“Diabet”

Elena Oskrogo

02/01/2021

1 Introduction

For the second project of the “Capstone” course, I’ve chosen an “Early-stage diabetes risk prediction” dataset. The dataset location is UCI Machine learning Repository http://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.

After the “Movielens: movie prediction” exercise, I would like to explore a different domain. With the Diabetes dataset, I have opportunities to practice machine leanings methods in disease diagnostics.

This dataset contains the sign and symptom data of newly diabetic or would be diabetic patients. Data has been collected using direct questionnaires from Sylhet Diabetes Hospital patients in Sylhet, Bangladesh, and approved by a doctor. UCI repository posted data on July 2020.

I’ve downloaded the available dataset to my laptop (file diabetes_data_upload.csv). Dataset has 520 observations and 17 attributes.

Each observation represents information about one patient and contains followed attributes:

- attribute 1 **Age** - age of patient;
- attribute 2 **Gender** - with values Male or Female;
- attribute 3 - 16 various symptoms - with values Yes or No;
- attribute 17 **Class** - diagnosis - with values Positive or Negative

List of provided symptoms is **Polyuria**, **Polydipsia**, **Sudden_Weight_Loss**, **Weakness**, **Polyphagia**, **Genital_Thrush**, **Visual_Blurring**, **Itching**, **Irritability**, **Delayed_Healing**, **Partial_Paresis**, **Muscle_Stiffness**, **Alopecia**, **Obesity** .

Here is the preview of first 6 lines of dataset:

```
## # A tibble: 6 x 17
##     Age Gender Polyuria Polydipsia Sudden_Weight_L~ Weakness Polyphagia Genital_Thrush
##   <dbl> <chr>  <chr>    <chr>      <chr>            <chr>    <chr>      <chr>
## 1    40 Male   No       Yes        No               Yes      No         No
## 2    58 Male   No       No         No               Yes      No         No
## 3    41 Male   Yes      No         No               Yes      Yes        No
## 4    45 Male   No       No         Yes              Yes      Yes        Yes
## 5    60 Male   Yes      Yes        Yes              Yes      Yes        No
## 6    55 Male   Yes      Yes        No               Yes      Yes        No
## # ... with 9 more variables: Visual_Blurring <chr>, Itching <chr>, Irritability <chr>,
## #   Delayed_Healing <chr>, Partial_Paresis <chr>, Muscle_Stiffness <chr>, Alopecia <chr>,
## #   Obesity <chr>, Class <chr>
```

The project aims to predict class values based on age, gender, and medical symptoms.

First, I performed a descriptive data analysis to understood data, research missing values, clean, transform if needed, and identify trends.

Second, I split the original dataset into training and validation sets.

After, I built the models to predict the class value based on available predictors. I trained my models on the training data set and verified accuracy on the validation dataset. To compare different models, I used root mean squared error (RMSE) as a loss function and model overall accuracy.

As I inserted variable values directly in the report text, I did not use the standard knit menu to create pdf output, but instead, I used command *rmarkdown::render("file_name")* to compile in pdf output.

Lastly, I provided some conclusions about my findings and suggestions regarding future development.

2 Methods/ analysis

2.1 Original data overview, detail dataset description

Dataset does not require data transformation: each line represents observation for one patient, and each attribute has only one information.

First, I've reviewed all attributes to find any missing values (N/A or null). I did not found any anomalies.

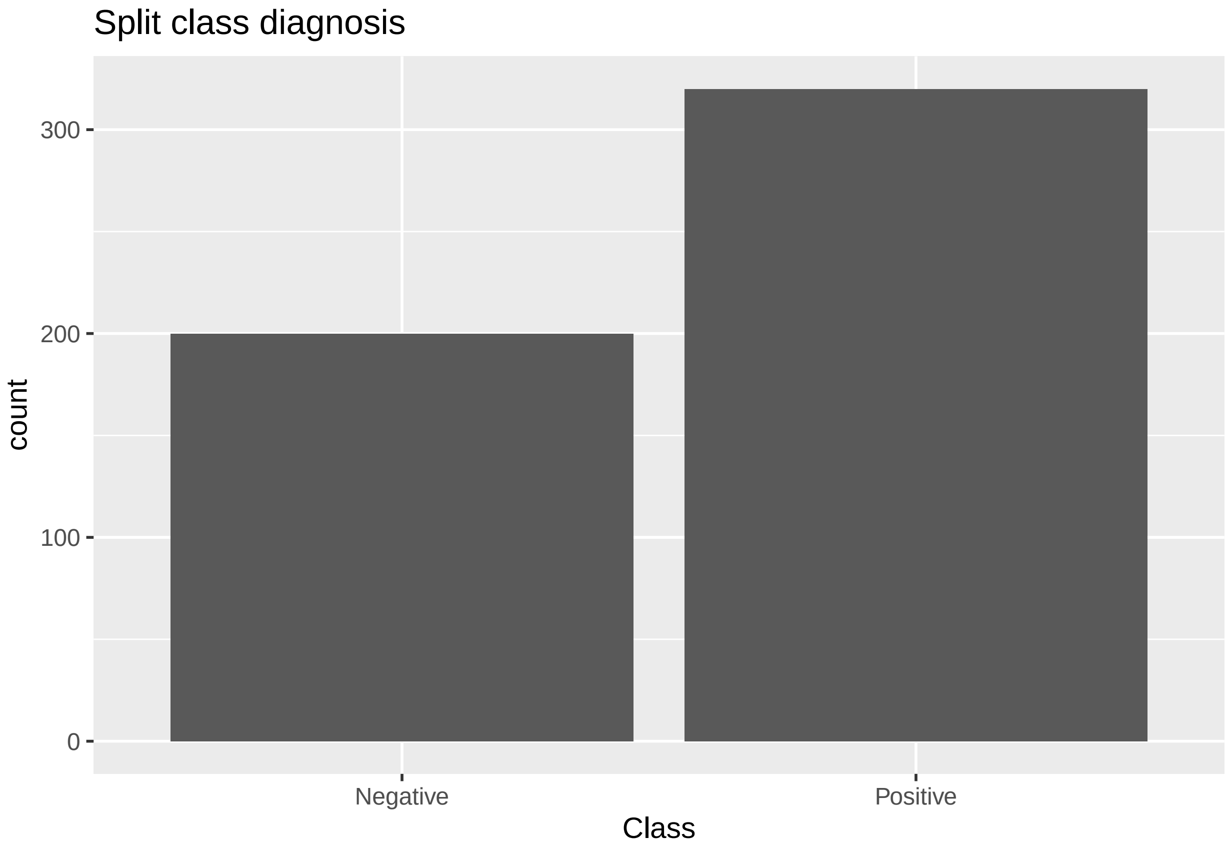
I've also checked that data does not have an incredible value as very high or negative age values, or symptoms have other matters than "Yes" or "No", or Class has different values than "Positive" or "Negative". Again, I did not found any anomalies.

