Assignment 2

Made by

Khadija Hesham

E-mail

khesh072@uottawa.com

Supervised by

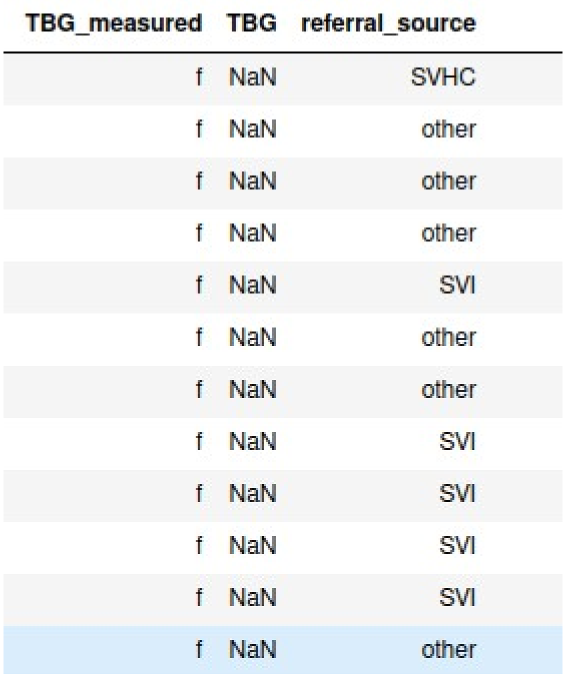
Dr, Bisi Runsewe

Part A):

Data Preparation:

Missing values handling:

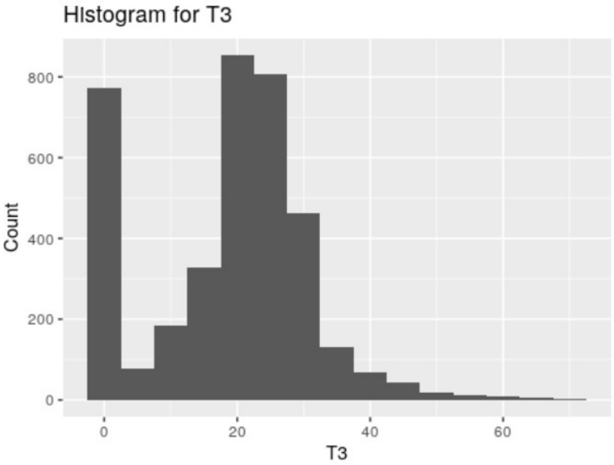
We have noticed that missing data has been represented by the “?” sign, hence we are converting this sign to NaN then to drop these data, or to fill them.

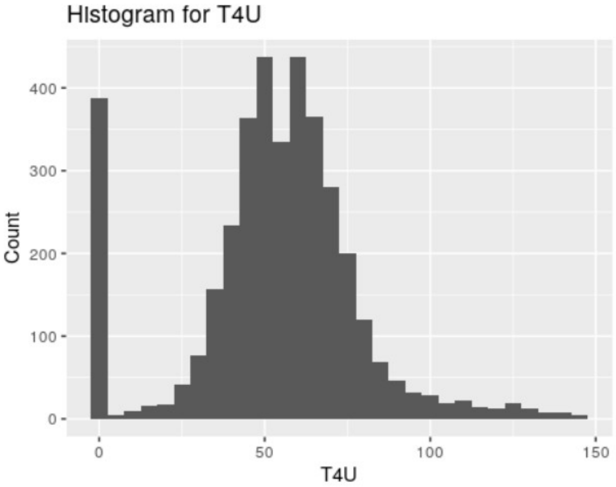


TBG_measured	TBG	referral_source
f	NaN	SVHC
f	NaN	other
f	NaN	other
f	NaN	other
f	NaN	SVI
f	NaN	other
f	NaN	other
f	NaN	SVI
f	NaN	SVI
f	NaN	SVI
f	NaN	SVI
f	NaN	other

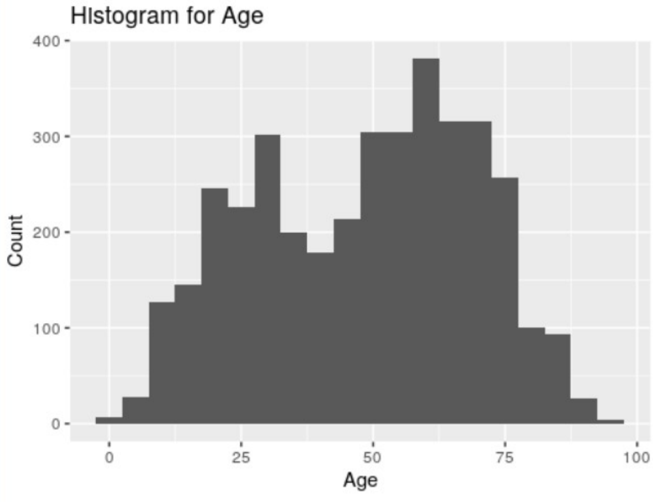
We have noticed that the “TBG” column has all nan data, so we have to get rid of it.

We have noticed that age, T4U and T3 all have approx. a bell shaped distribution.
T3 and T4U both of them have a right skewed distribution, hence we will use median to fill missing data.

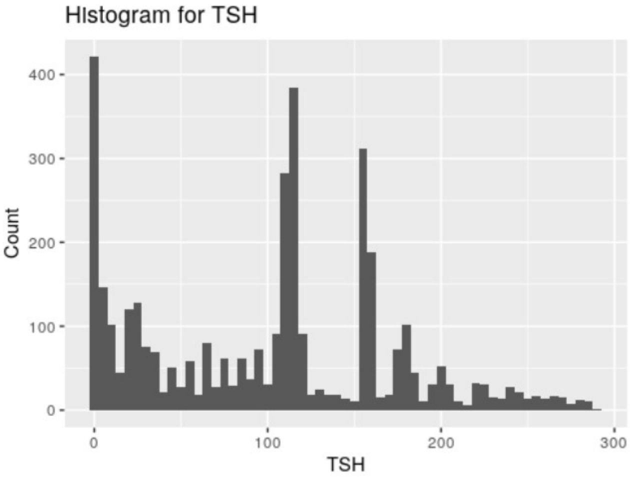


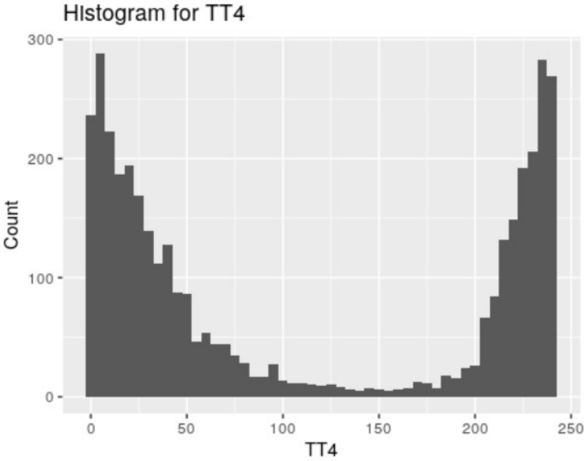


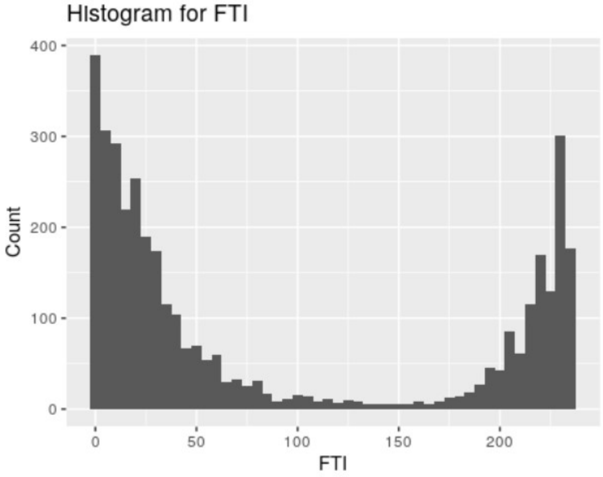
In addition to using mean in age column to fill missing data.



The rest of features do not follow a distribution.
For the rest of numerical features , we will drop missing data as all do not follow a distribution.

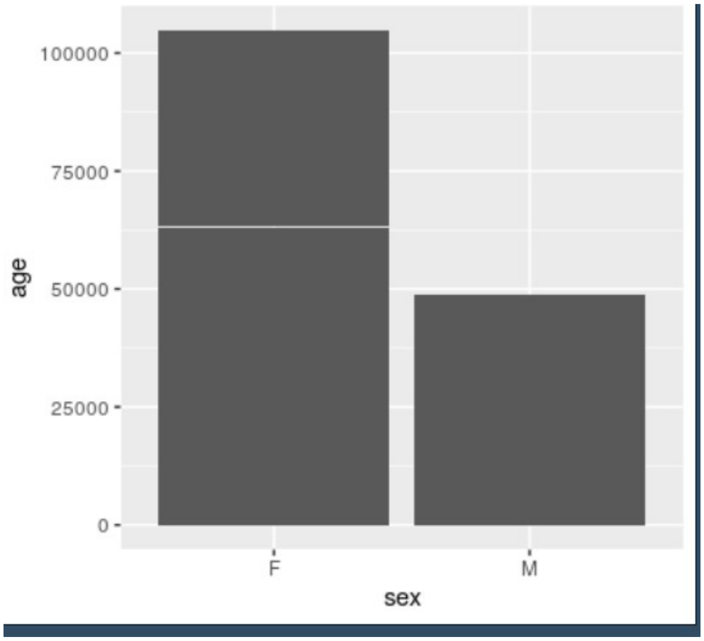






For categorical data, we will drop all missing values.

