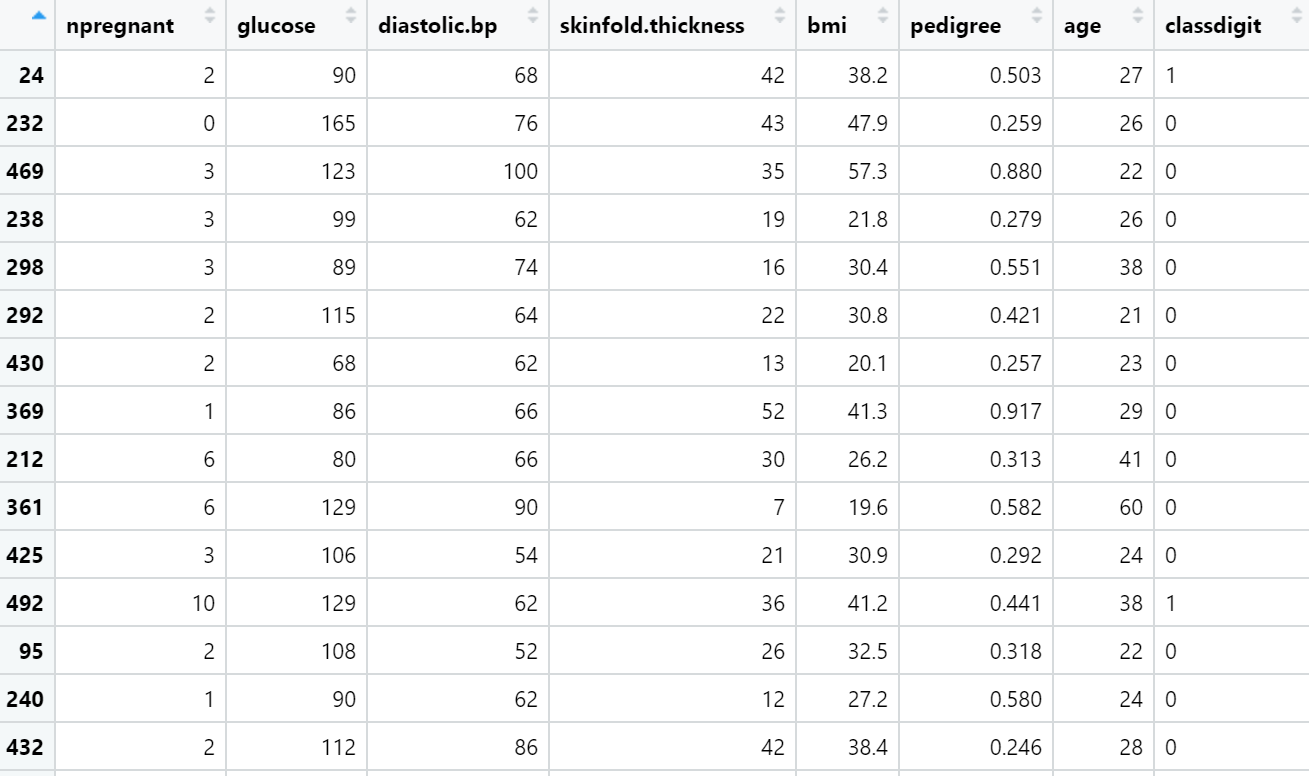
Process:

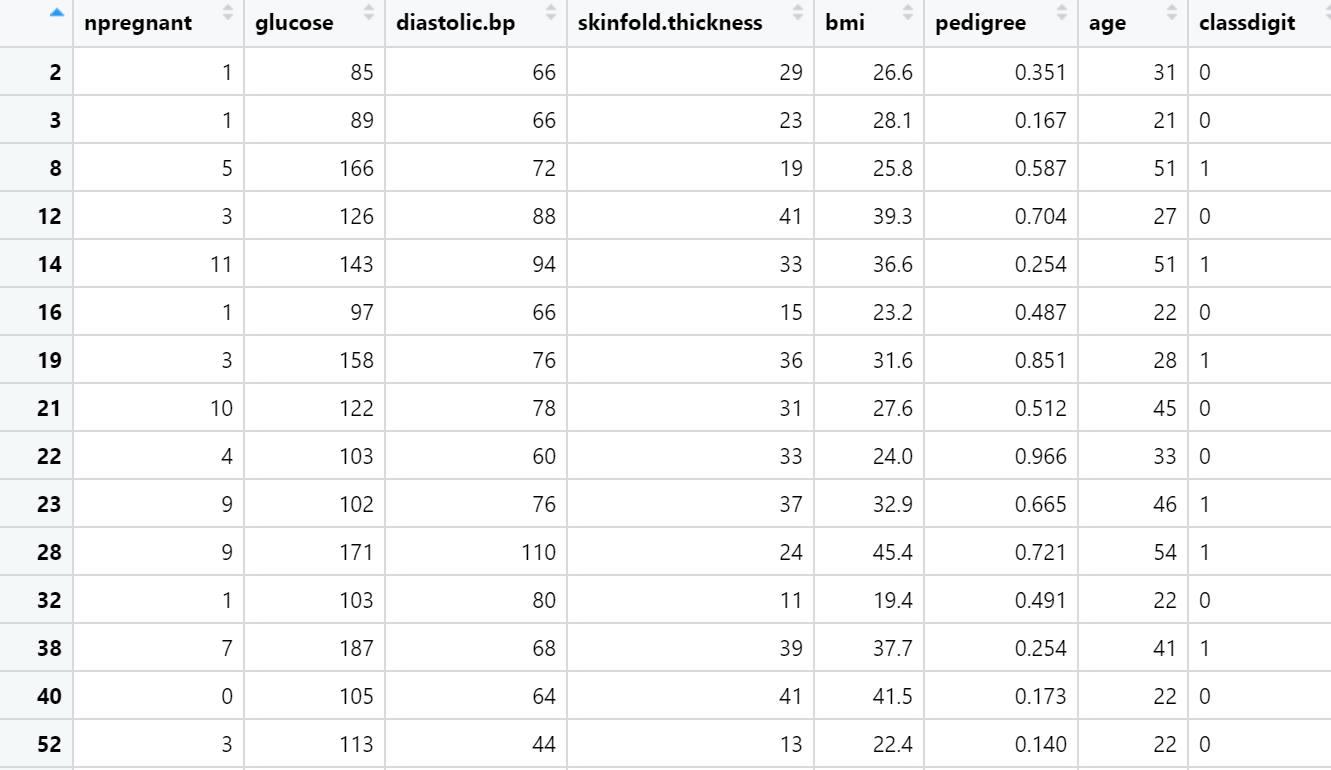
1. **Read data**: I read ‘pima’ dataset.
2. **Made the Train and Test data**: I divided 70% of data into train and 30 % of data into the test.
3. **The tree model:** I made a model with full variables to compare with the pruned tree model. And I made two confusion matrixes to compare two models.
4. **The Boosting model:** I made a boosting model and find the best shrinkage value for the boost model (n.tree=1000). I also calculated test accuracy for the model.
5. **Random Forest model:** I made a RF model. I got the result of the Random Forest model’s importance factor and test Accuracy
6. **Patrial dependence plot:** I plotted the partial dependence plot with the Random Forest model for 4 important variables.

Output:

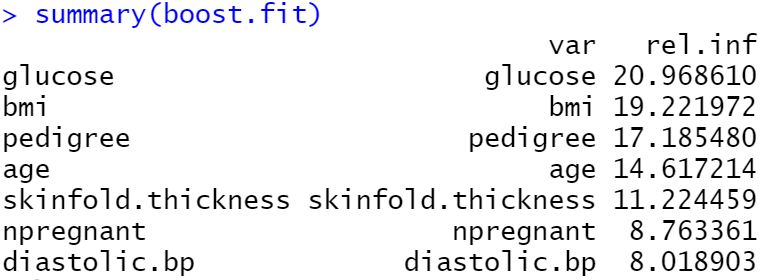
Train



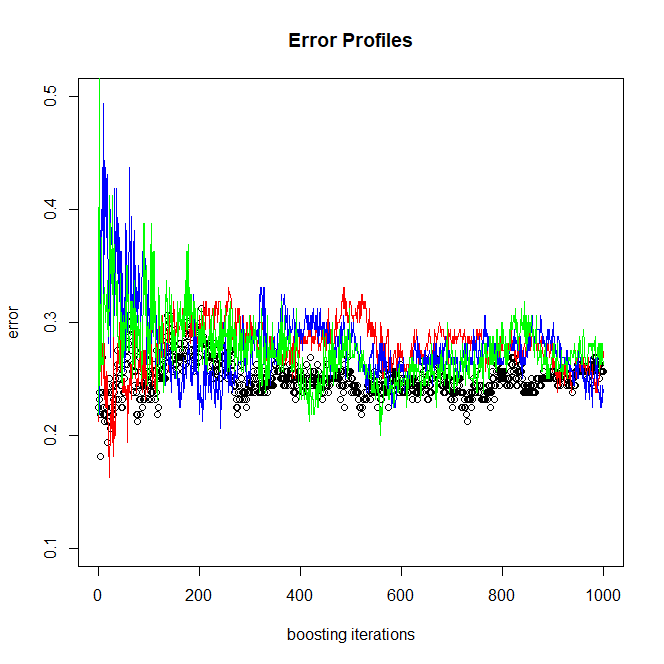
Test

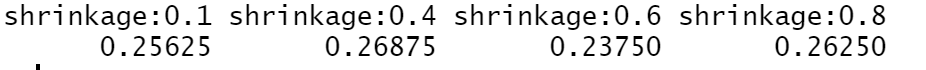


Feature importance

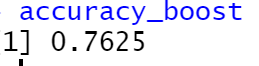


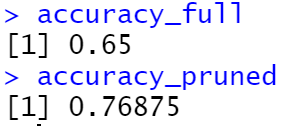
The best shrinkage value for the boost model(n.tree=1000)