So the dataset is ready for analysis.

The only transformation I've applied to columns name: purely cosmetic one - start each word with capitals letter to have the same visual presentation. I also replace " " by "_", as I experienced a problem with one function during the analysis stage.

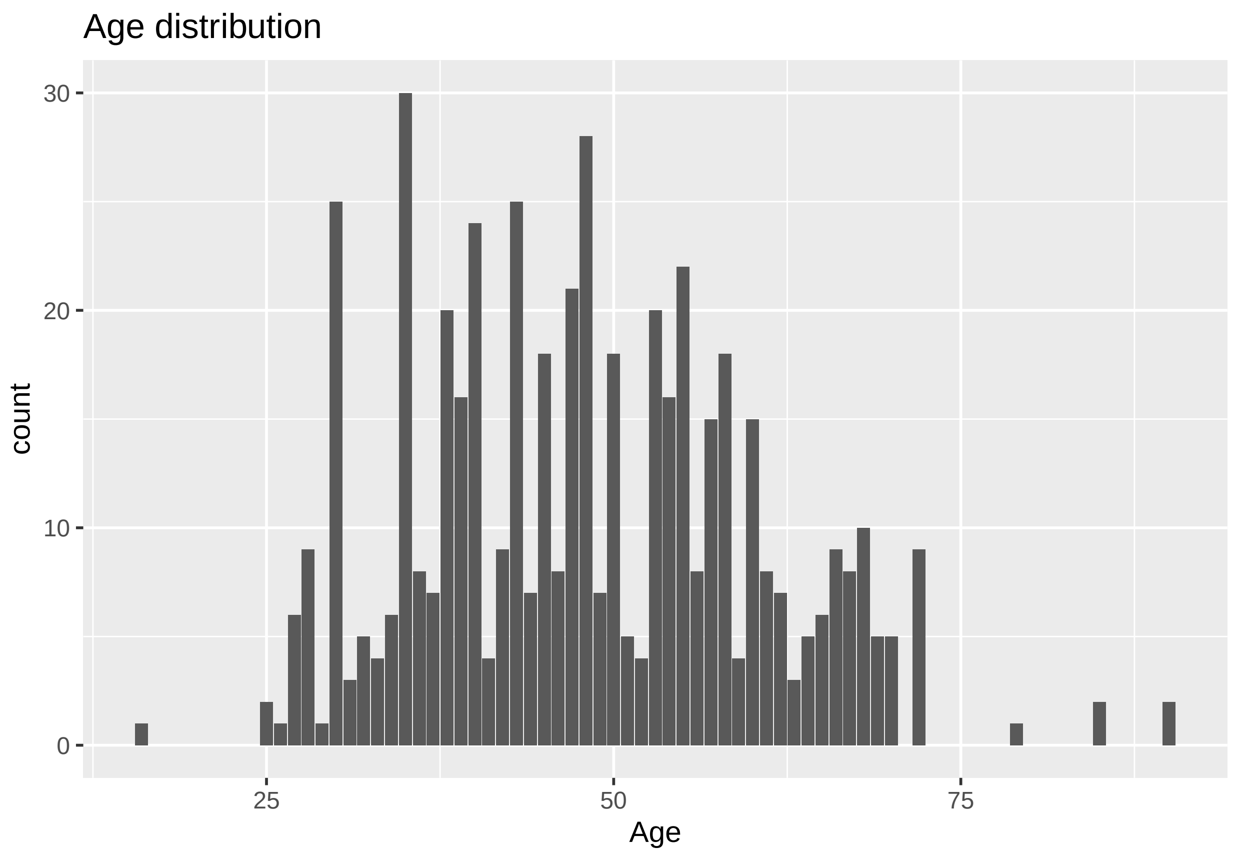
Let's first review each attribute individually.

Here is a general distribution between patients diagnosed with diabetes and healthy ones:

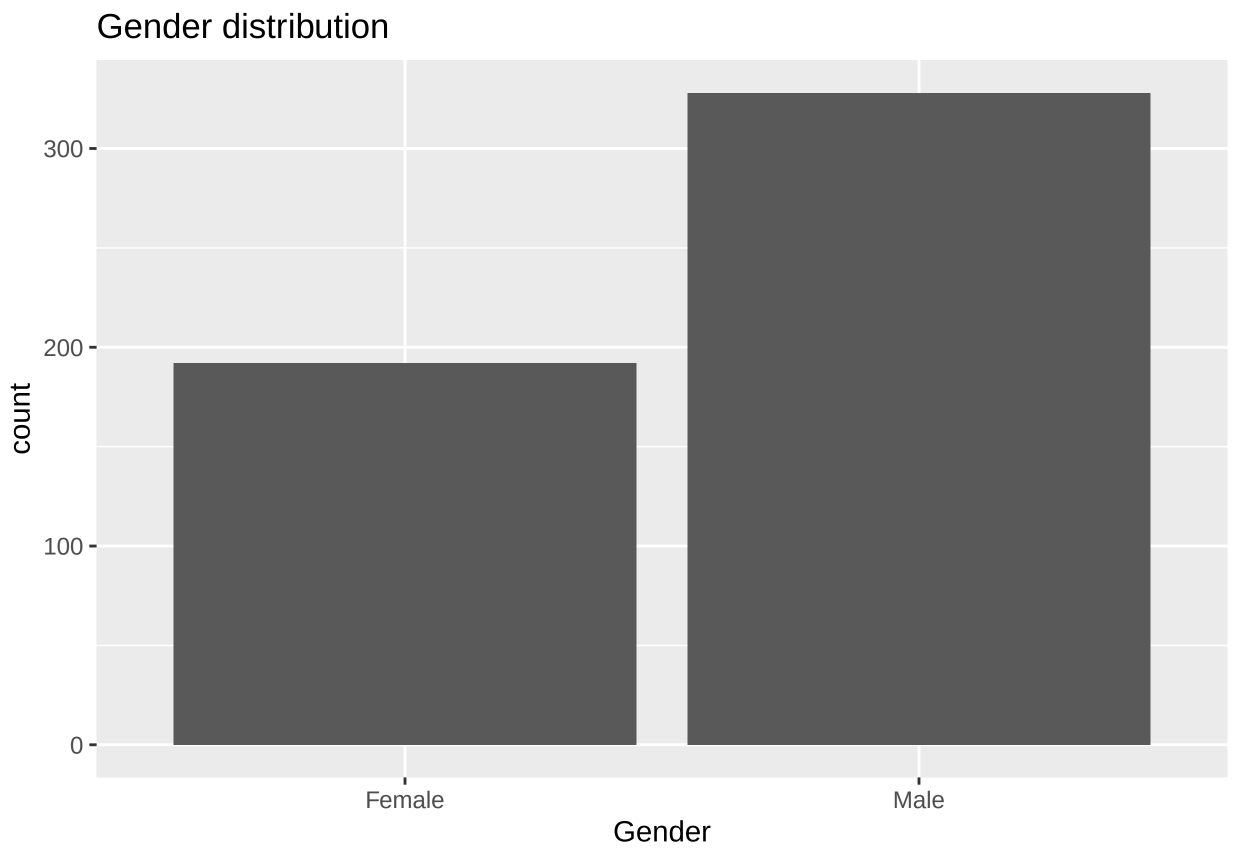


We can see that 320 where diagnosed with diabetes, which represents 61.5384615384615 %.

