Student Name: Qiuyu Luo

Date: 12/15/2020

**Final Report:** **Predicting likelihood of diabetes based on sign symptoms**

**I. Introduction**

Nowadays, diabetes has become one of the fastest growing chronic diseases in the world. In US, 34.2 million people have diabetes (10.5% of the US population) and 88 million people have prediabetes (34.5% of the adult US population). Because the asymptomatic phase of diabetes is relatively long and some of the sign symptoms in initial phase are mild, a patient can have diabetes for a long term before he or she get clinical diagnosis. However, without the diagnosis in the early phase, patients will gradually suffer from fatal complications like heart attacks, kidney damage, and other forms of multi-organ damage, these complications would be under good control if there is a proper assessment of the diabetes risk based on the sign symptoms. Besides, the expenditure of managing the diabetes can be a huge financial burden for some low and middle-income countries. Therefore, early likelihood detection of diabetes is beneficial for health of patients and managing national expenditure. This report will address the problems “which model is the best one to predict the likelihood of diabetes based on sign symptom” and “which element(s) have significant influence on prediction results”.

The dataset was collected using direct questionnaires from the patients of Sylhet Diabetes Hospital in Sylhet, Bangladesh and approved by a doctor. The dataset contains 520 observations and 17 attributes. Based on Table 1, the “Age” is a numeric variable, and the other variables are all binary categorical variables. The categorical variables are “Gender”(Male/Female), “Polyuria”(Yes/No), “Polydipsia”(Yes/No), “Sudden weight loss”(Yes/No), “Weakness”(Yes/No), “Polyphagia”(Yes/No), “Genital thrush”(Yes/No), “Visual blurring”(Yes/No), “Itching”(Yes/No), “Irritability”(Yes/No), “Delayed healing”(Yes/No), “Partial paresis”(Yes/No), “Muscle stiffness”(Yes/No), “Alopecia”(Yes/No), “Obesity”(Yes/No), “Class”(Positive/Negative). The target variable of the problem is “Class”, where Positive means patients have diabetes and Negative means patients don’t have diabetes. Therefore, the problems analyzed are classification problems. Except “Class”, other variables like the “Age” and “Gender” are all predictors in the addressed problems.

**II. Data mining methods and models development**

To select the best model based on the classification prediction results, three data mining methods are applied to this analysis:

* Logistic regression
* Classification tree
* Random forests

Before applying resampling method to train and test the model, the whole dataset is used to fit the model to present an overview of the models. The analysis is focus on prediction instead of inference, therefore models will be built using all predictors in the dataset.

For logistic regression, the number of Positive and Negative of target variable “Class” is 320 and 200 according to Table 1. The imbalance ratio is 0.625 and very close to 0.5, so in this analysis 0.5 is used as the threshold for logistic regression. After fitting the model with the whole dataset, the AIC is 205.7 and residual deviation is 171.7 based on Table 2.

For classification tree, the whole dataset is used to fit the model, then cv.tree is used to select the best size(number of terminal nodes) of the fitted tree model. Based on Table 3, the best size of this tree model is 14 or 13. Then prune.tree function is used to obtain the final tree model. Based on Table 4**,** the variables actually used in tree construction are “Polyuria”, “Gender”, “Alopecia”, “Age”, “Polydipsia”, “Irritability”, “Genital thrush”, “Muscle stiffness”, “Itching”, “delayed healing”. The residual mean deviance is 0.2609 and misclassification error rate is 0.0442. The Figure 5 shows how a classification is made, for example, if a patient has the symptom of polyuria and polydipsia, this patient will be classified as “Positive” based on the constructed tree model.

For random forests, due to the classification model the mtry equals to the square root of the number of predictors. Based on the Table 6, the mtry is 4, the number of trees is 500. The Out-of-bag error rate is 1.73%, the Negative class error is 0.025, the Positive class error is 0.0125.

**III. Model Evaluation**

To evaluate the performance of each model, the analysis applies 5-fold cross validation to split the training and testing dataset. Each time four folds will be used as training data and one fold will be used as testing data, and the same testing data is applied to all the models.

The performance evaluation measures used in the analysis are accuracy, recall (true positive rate, type II error) and precision (positive predictive value) from each model. With 5-fold cross validation, the analysis will compare the mean of accuracy, mean of recall and mean of precision to select the best model. In this dataset, the level “Positive” of target variable “Class” is defined as positive to calculate the recall and precision, because a patient with diabetes is diagnosed as non-diabetes is a more serious case, which can cause delay in treatment and result in following fatal complications.

