations: 9
```

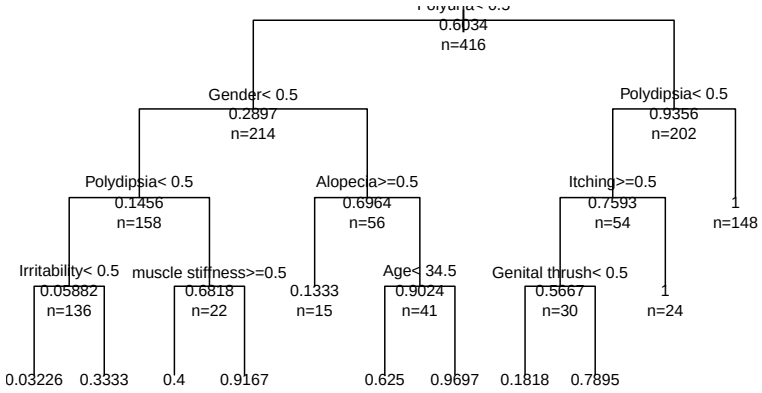
Gender, Polyuria, Polydipsia, Itching and Irritability are most releveant measurments - based on their low p_values.

Next, Confusion Matrix is used in order to measure performance of Logistic Regression model. Confusion matrix is a summary of prediction results and an appropriate approach for classification problems. Confusion Matrix allows to study sensitivity, specificity and balanced accuracy.

```
## Confusion Matrix and Statistics
##
##
## lr_prediction  0  1
##             0 33  5
##             1  2 64
##
##                Accuracy : 0.9327
##                  95% CI : (0.8662, 0.9725)
##     No Information Rate : 0.6635
##     P-Value [Acc > NIR] : 6.264e-11
##
##                   Kappa : 0.8524
##
##  Mcnemar's Test P-Value : 0.4497
##
##             Sensitivity : 0.9429
##             Specificity : 0.9275
##          Pos Pred Value : 0.8684
##          Neg Pred Value : 0.9697
##              Prevalence : 0.3365
##          Detection Rate : 0.3173
##    Detection Prevalence : 0.3654
##       Balanced Accuracy : 0.9352
##
##        'Positive' Class : 0
##
```

Model 2: Decision Tree

Decision tree is another model which can be built on categorical data. Decision tree predicts an outcome variable by partitioning the predictors. In the model below the complexity parameter (cp) has been adjusted in order to enlarge the tree. With relatively small sample size the lower cp does not effect the computing time but it can slightly improve the error.



If a person has Polydipsia and Polyuria then she is most likely to have diabetes.

Let's use Confusion Matrix as to measure performance of our Decision Tree model.

```
## Confusion Matrix and Statistics
##
##
## dt_prediction  0  1
##             0 33  7
##             1  2 62
##
##                Accuracy : 0.9135
##                  95% CI : (0.8421, 0.9597)
##     No Information Rate : 0.6635
##     P-Value [Acc > NIR] : 2.193e-09
##
##                   Kappa : 0.8128
##
##  Mcnemar's Test P-Value : 0.1824
##
##             Sensitivity : 0.9429
##             Specificity : 0.8986
##          Pos Pred Value : 0.8250
##          Neg Pred Value : 0.9687
##              Prevalence : 0.3365
##          Detection Rate : 0.3173
##    Detection Prevalence : 0.3846
##       Balanced Accuracy : 0.9207
##
##        'Positive' Class : 0
##
```

#Results

This is the summary table of performances for the two studied models: logistic regression and decision tree. Performance for random aampling is added for comparison.

```
##                Method Accuracy Sensitivity Specificity
## 1     Random sampling     0.49        0.57        0.45
## 2 Logistic regression     0.93        0.94        0.93
## 3       Decision tree     0.91        0.94        0.90
```

Conclusion

Both decision tree model and logistic regression model can achieve very good accuracy for the dataset. Logistic regression achieves the best result with accuracy of 91% and sensitivity of 95%.

The measurments importance for both models show that Polydipsia and Polyuria are particularly significant for diagnosing diabetes.

Following improvements for the models could be considered: - Dataset is somehow imbalanced with diagnosis distribution of ca. 2:3. Implementing techniques which help dealing with imbalanced data could improve the performance even more. - It is somehow doubtfull that there is no clear correlation between obesity and diabetes among patients in general. What might be true for this particular dataset might not be true (and most likely is not true) for datasets from other countries or other hospitals. Bigger dataset (sample size) would ensure higher credibility.

References

[1] World Health Organization: Diabetes https://www.who.int/news-room/fact-sheets/detail/diabetes. (2021-01-06) [2] UCI Machine Learning Repository: Early stage diabetes risk prediction dataset. https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset. (2021-01-06) [3] UpToDate: Evaluation of patients with polyuria https://www.uptodate.com/contents/evaluation-of-patients-with-polyuria. (2021-01-06) [4] Wikipedia: Polydipsia https://en.wikipedia.org/wiki/Polydipsia. (2021-01-06) [5] Diabetes.co.uk the global diabetes community: Polyphagia - Increased Appetite. https://www.diabetes.co.uk/symptoms/polyphagia.html (2021-01-06) [6] HealthyWA Health information for Western Australians: Thrush (genital) https://healthywa.wa.gov.au/Articles/S_T/Thrush-genital. (2021-01-06) [7] Wikipedia: Alopecia areata https://en.wikipedia.org/wiki/Alopecia_areata. (2021-01-06)