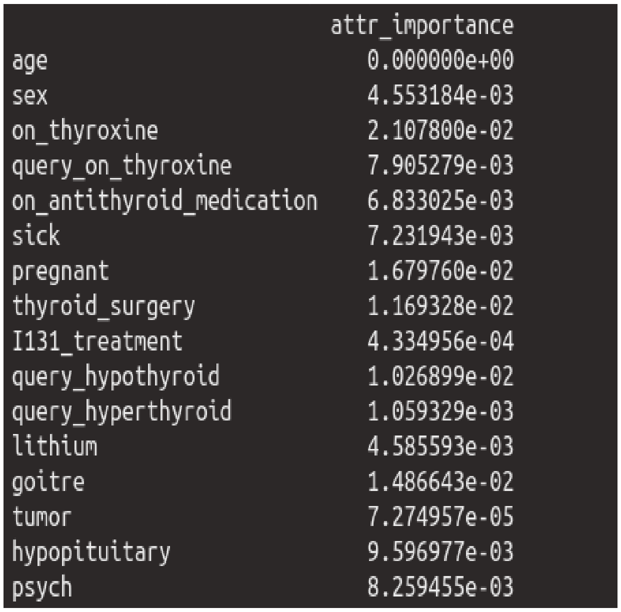
We have noticed that the females have much age in comparison with males.



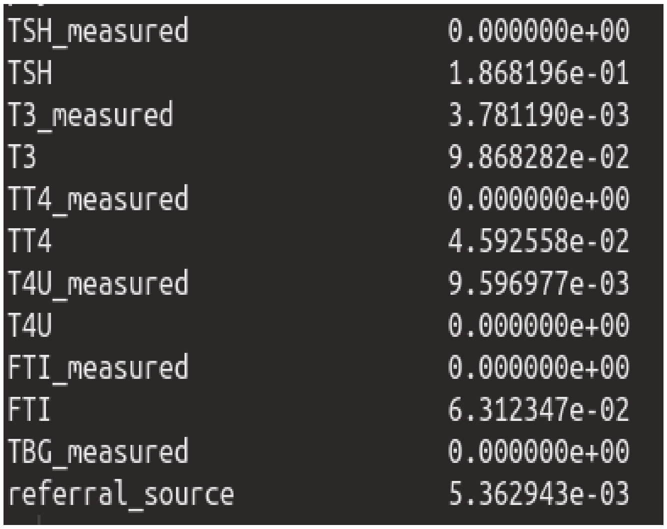
Attribute Selection:

- It enables the machine learning algorithm to train faster.
- It reduces the complexity of a model and makes it easier to interpret.
- It improves the accuracy of a model if the right subset is chosen.
- It reduces overfitting.

We will rank features by Feature Ranking Algorithm which has a package called “Fselector” in r. We have performed feature importance

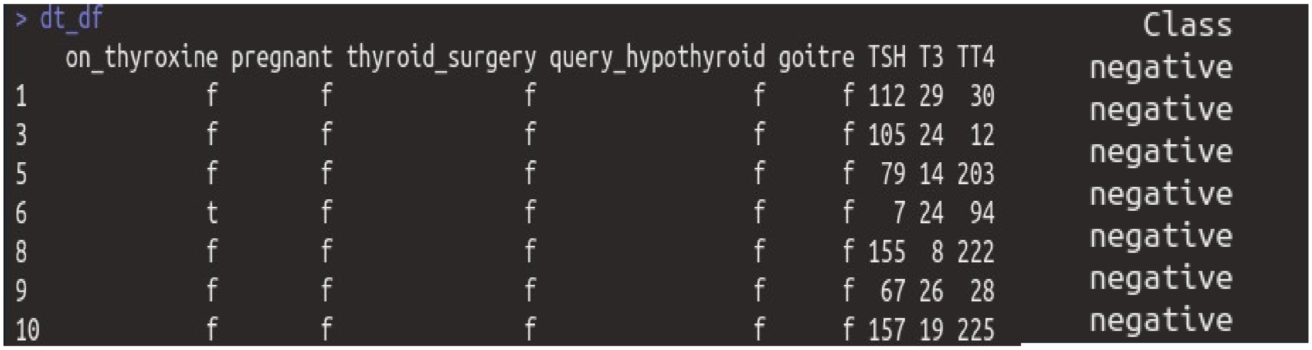


```
                          attr_importance
age                          0.000000e+00
sex                          4.553184e-03
on_thyroxine                 2.107800e-02
query_on_thyroxine           7.905279e-03
on_antithyroid_medication    6.833025e-03
sick                         7.231943e-03
pregnant                     1.679760e-02
thyroid_surgery              1.169328e-02
I131_treatment               4.334956e-04
query_hypothyroid            1.026899e-02
query_hyperthyroid           1.059329e-03
lithium                      4.585593e-03
goitre                       1.486643e-02
tumor                        7.274957e-05
hypopituitary                9.596977e-03
psych                        8.259455e-03
```



```
TSH_measured                 0.000000e+00
TSH                          1.868196e-01
T3_measured                  3.781190e-03
T3                           9.868282e-02
TT4_measured                 0.000000e+00
TT4                          4.592558e-02
T4U_measured                 9.596977e-03
T4U                          0.000000e+00
FTI_measured                 0.000000e+00
FTI                          6.312347e-02
TBG_measured                 0.000000e+00
referral_source              5.362943e-03
```

We will use features having importance above or equal 0.001.
Reduced data set:

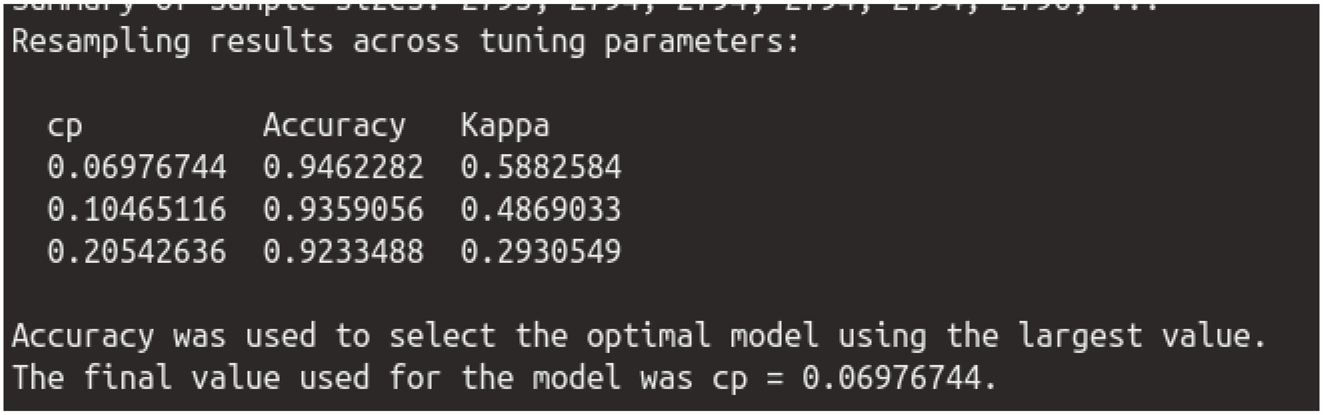


```
> dt_df                                                                         Class
   on_thyroxine pregnant thyroid_surgery query_hypothyroid goitre TSH T3 TT4    negative
1             f        f               f                 f      f 112 29  30    negative
3             f        f               f                 f      f 105 24  12    negative
5             f        f               f                 f      f  79 14 203    negative
6             t        f               f                 f      f   7 24  94    negative
8             f        f               f                 f      f 155  8 222    negative
9             f        f               f                 f      f  67 26  28    negative
10            f        f               f                 f      f 157 19 225    negative
```

Modeling:

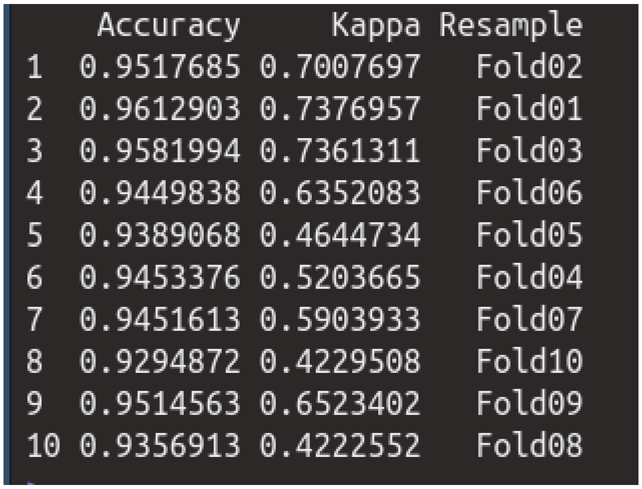
We are modeling with DT algorithm using 10 k fold.