The formula of accuracy, recall and precision:

* Accuracy:
* Recall:
* Precision:

For Logistic regression, Figure 7 presents the 5 confusion matrices from the 5-fold cross validation. The accuracy of each fold is 0.9010, 0.9259, 0.9464, 0.9109, 0.9082, and the mean of accuracy is 0.9185. The recall of each fold is 0.9138, 0.9524, 0.9444, 0.9286, 0.9123, and the mean of recall is 0.9303. The precision of each fold is 0.9138, 0.9231, 0.9714, 0.9420, 0.9286. The mean of precision is 0.9358.

For classification tree, Figure 8 presents the 5 confusion matrices from the 5-fold cross validation. The accuracy of each fold is 0.8713, 0.9167, 0.9286, 0.9208, 0.8980, and the mean of accuracy is 0.9071. The recall of each fold is 0.7931, 0.9365, 0.8889, 0.9143, 0.9298 and the mean of recall is 0.8925. The precision of each fold is 0.9787, 0.9219, 1.0000, 0.9697, 0.8983 and the mean of precision is 0.9537.

For classification tree, Figure 9 presents the 5 confusion matrices from the 5-fold cross validation. The accuracy of each fold is 0.9604, 0.9722, 0.9911, 1.0000, 0.9694, and the mean of accuracy is 0.9786. The recall of each fold is 0.9310, 0.9841, 0.9861, 1.0000, 0.9825 and the mean of recall is 0.9767. The precision of each fold is 1.0000, 0.9688, 1.0000, 1.0000,0.9655 and the mean of precision is 0.9869.

**IV. Result Analysis**

The Table 10 presents the comparison of the performance measures of the 3 models. It shows that the three models all have high prediction accuracy over 0.9, and among them random forests with 5-fold cross validation, mtry=4 and 500 trees has the best prediction performance with the highest mean of accuracy, mean of recall and mean of precision: 0.9786, 0.9767 and 0.9869. In this dataset, the accuracy of classification tree is the lowest, but on the other hand the classification tree plots can provide the good interpretability so it’s easier to figure out how an instance is classified.

For the second problem “which element(s) have significant influence on prediction results”, the results can be obtained through the variable importance plot. Based on Figure 11, when the variables “Polyuria” and “Polydipsia” is excluded, the mean decrease accuracy and mean decrease GINI is much larger compared with the rest variables. Therefore, the variables “Polydipsia” and “Polyuria” have more significant impact than other variables in the dataset.

**V. Conclusion and Implications**

In conclusion, the random forest has the best performance in predicting likelihood of diabetes based on sign symptoms from this dataset, it can classify almost 98% of the instances correctly. And the sign symptom polyuria and polydipsia are very significant elements for determining the prediction results.

With the conclusion, it would be very useful to build a user-friendly tool with the random forest algorithm to predict the risk of diabetes and prediabetes based on the sign symptoms. This tool can help the patients and health workers to achieve early diabetes risk assessment in rural areas where the medical facility is not available. The early assessment would be very helpful to prevent the fatal complications in some degree and it’s beneficial for saving both individual and national expenditure.

Because polydipsia and polyuria are two important sign symptoms among the sign symptoms of diabetes, if a person realize he or she has these two symptoms, it would be essential to pay more attention to the health situation and consult the professional health workers to prevent delayed treatment.