Age - age of patients is numeric value between 16 and 90, as provided by *summary()* function.

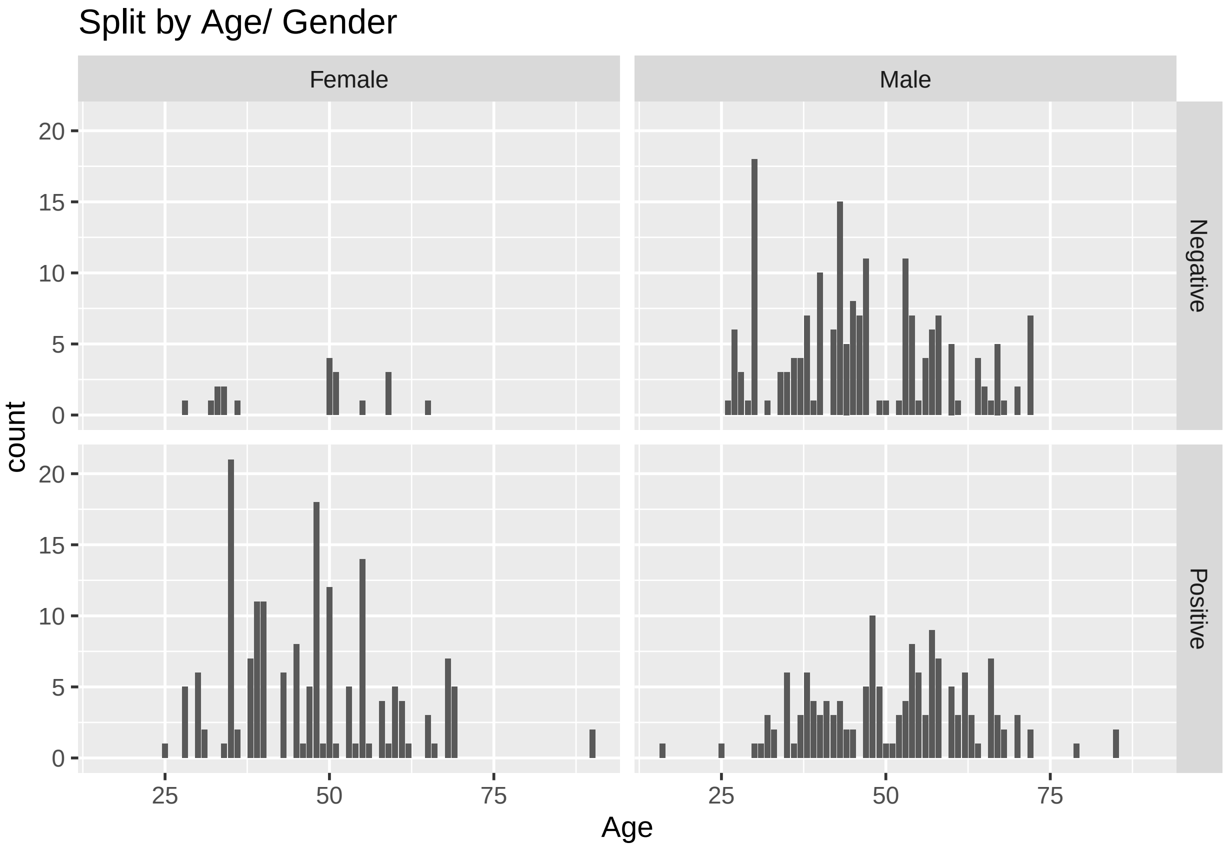


Gender - categorical data with distribution:



Male patients are more represented in this dataset.

Here is demographic distribution of dataset:

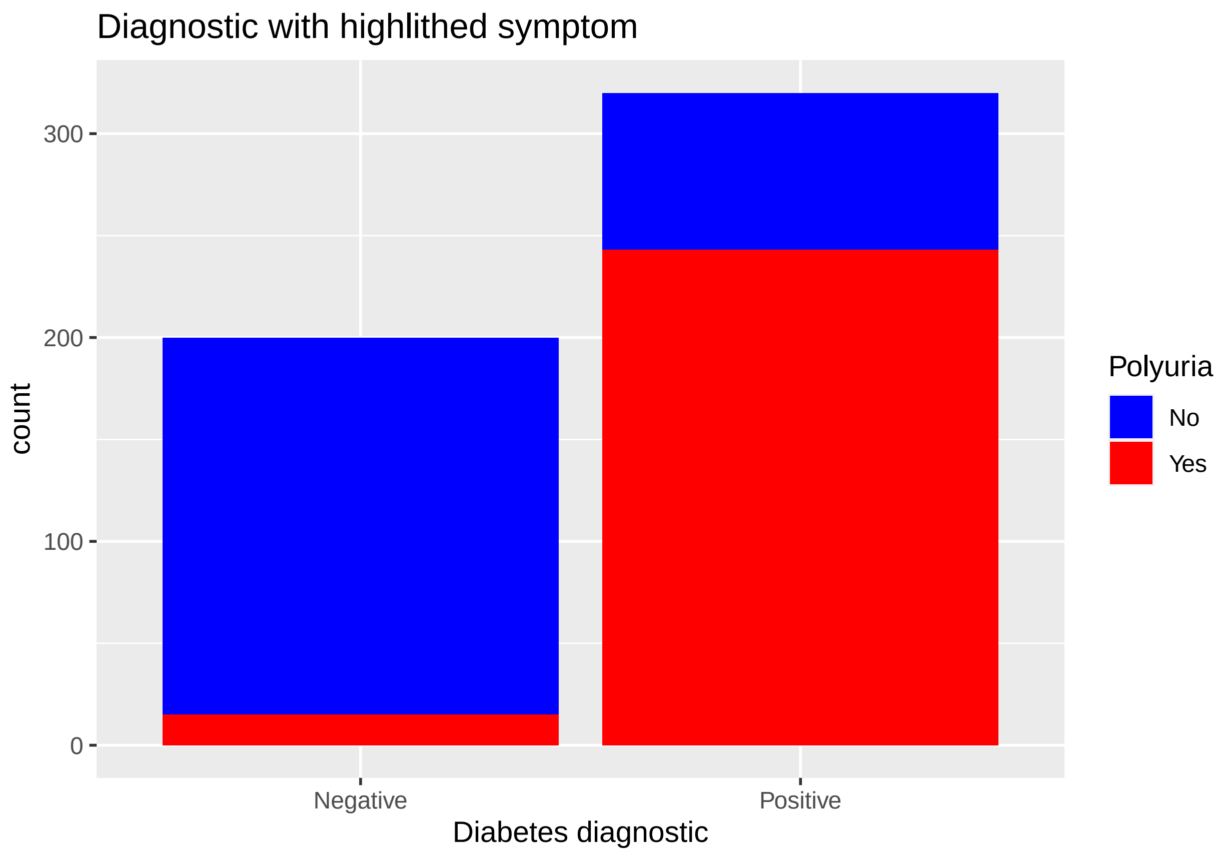


From visual distribution it is difficult to observe any trends related to Age. We can notice that Female patients were more frequently tested positive.

All symptoms are categorical values with values ‘Yes’ or ‘No’. For visualization I’ve used three types of plots: split **Class** by individual symptoms, second split individual symptoms by **Class** and third one split individual symptom with facet by **Gender**.

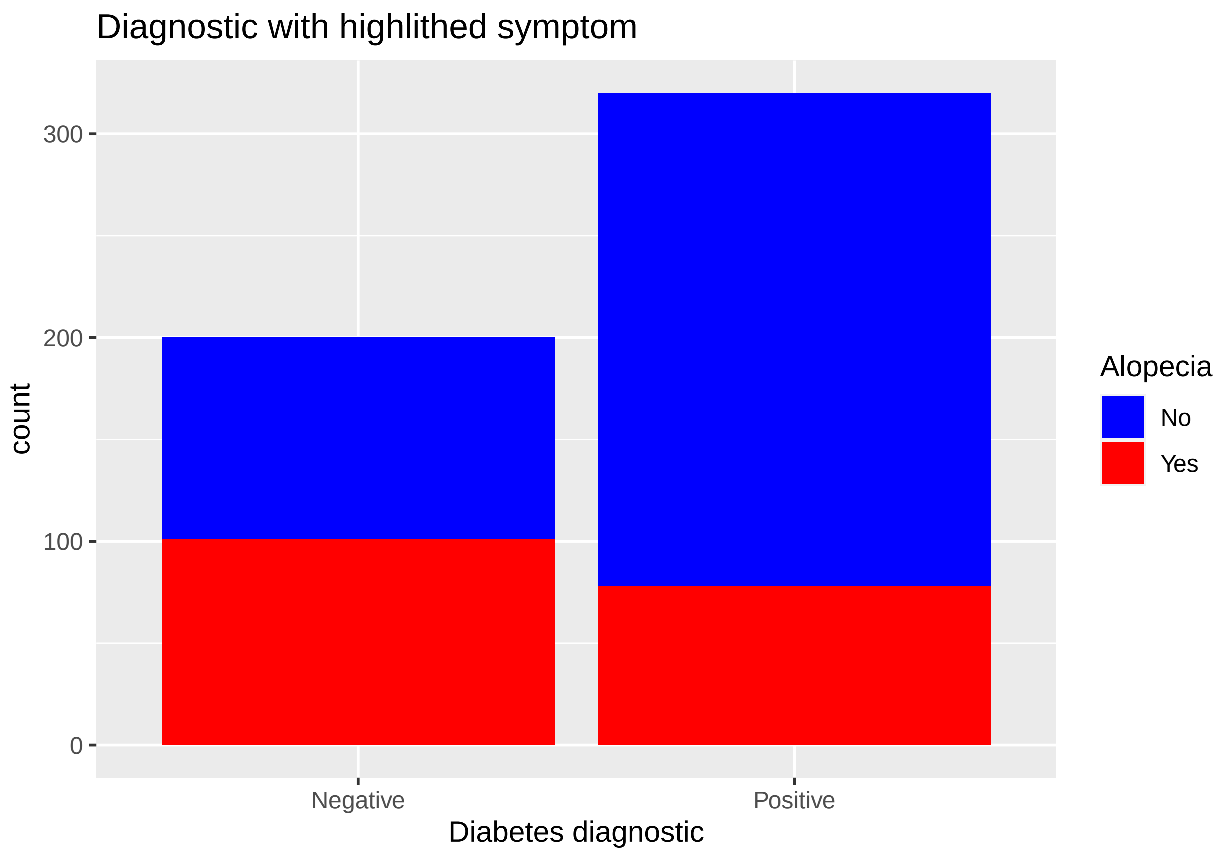
First type of plot help to represent how many people who where diagnosed as Negative or Positive have particular symptom. We can observe that this relationship depends on symptoms.

In case of **Polyuria** it will be:



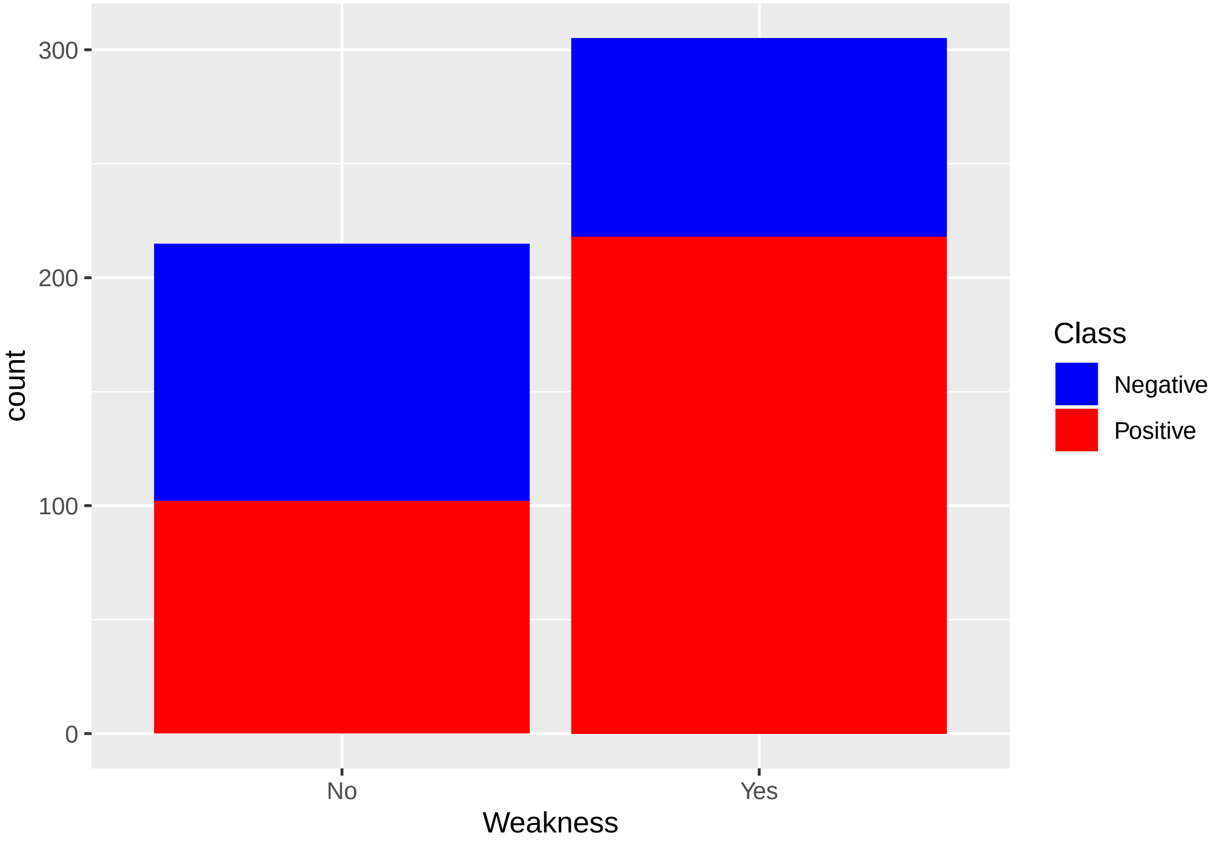
We can observe high number of people with diabetes who present **Polyuria** symptoms and relatively small part of people who does not have diabetes and have **Polyuria** symptom.

Same plot for **Alopecia** will be presented as :

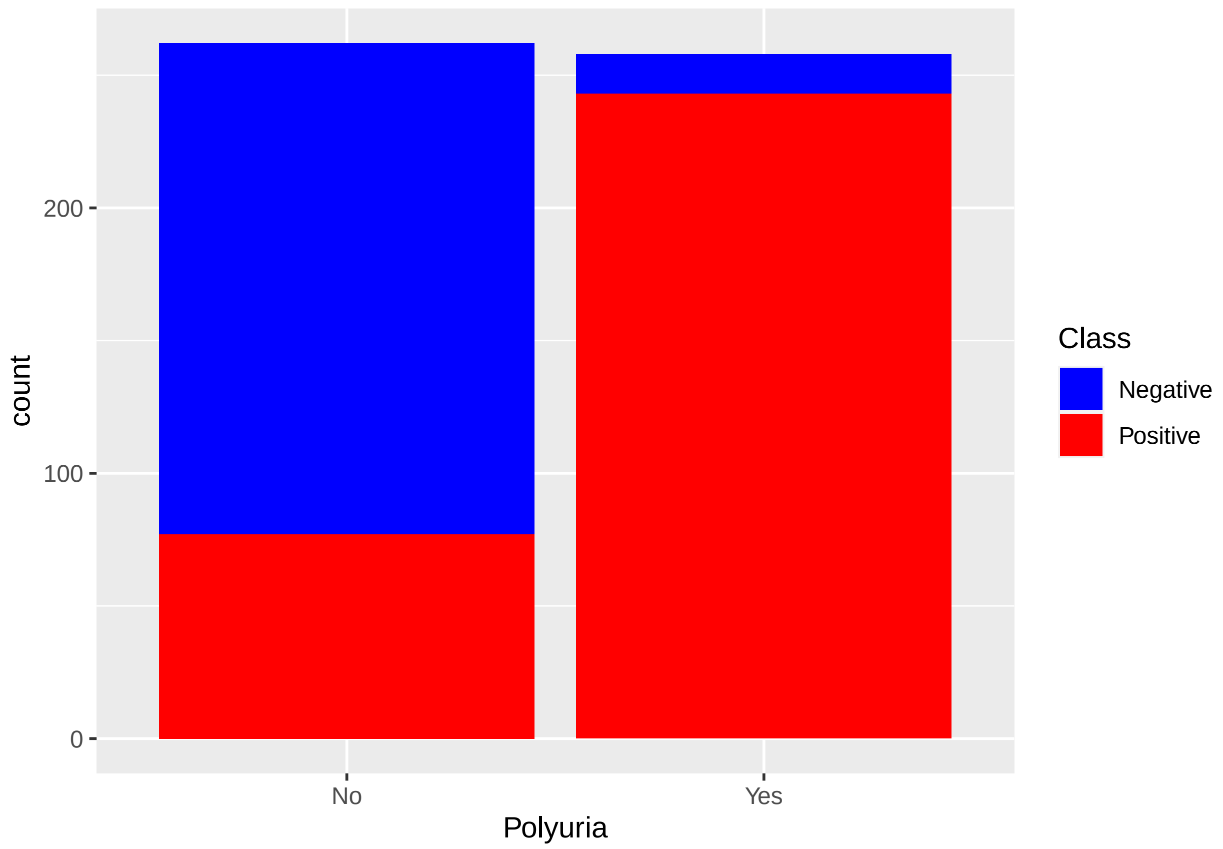


Here we can observe that half of people who does not have diabetes had **Alopecia** symptom and small part of patient with diabetes will have **Alopecia**. This type of bar help to visualize how frequently particular symptom present in case of diabetes.