Resampling results:



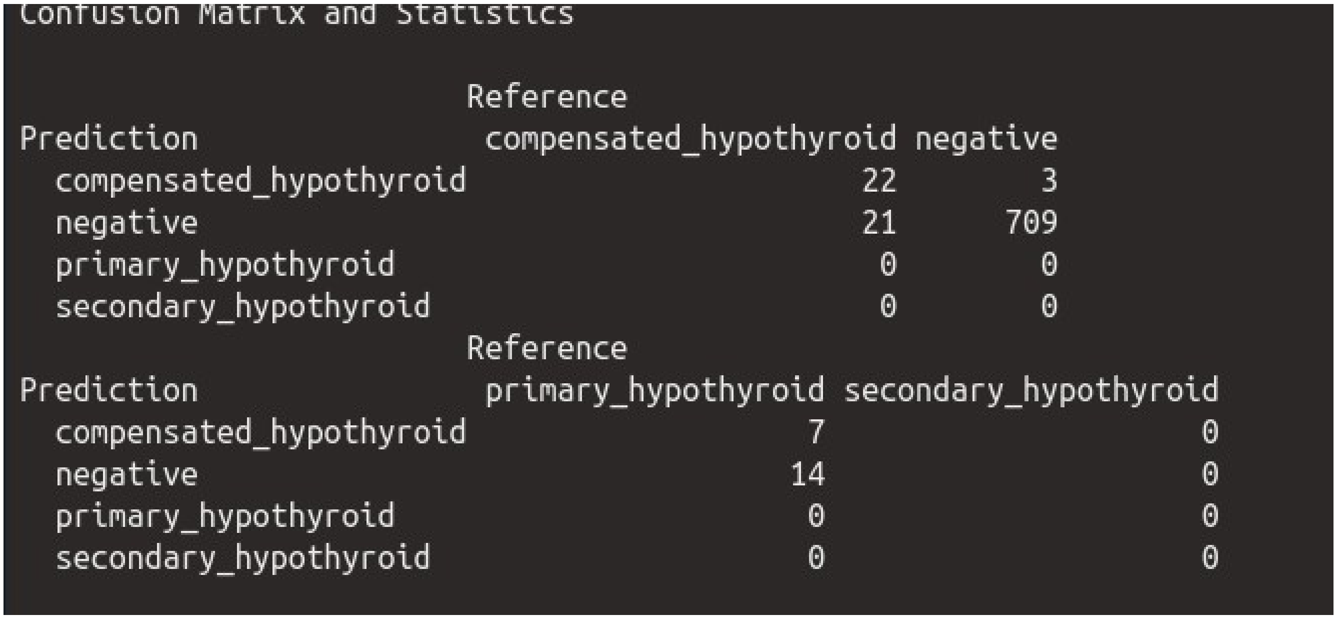
```
Summary of sample sizes: 2793, 2794, 2794, 2794, 2794, 2796, ...
Resampling results across tuning parameters:

  cp          Accuracy   Kappa
  0.06976744  0.9462282  0.5882584
  0.10465116  0.9359056  0.4869033
  0.20542636  0.9233488  0.2930549

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.06976744.
```

Metrics per fold:



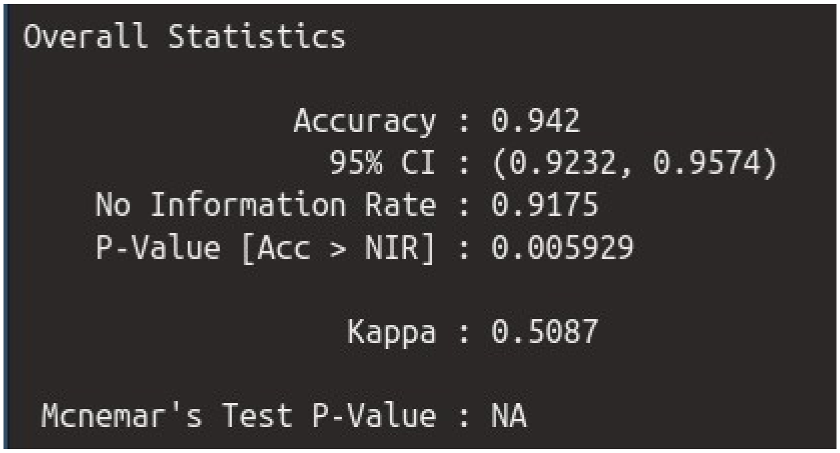
```
   Accuracy     Kappa Resample
1  0.9517685 0.7007697   Fold02
2  0.9612903 0.7376957   Fold01
3  0.9581994 0.7361311   Fold03
4  0.9449838 0.6352083   Fold06
5  0.9389068 0.4644734   Fold05
6  0.9453376 0.5203665   Fold04
7  0.9451613 0.5903933   Fold07
8  0.9294872 0.4229508   Fold10
9  0.9514563 0.6523402   Fold09
10 0.9356913 0.4222552   Fold08
```

Confusion Matrix and Statistics:

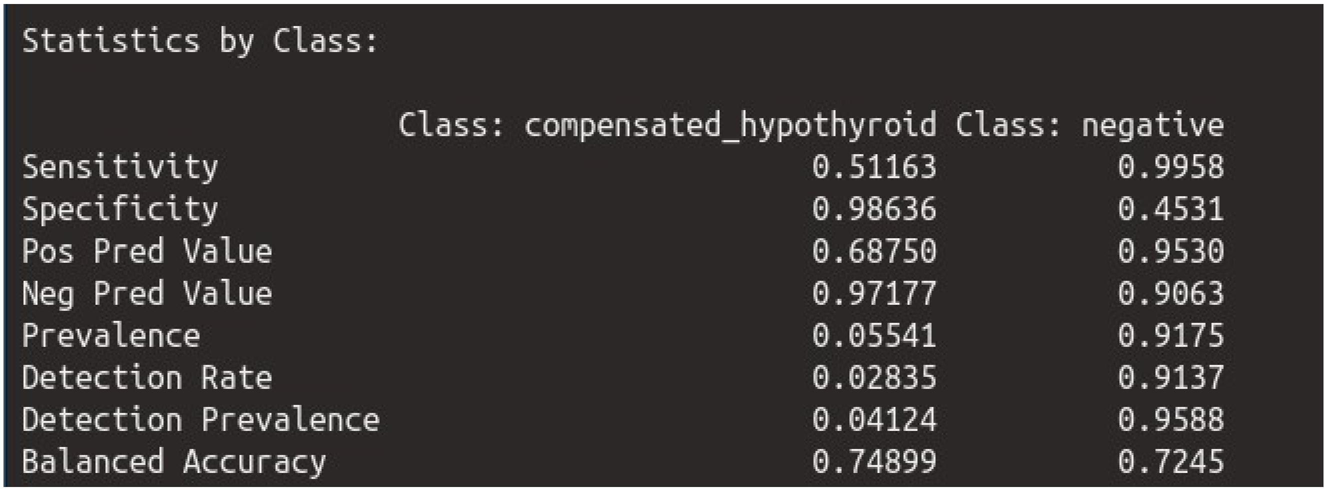


```
Confusion Matrix and Statistics


                         Reference
Prediction                compensated_hypothyroid negative
  compensated_hypothyroid                      22        3
  negative                                     21      709
  primary_hypothyroid                           0        0
  secondary_hypothyroid                         0        0
                         Reference
Prediction                primary_hypothyroid secondary_hypothyroid
  compensated_hypothyroid                   7                     0
  negative                                 14                     0
  primary_hypothyroid                       0                     0
  secondary_hypothyroid                     0                     0
```

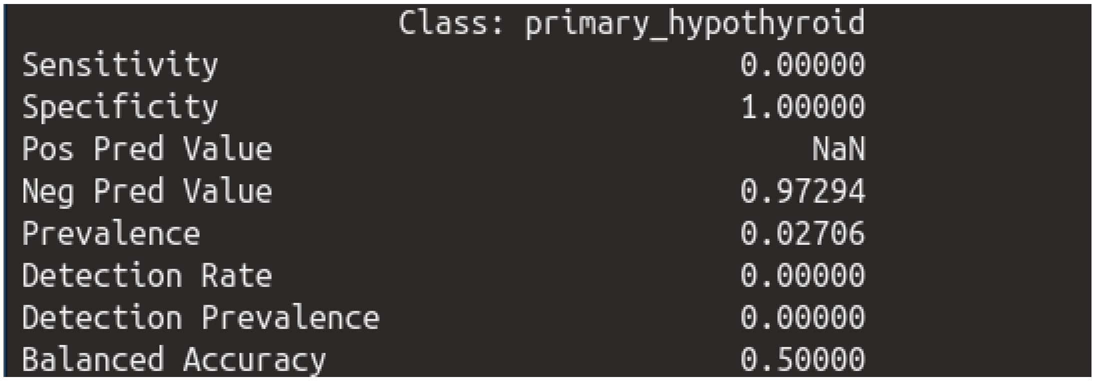
Accuracy:



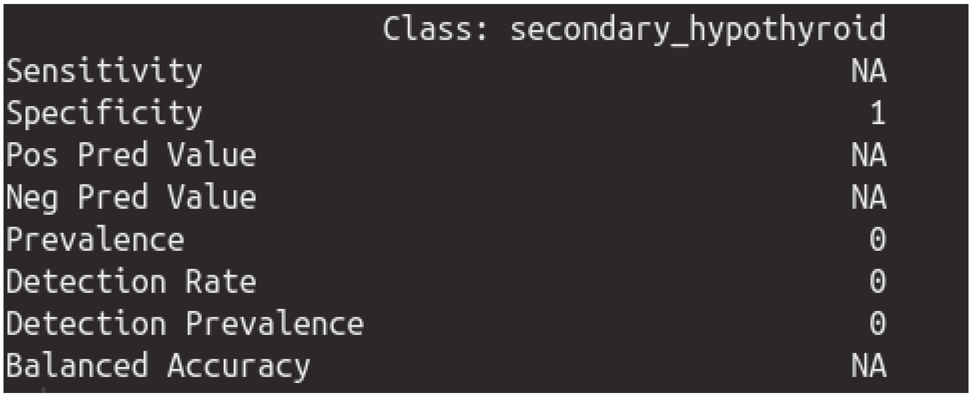
```
Overall Statistics

               Accuracy : 0.942
                 95% CI : (0.9232, 0.9574)
    No Information Rate : 0.9175
    P-Value [Acc > NIR] : 0.005929

                  Kappa : 0.5087

 Mcnemar's Test P-Value : NA
```



```
Statistics by Class:

                     Class: compensated_hypothyroid Class: negative
Sensitivity                                 0.51163          0.9958
Specificity                                 0.98636          0.4531
Pos Pred Value                              0.68750          0.9530
Neg Pred Value                              0.97177          0.9063
Prevalence                                  0.05541          0.9175
Detection Rate                              0.02835          0.9137
Detection Prevalence                        0.04124          0.9588
Balanced Accuracy                           0.74899          0.7245
```

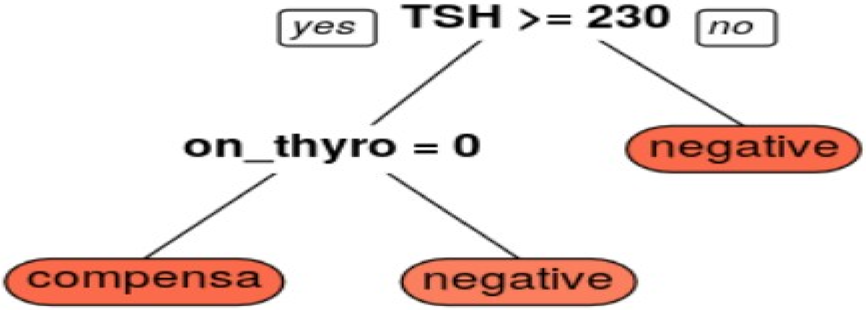


```
                     Class: primary_hypothyroid
Sensitivity                             0.00000
Specificity                             1.00000
Pos Pred Value                              NaN
Neg Pred Value                          0.97294
Prevalence                              0.02706
Detection Rate                          0.00000
Detection Prevalence                    0.00000
Balanced Accuracy                       0.50000
```



```
                     Class: secondary_hypothyroid
Sensitivity                                    NA
Specificity                                     1
Pos Pred Value                                 NA
Neg Pred Value                                 NA
Prevalence                                      0
Detection Rate                                  0
Detection Prevalence                            0
Balanced Accuracy                              NA
```

Tree Plot:

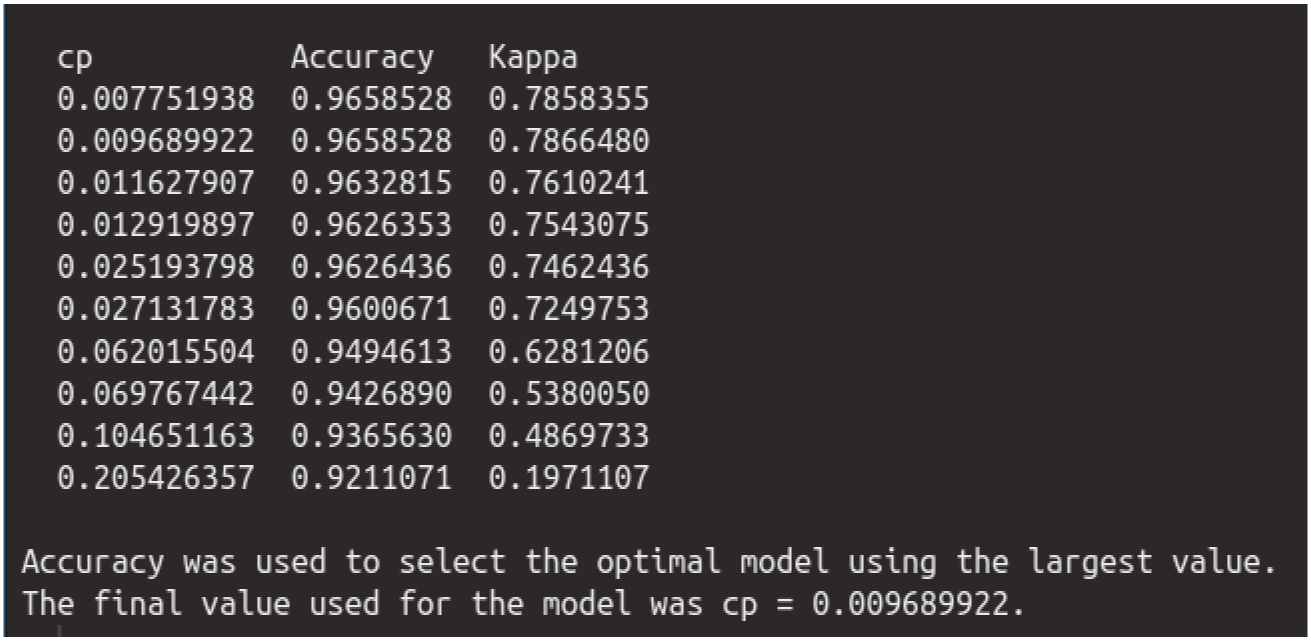


We can get some rules as :
If the TSH value is not above or equal 230 , the class is negative.
If the TSH value is above or equal 230 and the on_thyro value is not equal to 0 , the class is negative.
If the TSH value is above or equal 230 and the on_thyro value is equal to 0 , the class is compensa.

Model Evaluation:

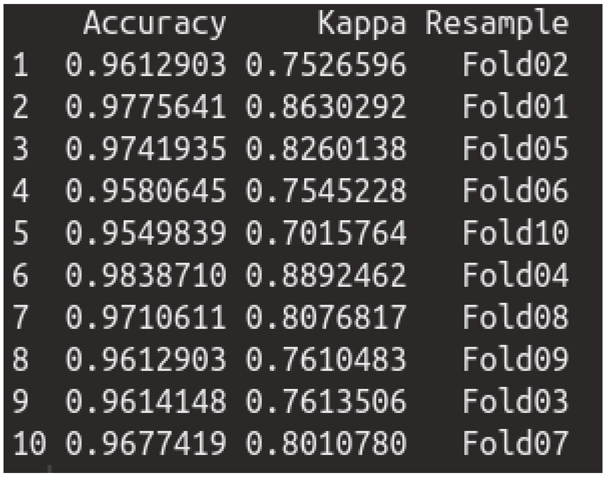
We will try to improve the decision tree by pruning by adding a max depth of 10:

Metrics after cross validation:

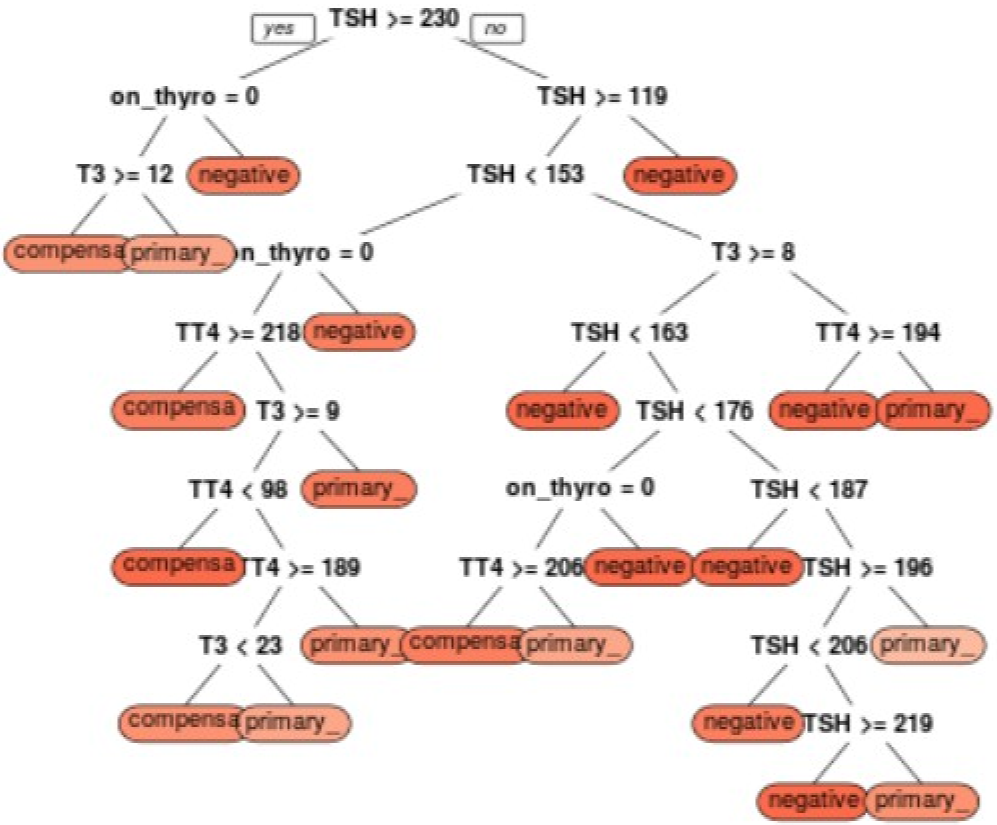


```
  cp           Accuracy   Kappa    
  0.007751938  0.9658528  0.7858355
  0.009689922  0.9658528  0.7866480
  0.011627907  0.9632815  0.7610241
  0.012919897  0.9626353  0.7543075
  0.025193798  0.9626436  0.7462436
  0.027131783  0.9600671  0.7249753
  0.062015504  0.9494613  0.6281206
  0.069767442  0.9426890  0.5380050
  0.104651163  0.9365630  0.4869733
  0.205426357  0.9211071  0.1971107

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was cp = 0.009689922.
```