**Appendix**

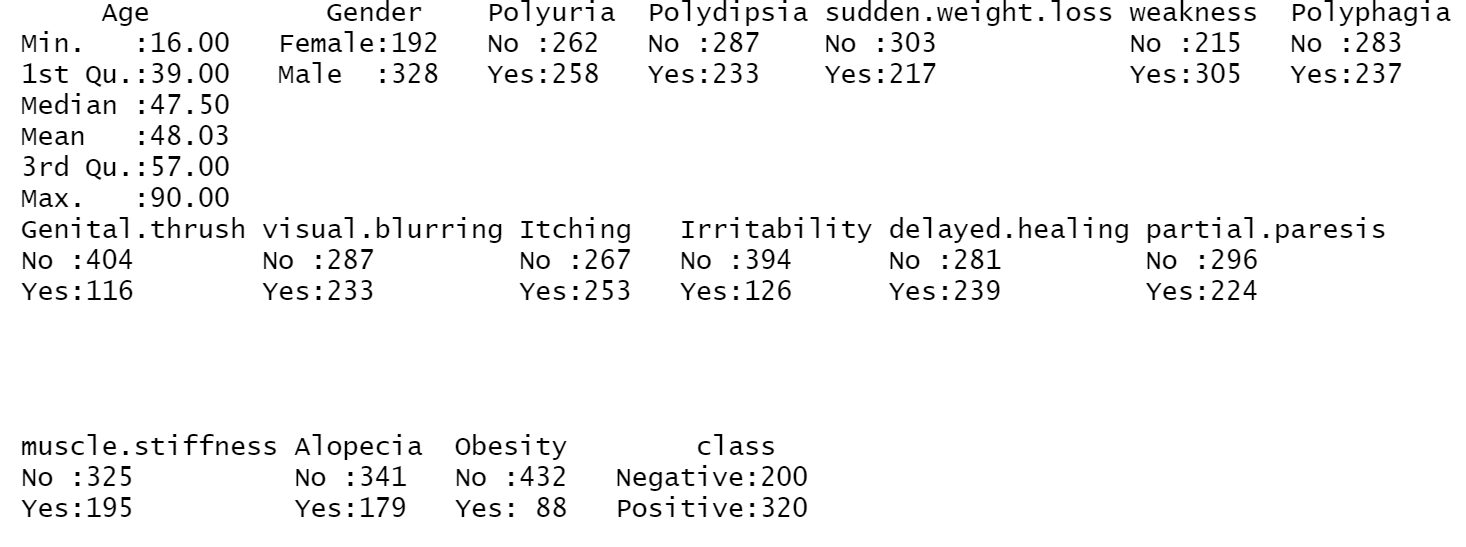


Table 1: Attributes list

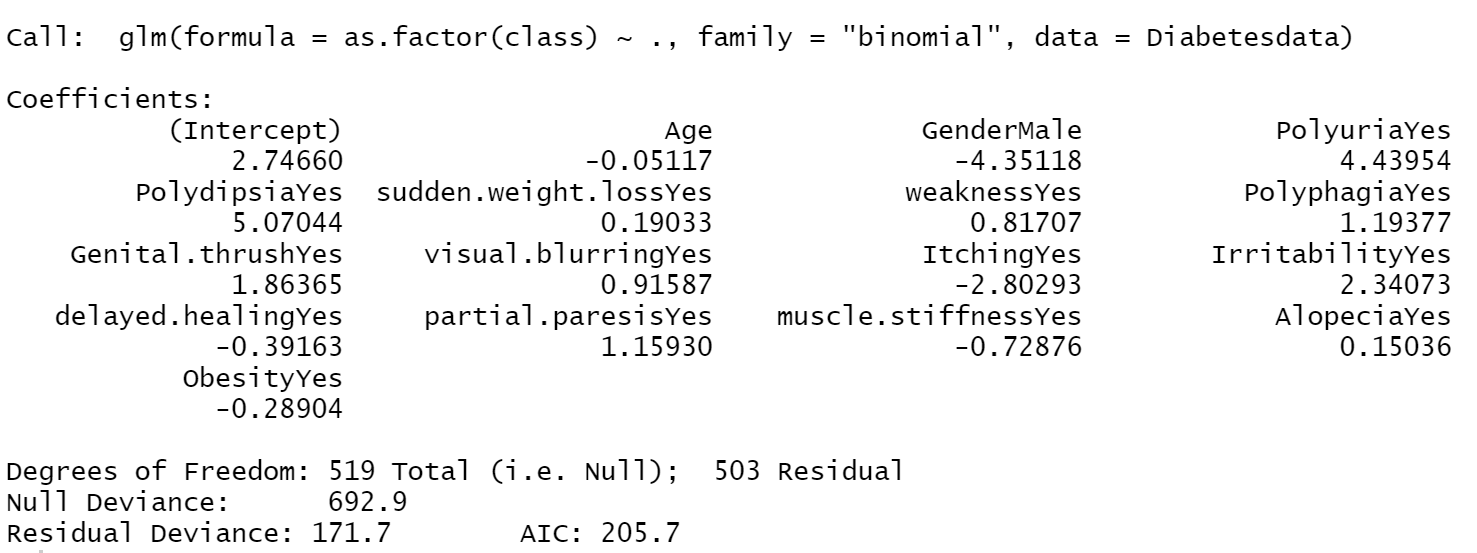


Table 2: Summary of logistic regression model