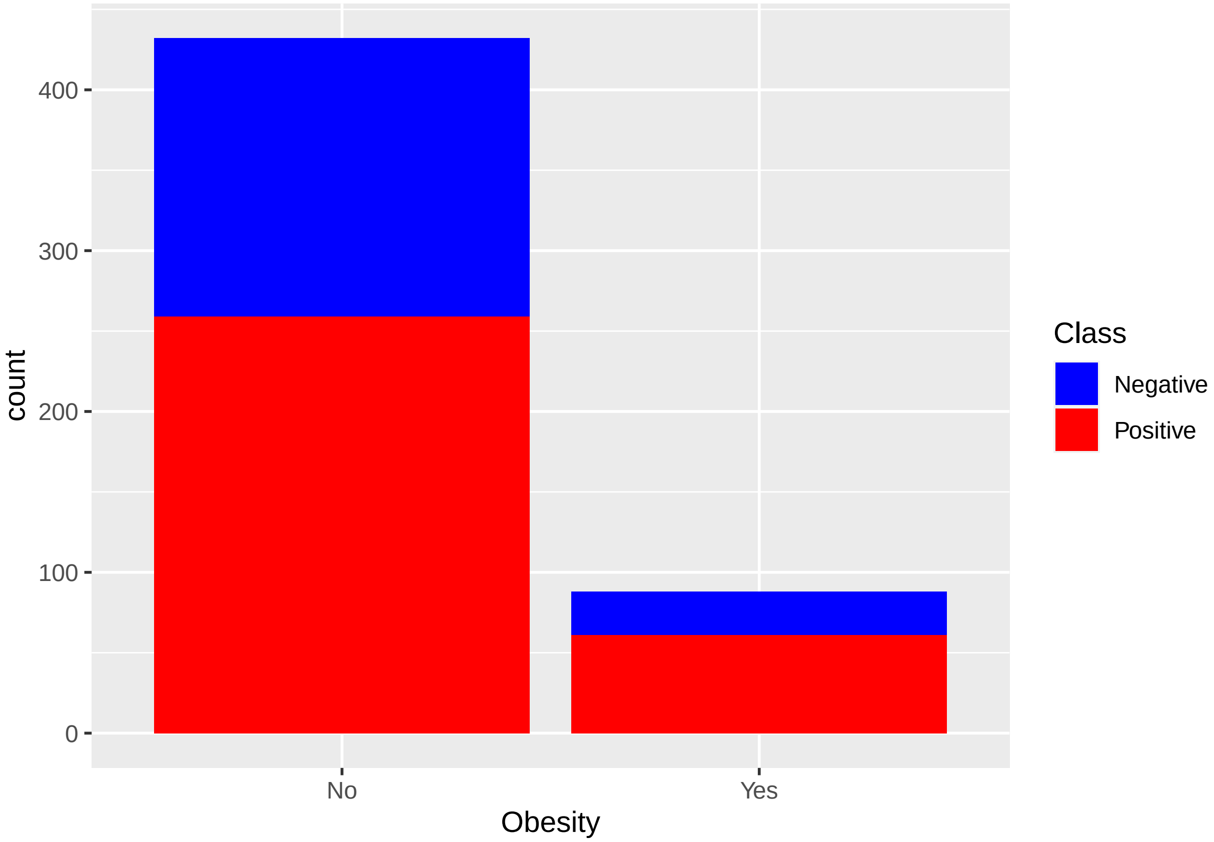
Second type of plot allows to visualize how often symptom is present and how many positive cases reported for people with this sympthom. One of most frequent symptoms is **Weakness** :



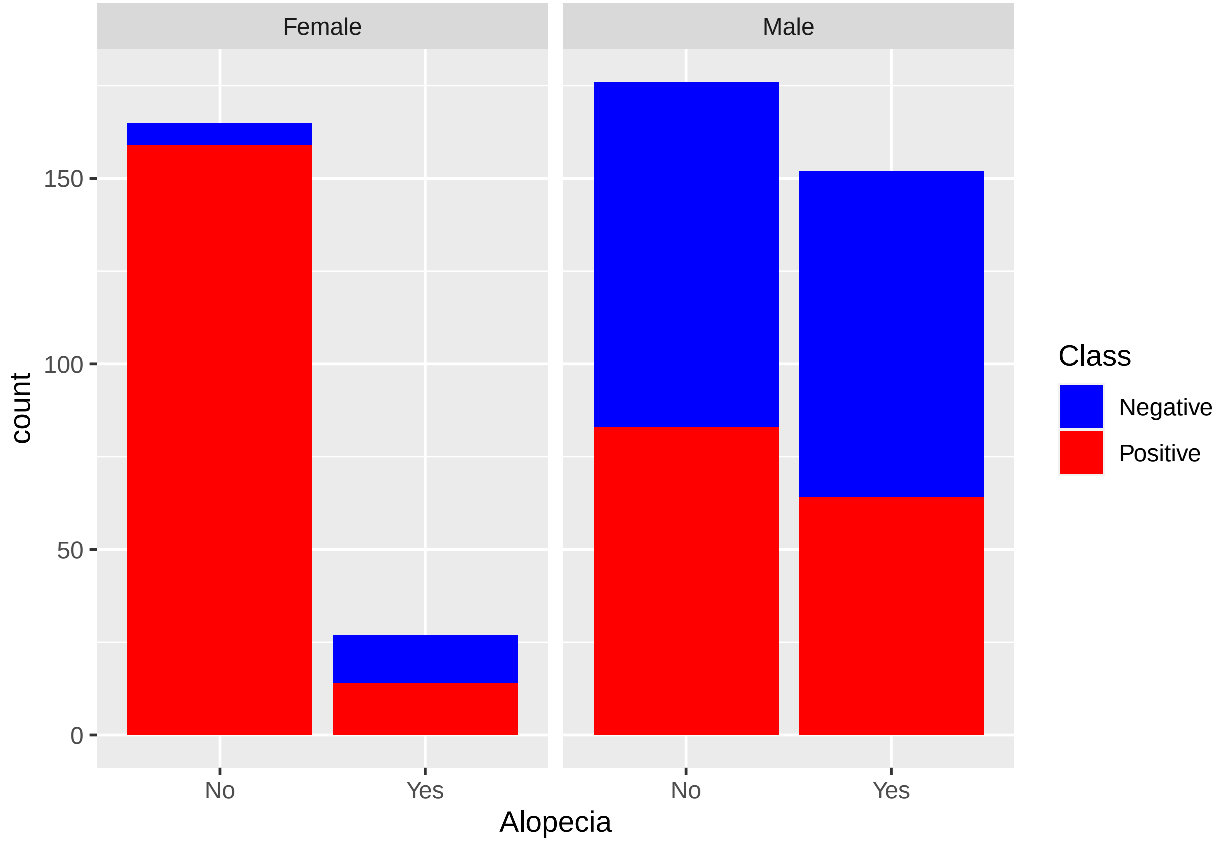
We can see that i.e. nearly 50% of patients has **Polyuria** :



From the other side **Obesity** is one of less frequent symptoms:



Third type of plot, allow to see if **Gender** has influence on symptoms, as i.e. in case of **Alopecia** :



We can clearly see that man have **Alopecia** more frequently than woman and that **Alopecia** is not the most frequent diabetes symptom. This visualization provide some interesting inside on symptoms and diabetes diagnostic but does not define clear picture which symptoms is the most important, it is also not easy to visualize impact of combination of several symptoms.

To estimate how each attribute influence **Class** value and if we have any influence between attribute I've used Anova and Chi Test methods.