Prediction for each fold:



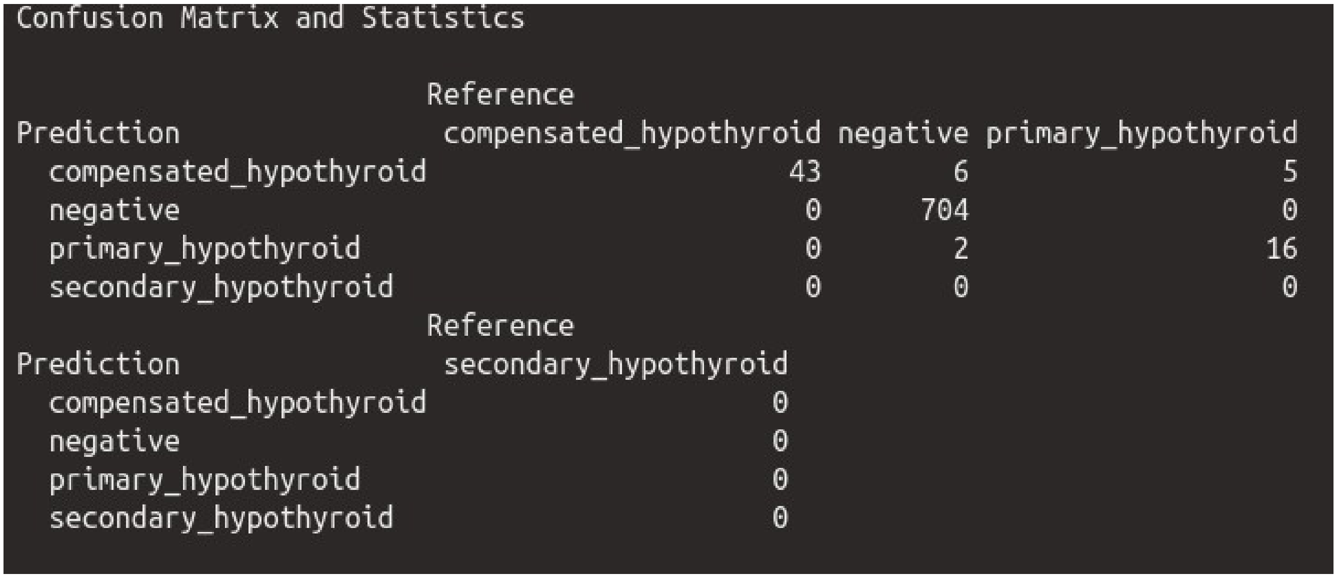
```
    Accuracy     Kappa Resample
1  0.9612903 0.7526596   Fold02
2  0.9775641 0.8630292   Fold01
3  0.9741935 0.8260138   Fold05
4  0.9580645 0.7545228   Fold06
5  0.9549839 0.7015764   Fold10
6  0.9838710 0.8892462   Fold04
7  0.9710611 0.8076817   Fold08
8  0.9612903 0.7610483   Fold09
9  0.9614148 0.7613506   Fold03
10 0.9677419 0.8010780   Fold07
```



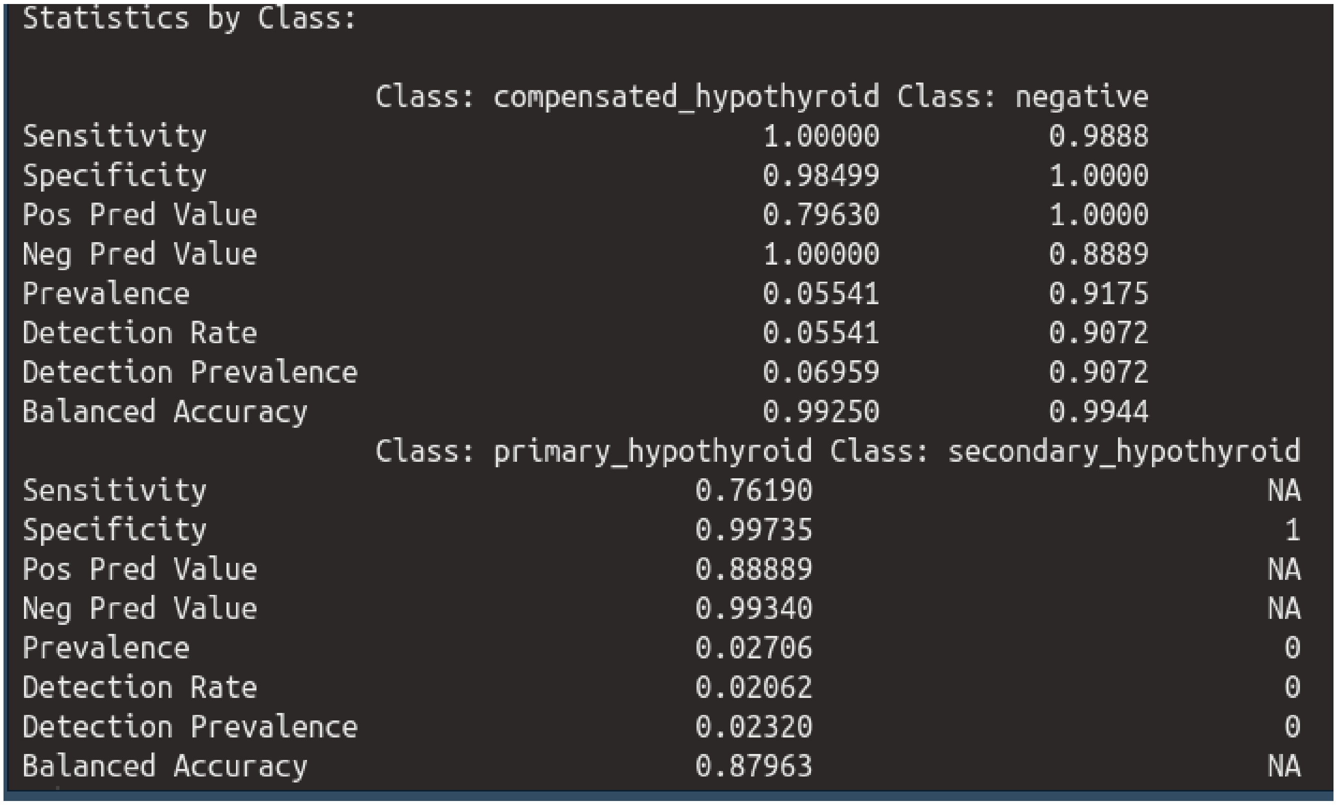
We can get some rules from the best model above:

If the TSH value was lower than 230 but greater than or equal 119, the class is negative.
if the TSH value was above or equal 230 and if the on_thyro value was zero, the class is negative.
if the TSH value was above or equal 230 , if the on_thyro value was zero, and if the T3 value was above 12, the class is compensa, while if the T3 value was below 12, the class is primary.

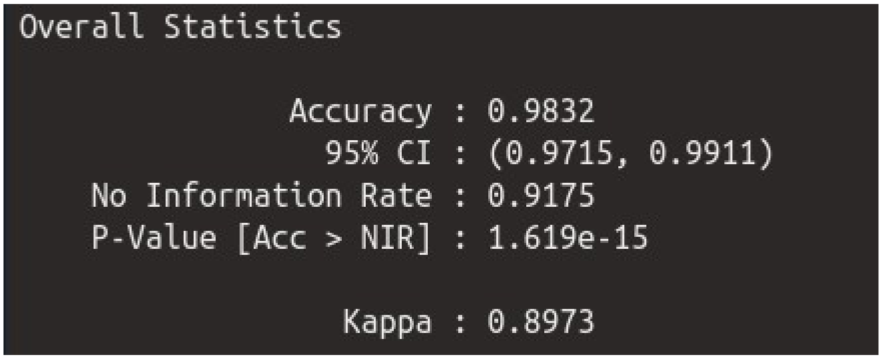
Model Performance:



```
Confusion Matrix and Statistics

                         Reference
Prediction                compensated_hypothyroid negative primary_hypothyroid
  compensated_hypothyroid                      43        6                   5
  negative                                      0      704                   0
  primary_hypothyroid                           0        2                  16
  secondary_hypothyroid                         0        0                   0
                         Reference
Prediction                secondary_hypothyroid
  compensated_hypothyroid                     0
  negative                                    0
  primary_hypothyroid                         0
  secondary_hypothyroid                       0
```