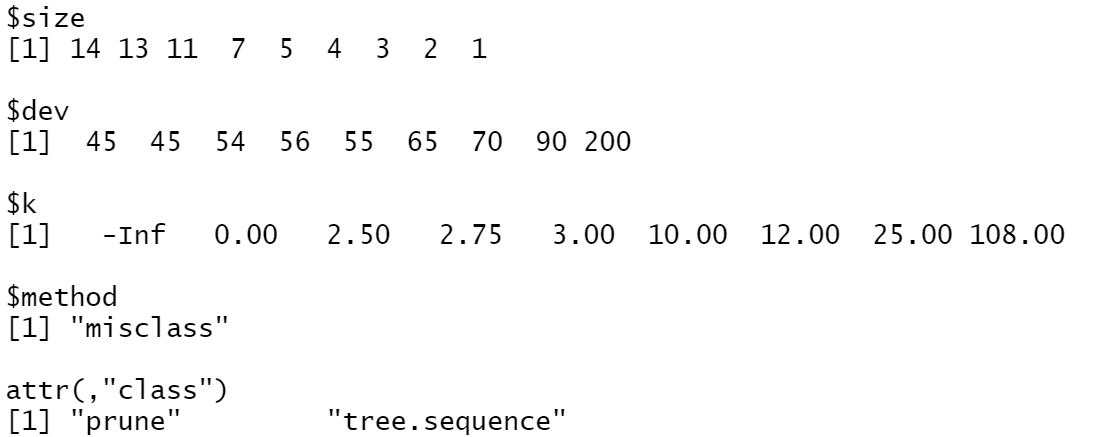


Table 3: cv.tree function

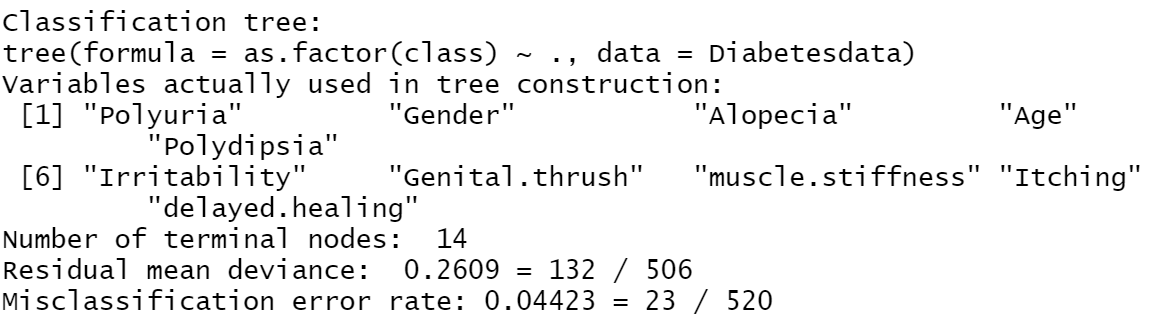


Table 4: Summary of classification tree

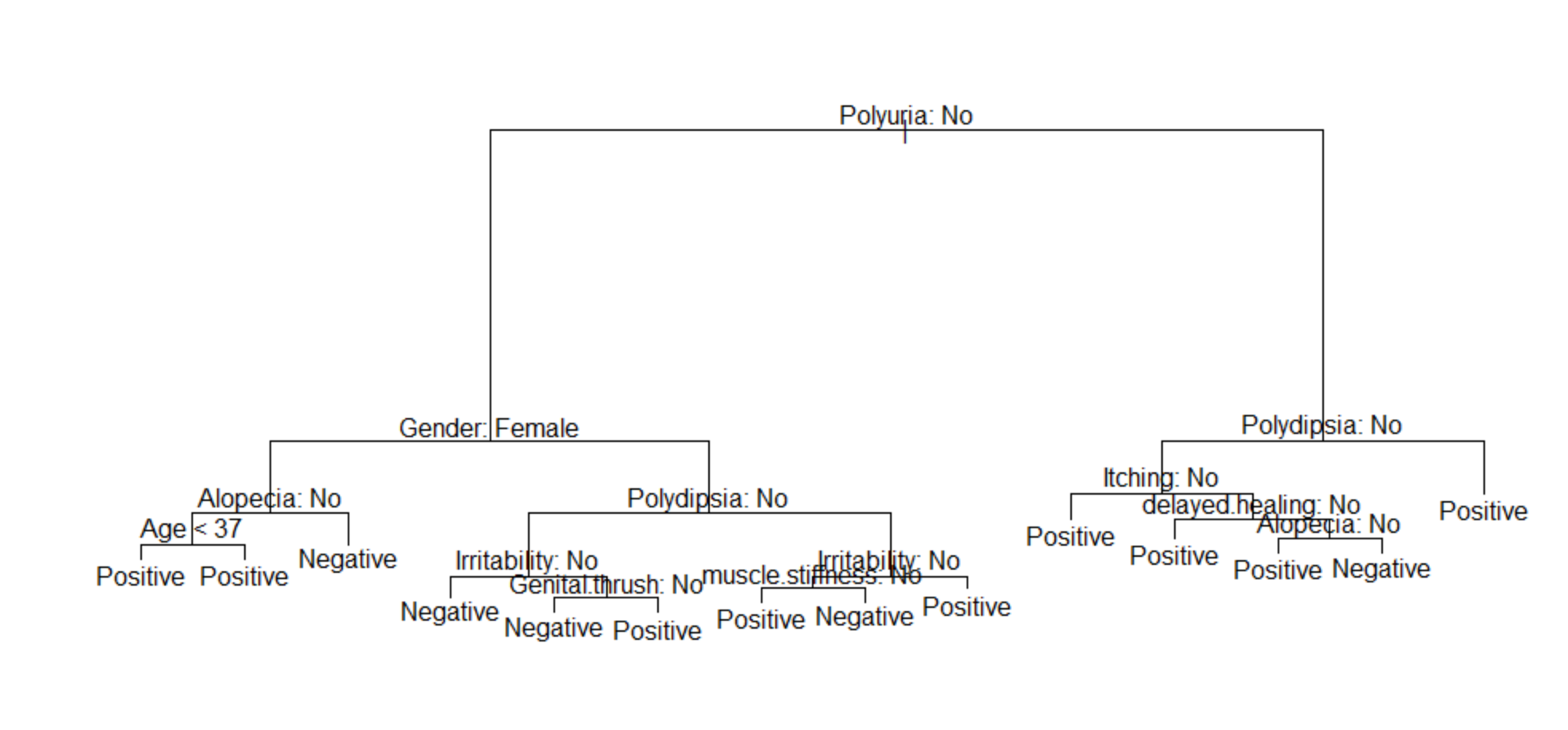