As **Age** is numeric value and **Class** is categorical one I've used ANOVA to estimate if **Age** is significant one for **Class** prediction. The results of ANOVA (function *aov()* for **Age** ~ **Class**):

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## Class         1    905   905.1   6.191 0.0132 *
## Residuals   518  75729   146.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

As Pr(>F) is less than 0.05, we can conclude that **Age** has significant influence on **Class**. This conclusion was not clearly identify at data visualization part.

I've also used ANOVA to check if **Age** has influence on others attributes, here is output of ANOVA *summary()*:

```
##                     Df Sum Sq Mean Sq F value   Pr(>F)
## Gender               1    303     303   3.162 0.075971 .
## Polyuria           1   3879    3879  40.487 4.45e-10 ***
## Polydipsia         1    168     168   1.756 0.185679
## Sudden_Weight_Loss   1     12      12   0.127 0.722222
## Weakness           1   2591    2591  27.050 2.89e-07 ***
## Polyphagia         1   5427    5427  56.645 2.42e-13 ***
## Genital_Thrush     1    430     430   4.486 0.034660 *
## Visual_Blurring    1   8015    8015  83.661  < 2e-16 ***
## Itching            1   1397    1397  14.583 0.000151 ***
## Irritability       1    533     533   5.562 0.018737 *
## Delayed_Healing    1    290     290   3.026 0.082535 .
## Partial_Paresis    1    443     443   4.620 0.032071 *
## Muscle_Stiffness   1    389     389   4.062 0.044390 *
## Alopecia           1   4120    4120  43.009 1.35e-10 ***
## Obesity            1    354     354   3.694 0.055159 .
## Residuals          504  48284      96
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Symptoms marked with one or several * are significantly dependent on **Age**. It looks quite logical, as all those symptoms appear with increasing age.

I’ve also verified if **Gender** is significant attributes for **Class**, as both are categorical variable I’ve used Chi test. Here are results:

```
##
##  Pearson’s Chi-squared test
##
## data:  wide_data$Gender and wide_data$Class
## X-squared = 104.94, df = 1, p-value < 2.2e-16
```

```
##                 wide_data$Class
## wide_data$Gender  Negative  Positive
##           Female -6.382374  5.045710
##           Male    4.883104 -3.860432
```

From results we can see that **Gender** is significant predictor for **Class** as p-value is less than 0.05, also from residuals display we can see that individually **Gender** Female is most significant for Positive cases and **Male** is most significant for Negative case.

I’ve performed Chi Test for all symptoms attributes and attribute **Class**. Here is the list of significant feautures (p-value < 0.05):

```
## # A tibble: 12 x 2
##    Sympthom             PVal
##    <chr>               <dbl>
##  1 Gender              0
##  2 PolyUria            0
##  3 Alopecia            0
##  4 Sudden_Weight_Loss 0
##  5 Irritability       0
##  6 Polydipsia         0
##  7 Weakness           0
##  8 Polyphagia         0
##  9 Genital_Thrush     0.0119
## 10 Visual_Blurring    0
## 11 Partial_Paresis    0
## 12 Muscle_Stiffness   0.0052
```

I also compare if symptoms have dependency between them:

```
## # A tibble: 15 x 15
##    col1      A    DH     G    GT    Ir    It    MS     O PartP PolyD PolyP PolyU    VB
##    <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
##  1 A       0     0     0     0     0.3   0     0.4   0.5   0     0     0.2       0   0.7
##  2 DH      0     0     0     0     0     0     0     0.1   0     0     0         0   0
##  3 G       0     0     0     0     0.8   0.2   0     0.9   0     0     0         0   0
##  4 GT      0     0     0     0     0     0     0     0.2   0     0.5   0.1       0   0
##  5 Ir      0.3   0     0.8   0     0     0     0     0     0     0     0         0   0.1
##  6 It      0     0     0.2   0     0     0     0     1     0     0     0         0   0
##  7 MS      0.4   0     0     0     0     0     0     0     0     0     0         0   0
##  8 O       0.5   0.1   0.9   0.2   0     1     0     0     0.8   0     0.5       0   0
##  9 PartP   0     0     0     0     0     0     0     0.8   0     0     0         0   0
## 10 PolyD   0     0     0     0.5   0     0     0     0     0     0     0         0   0
## 11 PolyP   0.2   0     0     0.1   0     0     0     0.5   0     0     0         0   0
## 12 PolyU   0     0     0     0     0     0     0     0     0     0     0         0   0
## 13 SWL     0     0     0     0     0     0.9   0     0     0     0     0         0   0.1
## 14 VB      0.7   0     0     0     0.1   0     0     0     0     0     0         0   0
## 15 W       0     0     0     0.5   0     0     0     0.3   0     0     0         0   0
## # ... with 1 more variable: W <dbl>
```

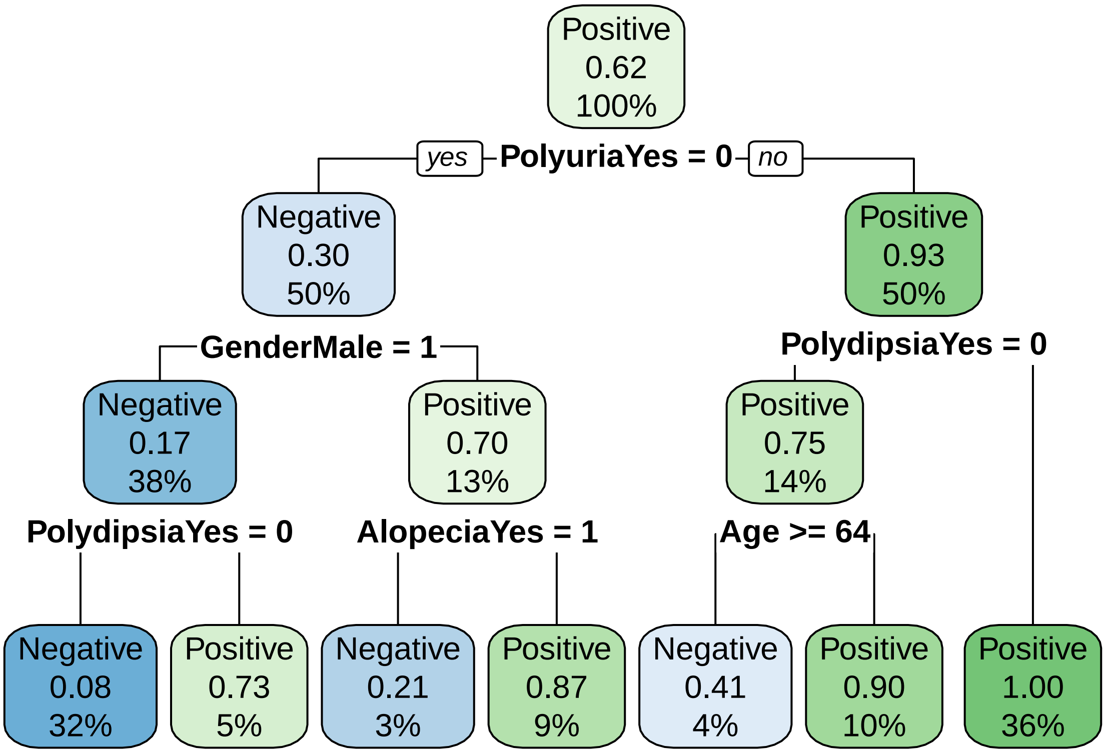
PolyUria is only symptom that is influenced by all others attributes:

```
##  col1     A    DH     G    GT    Ir    It    MS     O PartP PolyD PolyP PolyU   SWL    VB
##   "W" "0.7" "0.1" "0.9" "0.5" "0.8" "1.0" "0.4" "1.0" "0.8" "0.5" "0.5"   "0" "0.9" "0.7"
##     W
## "0.5"
```

2.2 Analysis, model building

Prior to build models I've split original dataset into training and test partition in proportion 80% ~ 20% and defined RMSE function. I've used package carret and tried several methods as ***Knn***, ***Rborist***, ***Random Forest***, ***SVM*** and ***Rpart***. All methods where trained and tuned on training dataset.

After I evaluated each model on test dataset, calculated RMSE and model overall accuracy from confusion matrix. For some models i was able (except___*Knn* **and** *SVM*___) to use *varImp()* function and build the list of most important attributes. This list can be used as guide for patients when they need to consult doctor. *Rpart* method also offer interesting visualization with tree decision which allows predict probability of diabetes based on combination of symptoms.



Here is summary overview:

method	RMSE	accuracy
Rborist	0.0000000	1.0000000
Random Forest	0.0980581	0.9903846
SVM	0.2192645	0.9519231
Rpart	0.2192645	0.9519231
knn	0.3668997	0.8653846

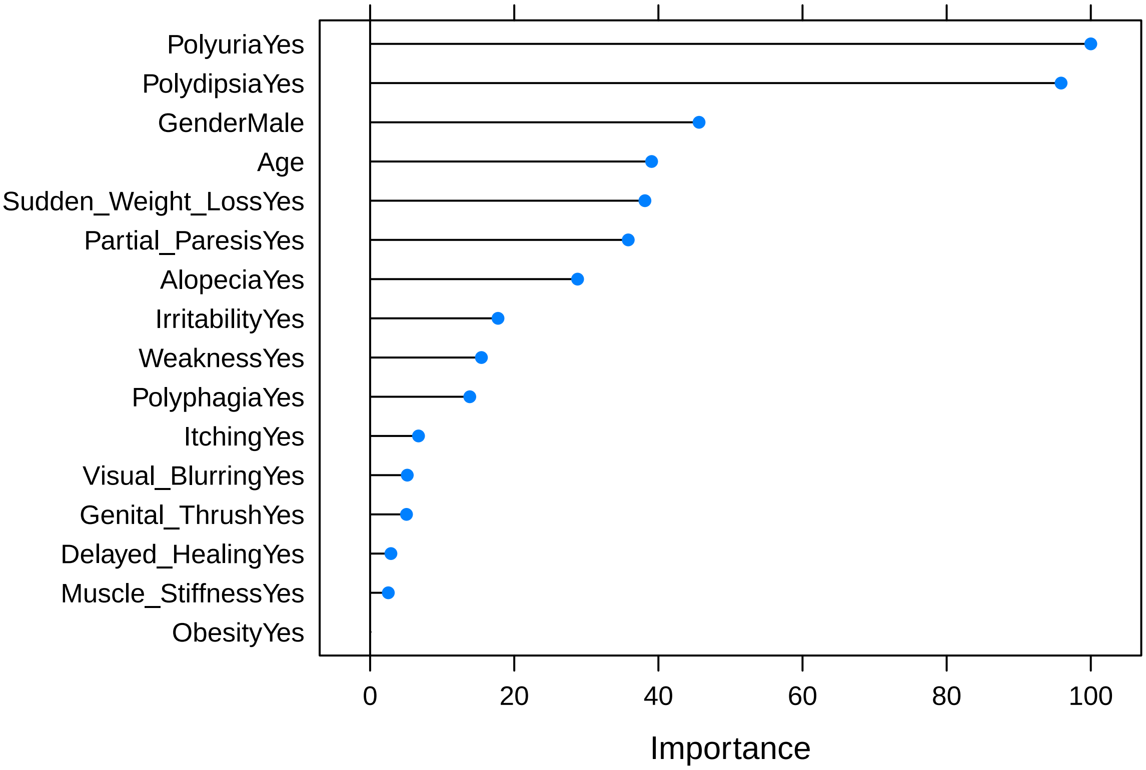
From this table we can see Rborist was one of the slowest to calculate, all others had very similar time execution.

3 Results

The final model was build with Rborist method of caret package. **Class** was predicted against all available attributes.

```
## Random Forest
##
## 416 samples
##  16 predictor
##   2 classes: 'Negative', 'Positive'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 416, 416, 416, 416, 416, 416, ...
## Resampling results across tuning parameters:
##
##   minNode  Accuracy   Kappa
##    3       0.9538272  0.9028904
##   50       0.8932646  0.7721688
##
## Tuning parameter 'predFixed' was held constant at a value of 2
## Accuracy was used to select the optimal model using the largest value.
## The final values used for the model were predFixed = 2 and minNode = 3.
```

The most important attributes:


PolyuriaYes
PolydipsiaYes
GenderMale
Age
Sudden_Weight_LossYes
Partial_ParesisYes
AlopeciaYes
IrritabilityYes
WeaknessYes
PolyphagiaYes
ItchingYes
Visual_BlurringYes
Genital_ThrushYes
Delayed_HealingYes
Muscle_StiffnessYes
ObesityYes
0
20
40
60
80
100
Importance

4 Conclusion

Dataset represents a static situation; the time dimension is missing. It will be interesting to see the evolution of symptoms with time. Also we do not know which type of diabetes is involved, or it is mix. Dataset was not trained for all ages, i.e. it was not trained for children younger than 16. It will be nice to build similar decision tree for final model as for rpart model.

References

Islam, MM Faniqul, et al. ‘Likelihood prediction of diabetes at early stage using data mining techniques.’ Computer Vision and Machine Intelligence in Medical Image Analysis. Springer, Singapore, 2020. 113-125.