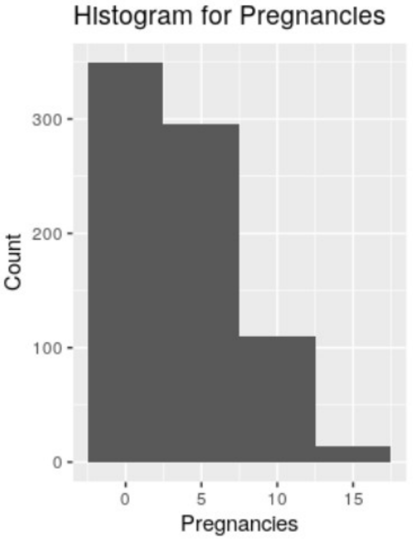
```
Statistics by Class:

                     Class: compensated_hypothyroid Class: negative
Sensitivity                                 1.00000          0.9888
Specificity                                 0.98499          1.0000
Pos Pred Value                              0.79630          1.0000
Neg Pred Value                              1.00000          0.8889
Prevalence                                  0.05541          0.9175
Detection Rate                              0.05541          0.9072
Detection Prevalence                        0.06959          0.9072
Balanced Accuracy                           0.99250          0.9944
                     Class: primary_hypothyroid Class: secondary_hypothyroid
Sensitivity                             0.76190                           NA
Specificity                             0.99735                            1
Pos Pred Value                          0.88889                           NA
Neg Pred Value                          0.99340                           NA
Prevalence                              0.02706                            0
Detection Rate                          0.02062                            0
Detection Prevalence                    0.02320                            0
Balanced Accuracy                       0.87963                           NA
```

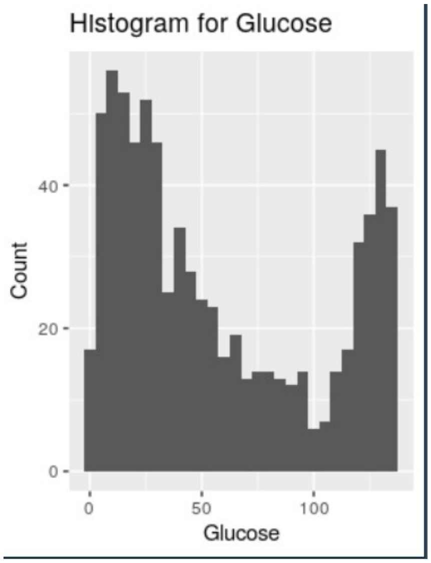


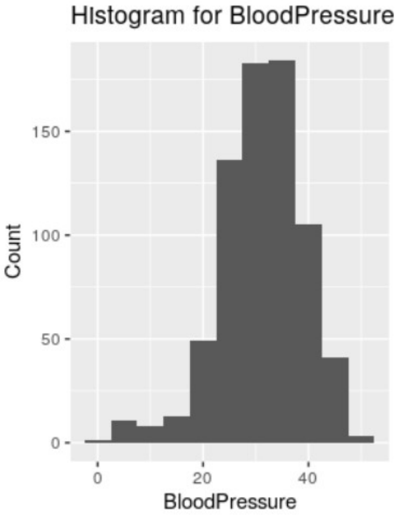
```
Overall Statistics

               Accuracy : 0.9832
                 95% CI : (0.9715, 0.9911)
    No Information Rate : 0.9175
    P-Value [Acc > NIR] : 1.619e-15

                  Kappa : 0.8973
```

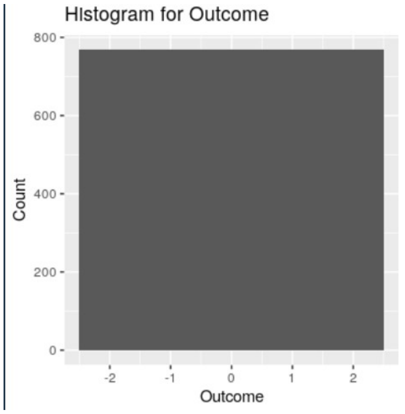
It’s shown that tree pruning has resulted in a significant improvement in metrics and accuracy has been increased from 94.2% up to 98.32%

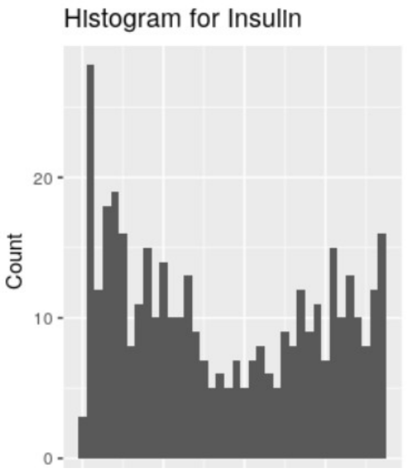
Part B):