Figure 5: Classification tree plot

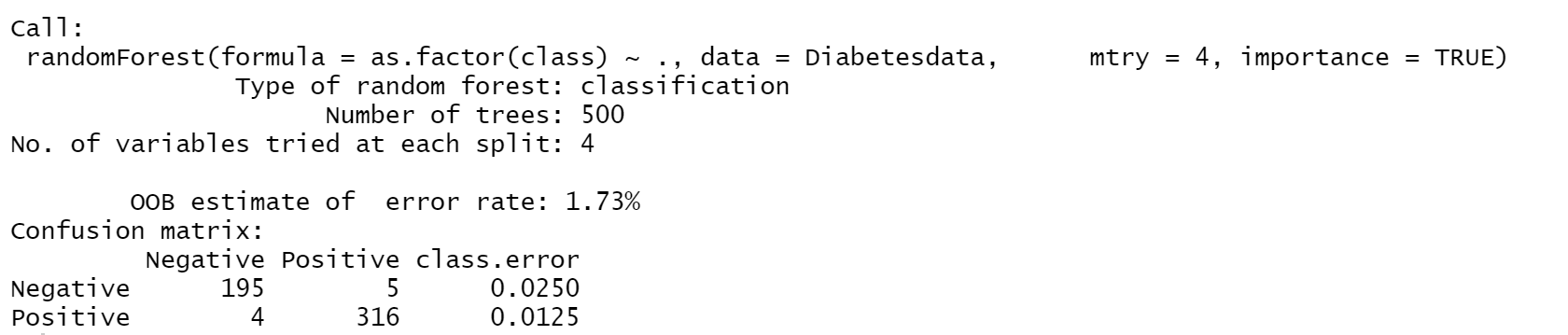


Table 6: Summery of random forest

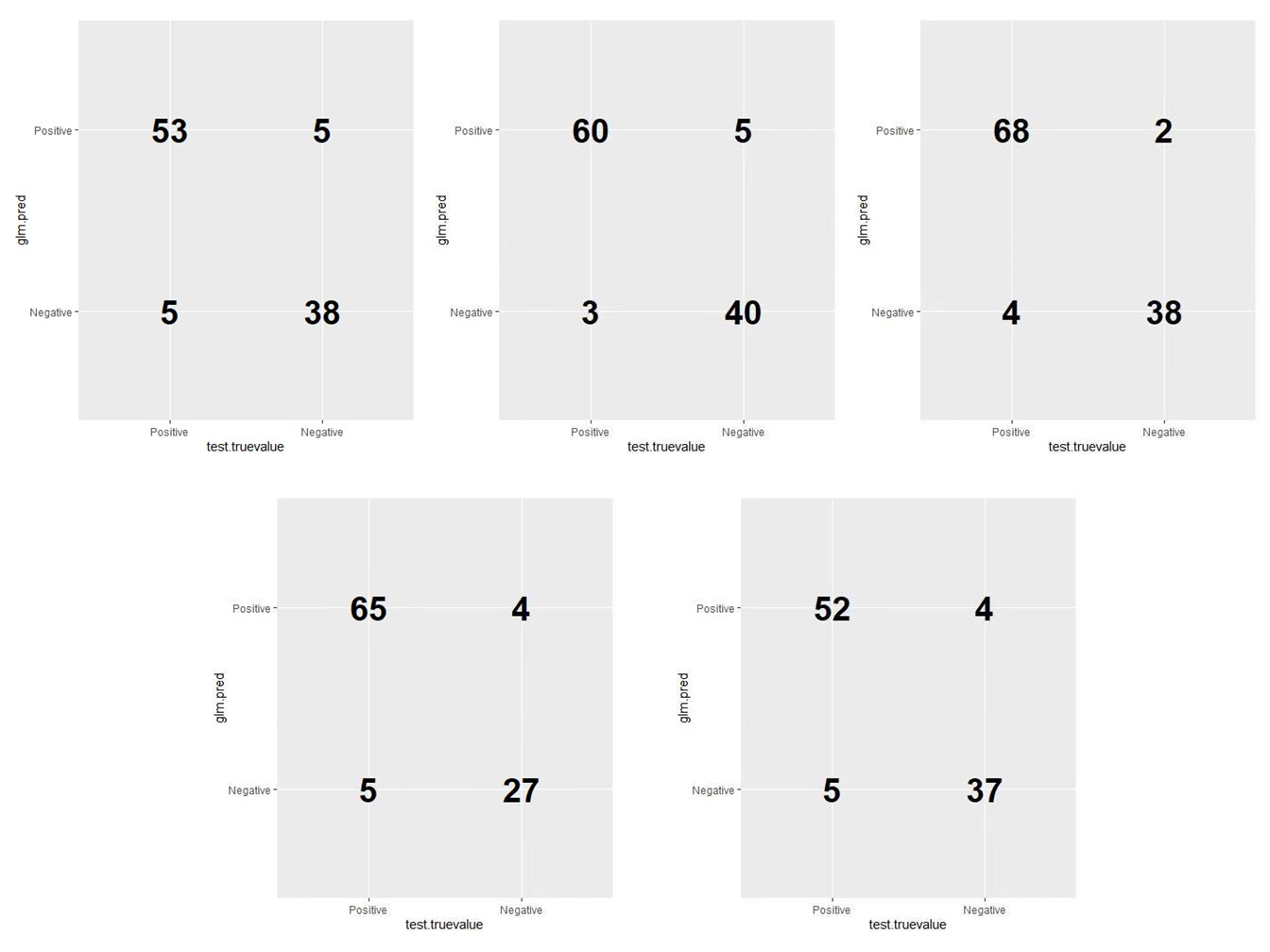


Figure 7: Confusion matrices of logistic regression

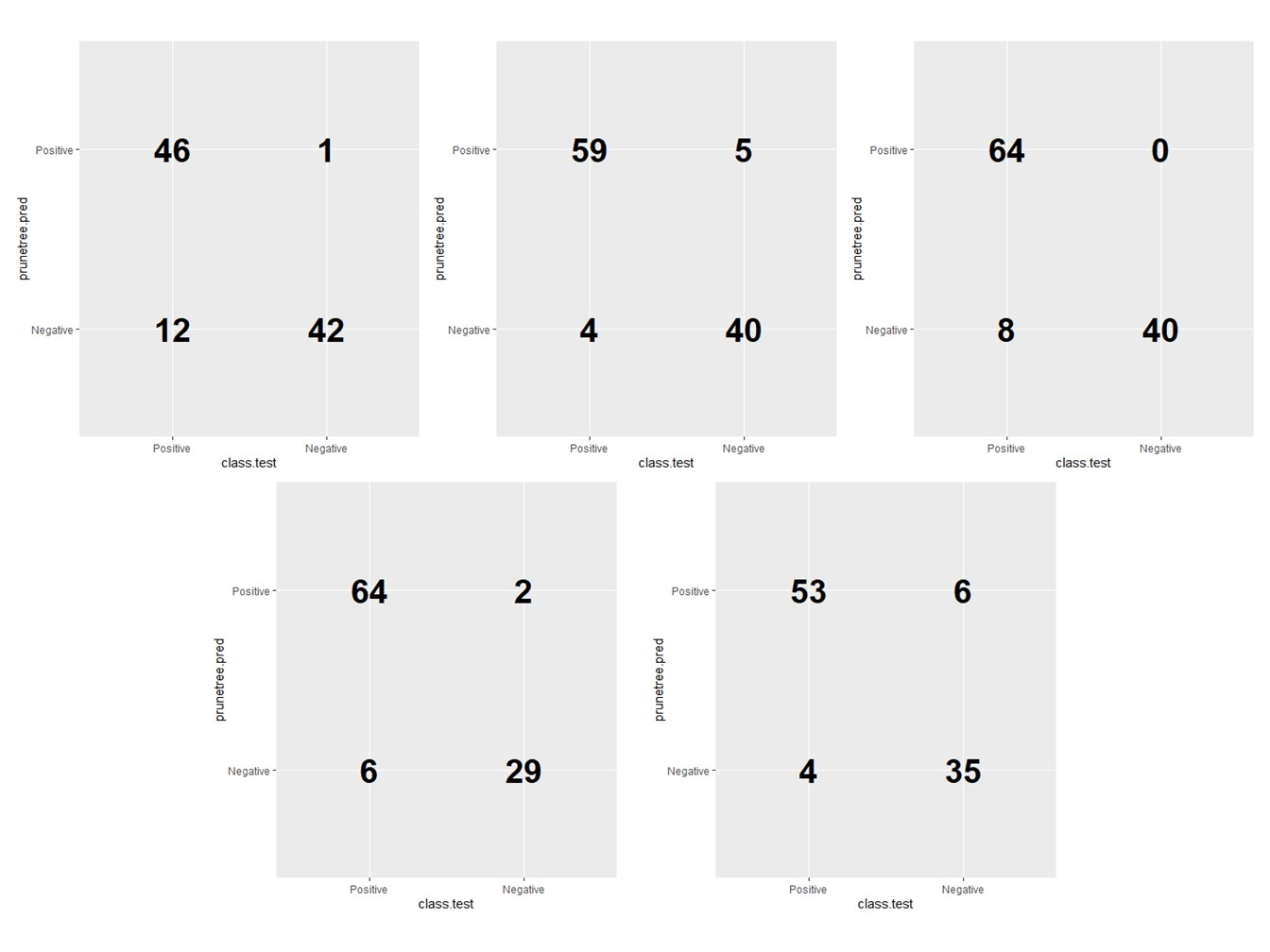


Figure 8: Confusion matrices of classification tree



Figure 9: Confusion matrices of random forest

|  |  |  |  |
| --- | --- | --- | --- |
|  | Logistic Regression | Classification Tree | Classification Forest |
| Accuracy | 0.9185 | 0.9071 | 0.9786 |
| Recall | 0.9303 | 0.8925 | 0.9767 |
| Precision | 0.9358 | 0.9537 | 0.9869 |