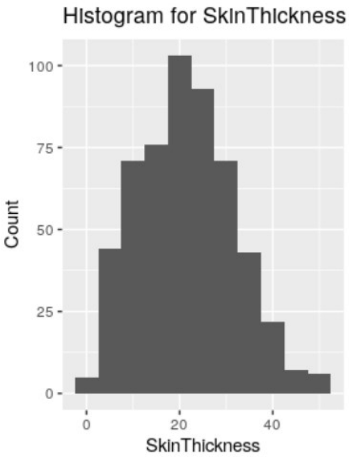


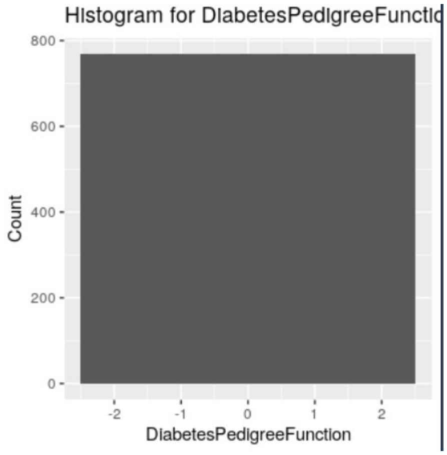


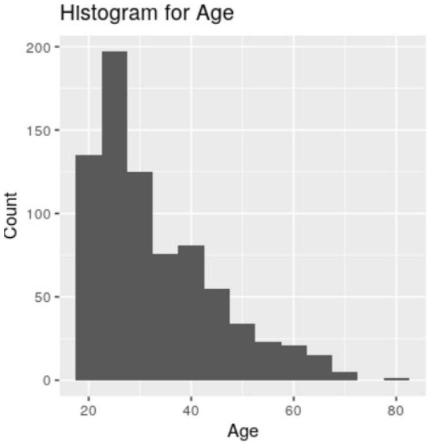


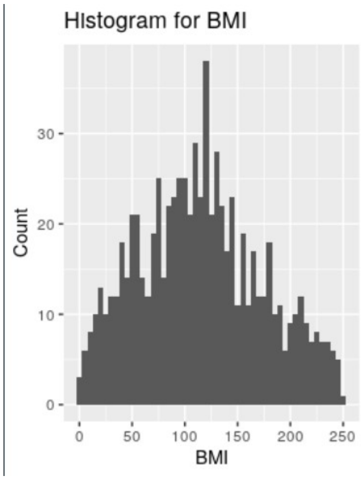








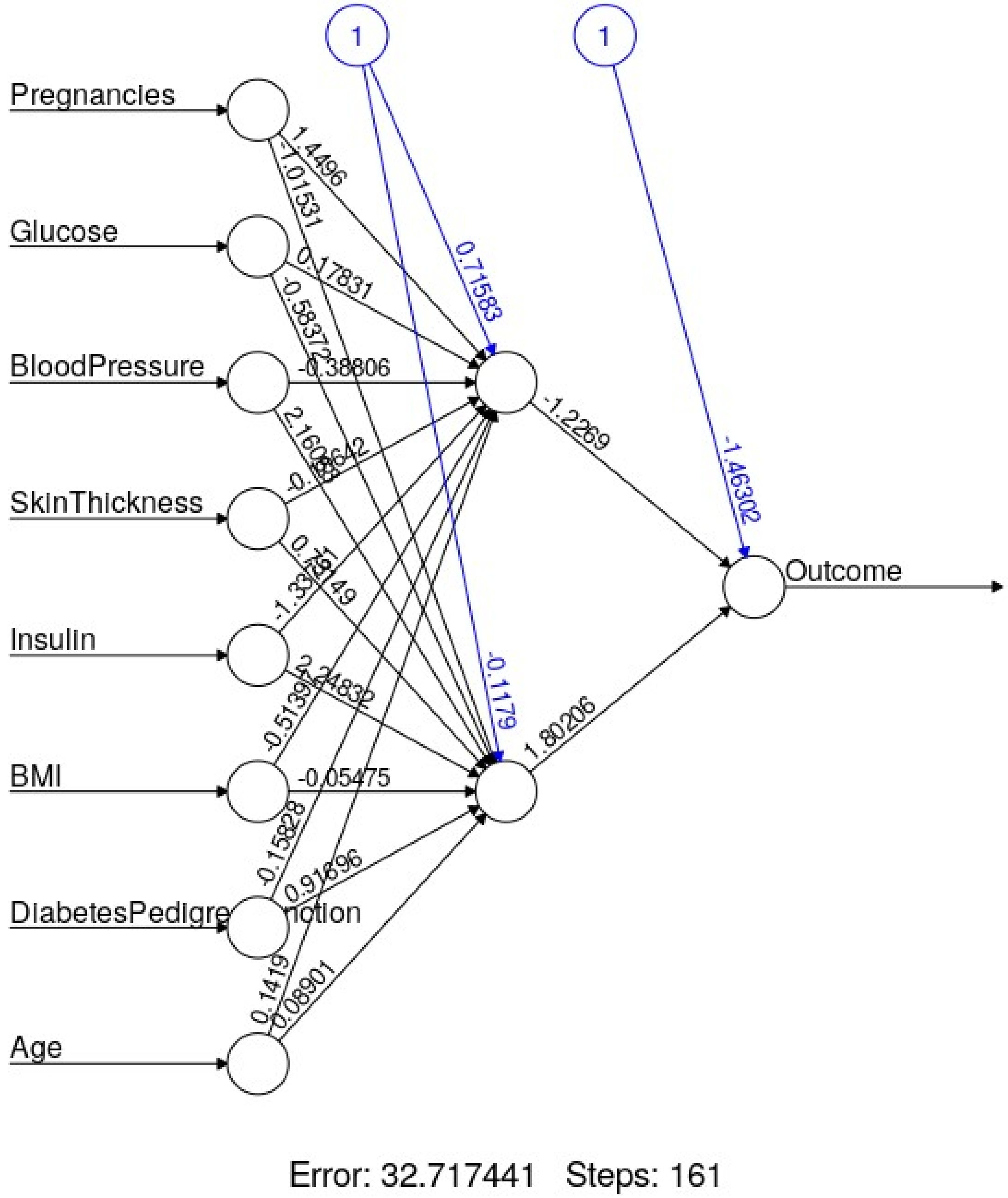




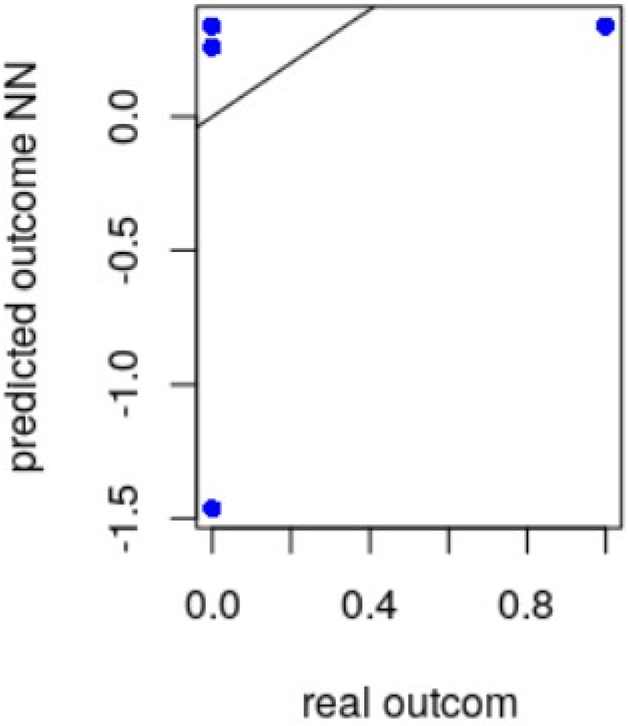
We will fill nan values in BIM, blood pressure and skin thickness by the mean.

We will fill nan values in age and pregnancy by median, other wise we will drop all missing values in the rest of features.
We have splitted a 25% of data for testing and made min max transformation to the data, then we have modeled with NN algorithm with 2 hidden neurons.

NN plot:

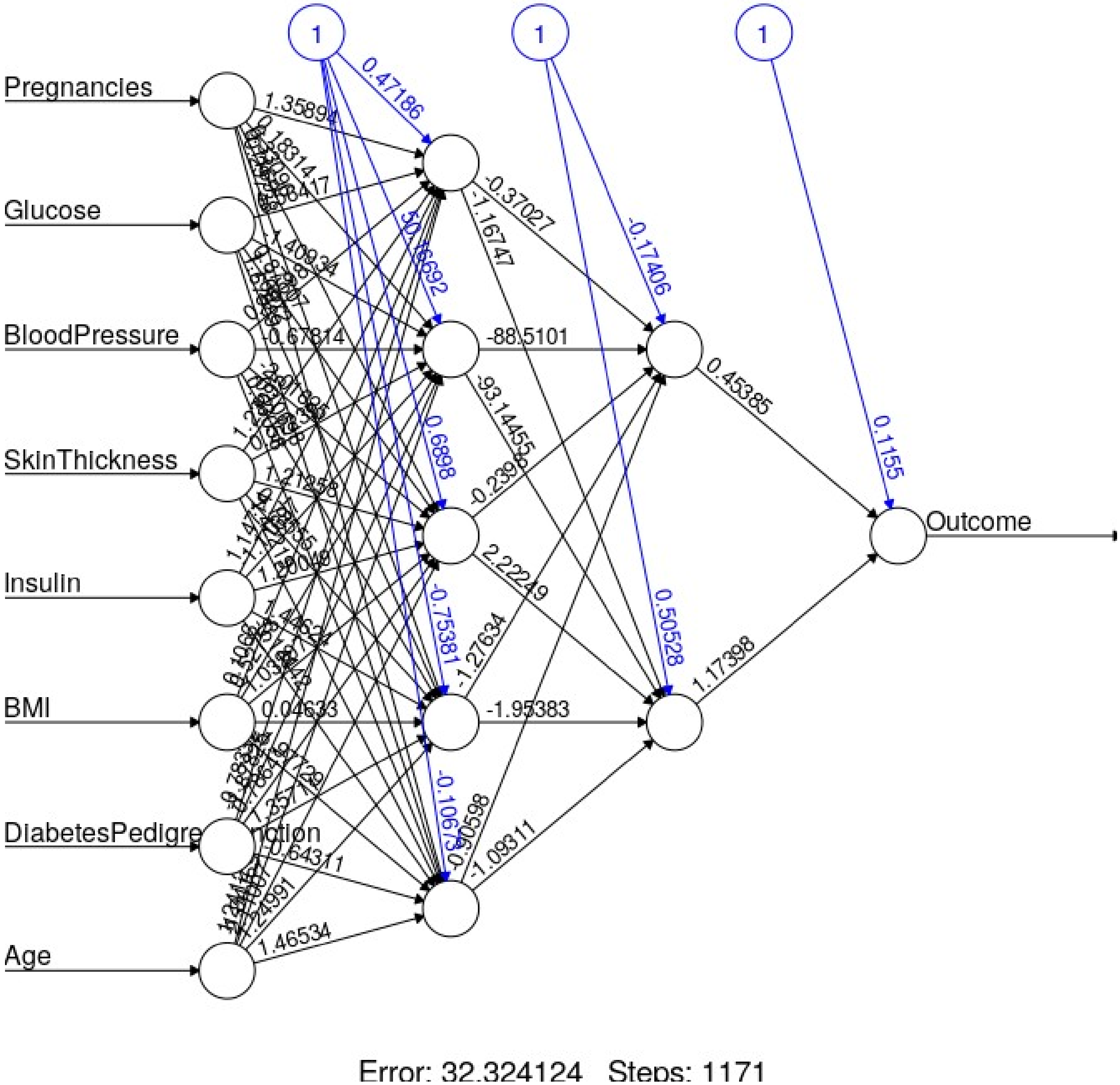


Evaluation:



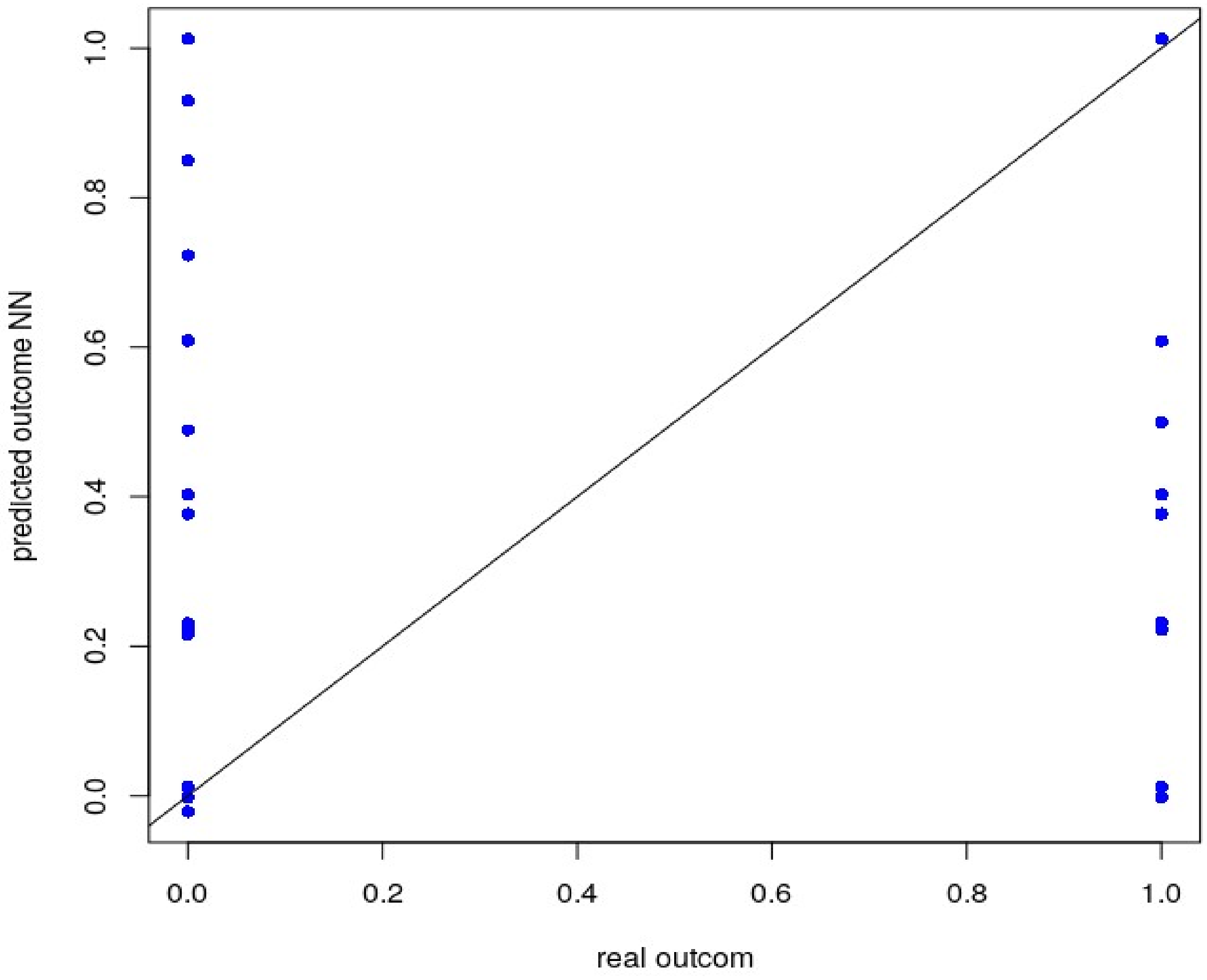
The Root Mean Square Error (RMSE) is 0.4921767