Table 10: Prediction performance comparison

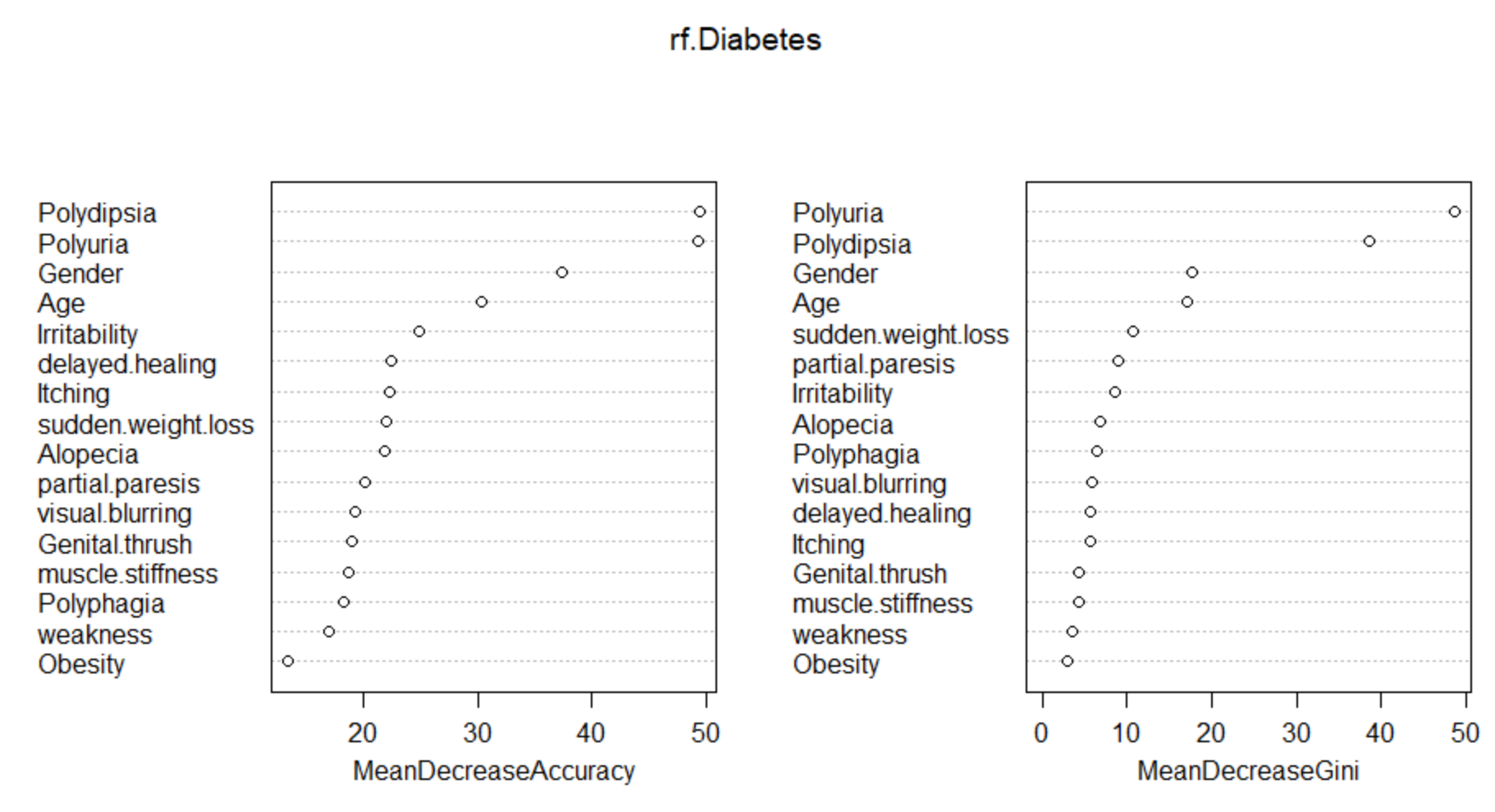


Figure 11: Variable importance plot

**Reference**

1. “National Diabetes Statistics Report, 2020”. Atlanta, GA: Centers for Disease Control and Prevention, U.S. Dept of Health and Human Services, 2020, accessed Dec 5, 2020, https://www.cdc.gov/diabetes/data/statistics-report/index.html.

2. “Early stage diabetes risk prediction dataset”. University of California, Irvine: Machine Learning Repository, Center for Machine Learning and Intelligent Systems, 2020, accessed Dec 10, 2020, https://archive.ics.uci.edu/ml/datasets/Early+stage+diabetes+risk+prediction+dataset.#.

3. Islam M.M.F., Ferdousi R., Rahman S., Bushra H.Y. (2020) “Likelihood Prediction of Diabetes at Early Stage Using Data Mining Techniques”. In: Gupta M., Konar D., Bhattacharyya S., Biswas S. (eds) Computer Vision and Machine Intelligence in Medical Image Analysis. Advances in Intelligent Systems and Computing, vol 992. Springer, Singapore. https://doi.org/10.1007/978-981-13-8798-2\_12.

4. “Plot confusion matrix in R using ggplot”. Stack Overflow, 2020, accessed Dec 12, 2020. https://stackoverflow.com/questions/37897252/plot-confusion-matrix-in-r-using-ggplot.