Retrain with 2 layers and 5 neurons:


1
1
1
Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunction
Age
Outcome
0.47186
1.35894
0.183141
-1.40934
-0.67814
1.21858
1.38049
0.04633
0.64311
1.46534
-0.37027
-1.16747
-0.17406
-88.5101
-93.14455
0.6898
-0.2398
2.22249
-0.75381
-1.27634
-1.95383
-0.10673
-1.09311
0.50528
0.45385
1.17398
0.1155
Error: 32.324124 Steps: 1171

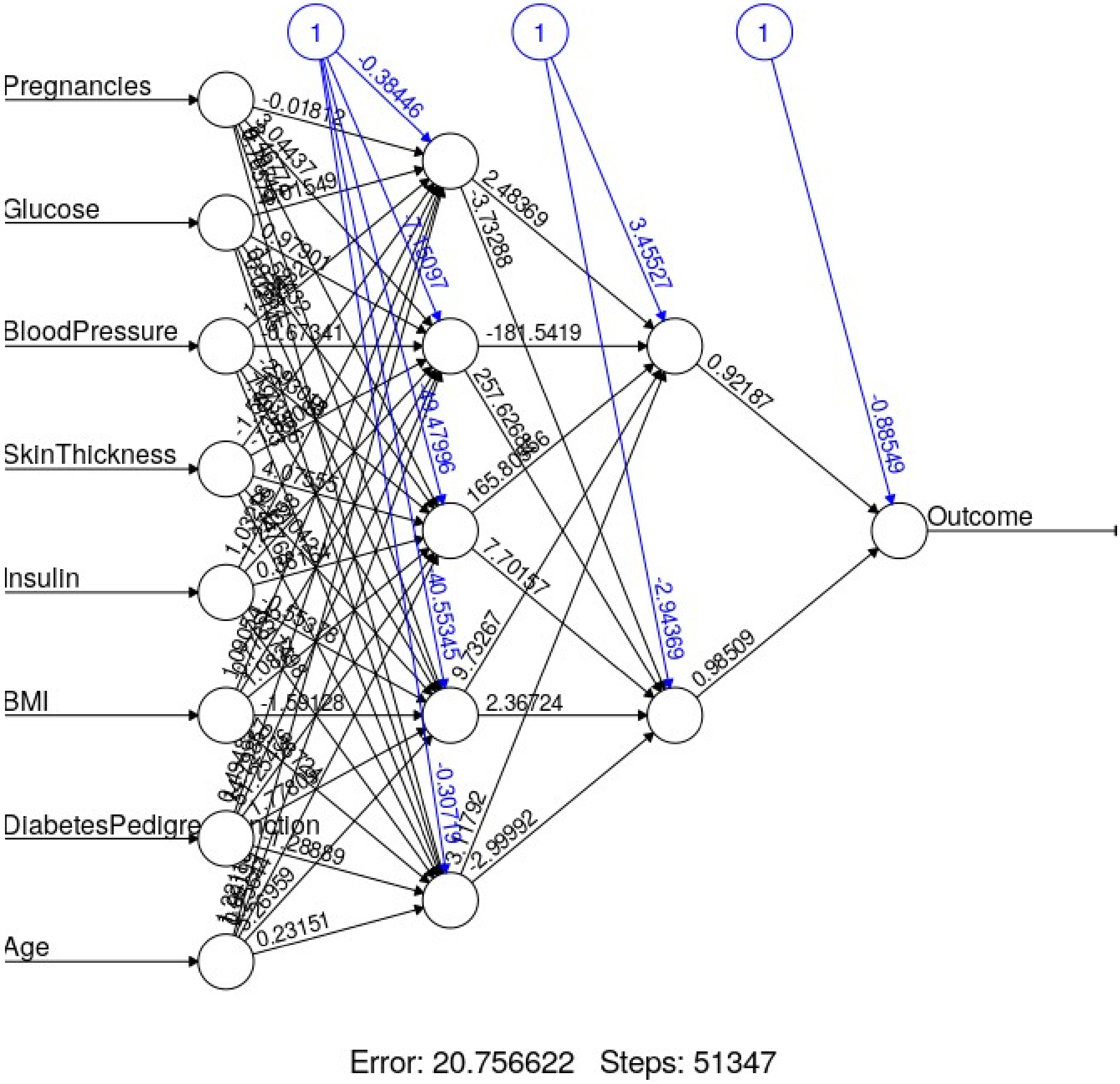
0.4994945

Evaluation:

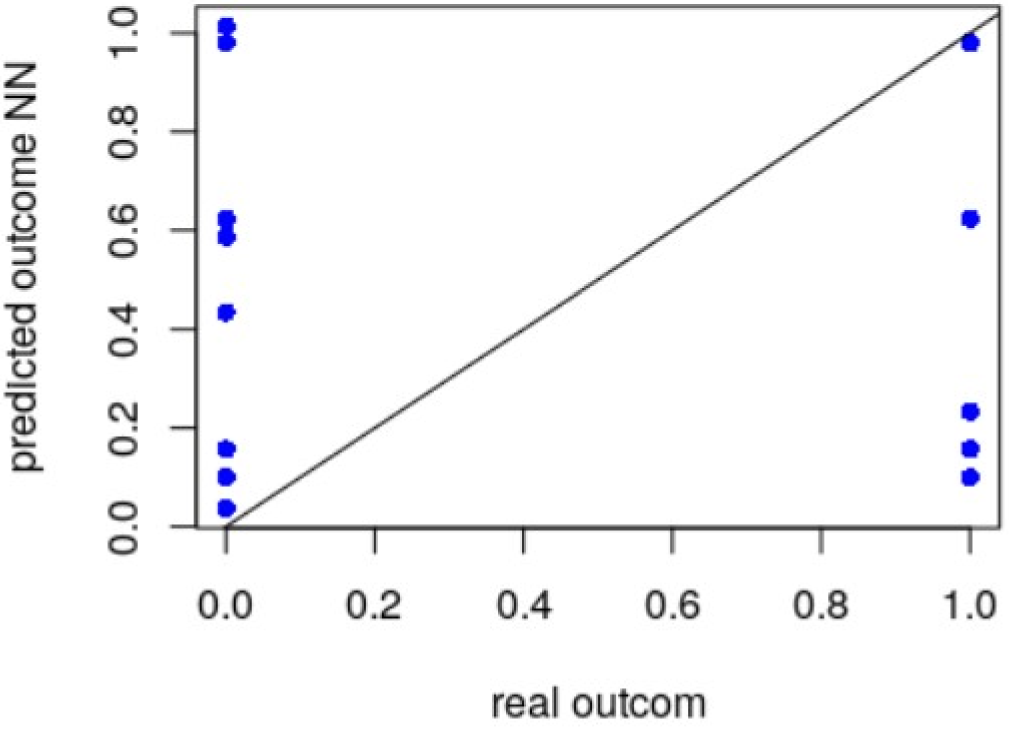


The Root Mean Square Error (RMSE) is 0.4994945, and in comparison with the previous RMSE, there is no significant change.

Retrain with logistic activation function with 2 hidden layes each have 5,2 neurons respectively, having learning rate from (-1,1.2) and a threshold of 0.05:


1
1
1
Pregnancies
Glucose
BloodPressure
SkinThickness
Insulin
BMI
DiabetesPedigreeFunction
Age
Outcome
-0.38446
-0.01812
3.04437
-0.97901
7.80097
-0.67341
49.47996
4.07555
-1.59128
40.55345
-0.30719
1.28889
0.23151
2.48369
-3.73288
-181.5419
7.70157
9.73267
2.36724
3.71792
-2.99992
3.45527
-2.94369
0.92187
0.98509
-0.88549
Error: 20.756622 Steps: 51347

Evaluation:



The Root Mean Square Error (RMSE) is `0.4962181` , there is no significant change.

References:

[1] Lab code and lecture notes.
[2] https://www.analyticsvidhya.com/blog/2016/12/introduction-to-feature-selection-methods-with-an-example-or-how-to-select-the-right-variables/
[3] https://statisticsglobe.com/loop-through-data-frame-columns-rows-in-r/
[4] https://stackoverflow.com/questions/10085806/extracting-specific-columns-from-a-data-frame
[5] https://www.datacamp.com/community/tutorials/neural-network-models-r