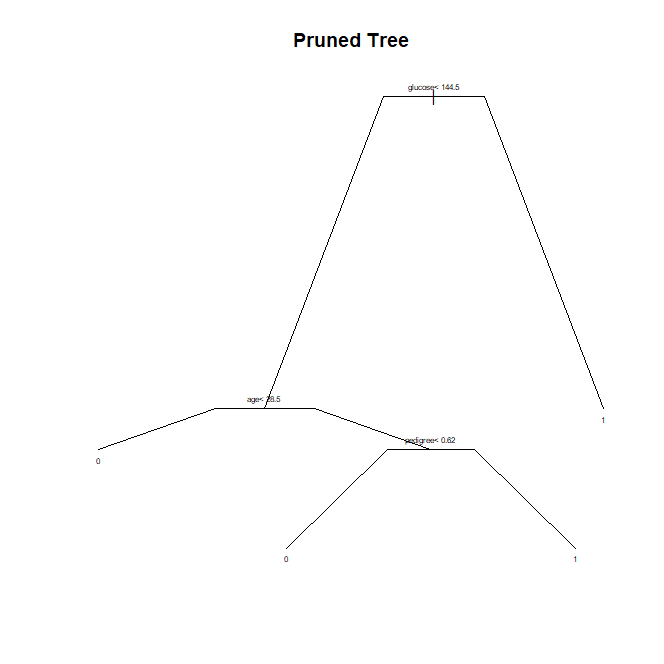


Test prediction accuracy of boost

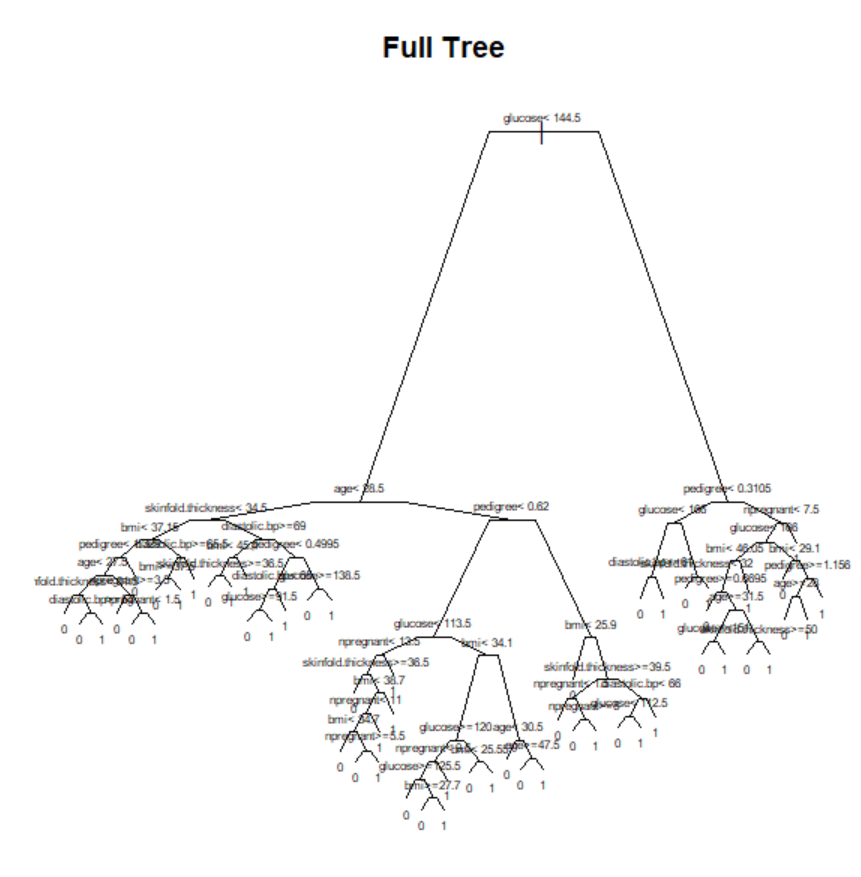




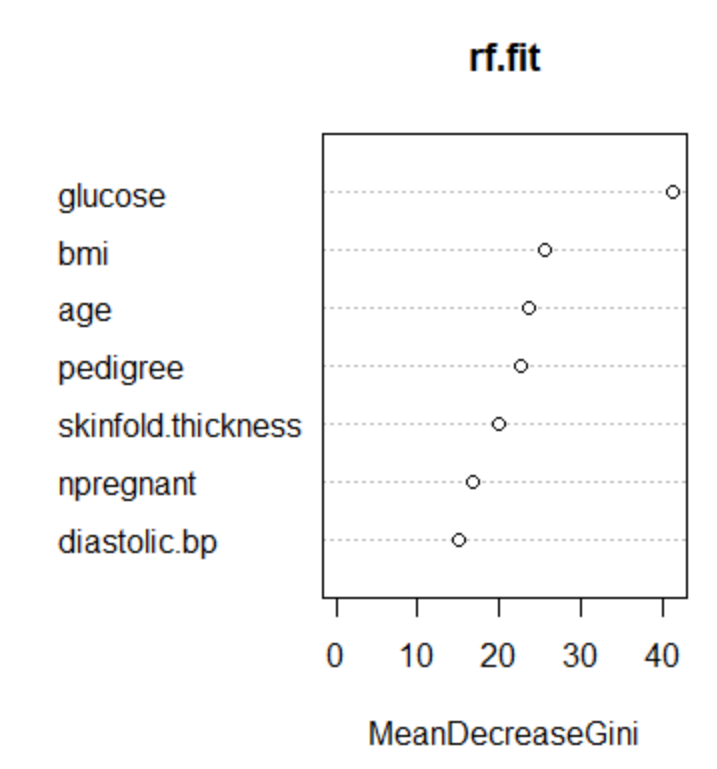
Pruned Tree shape

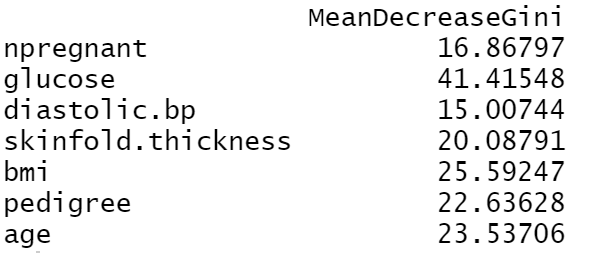


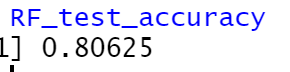
Full tree shape



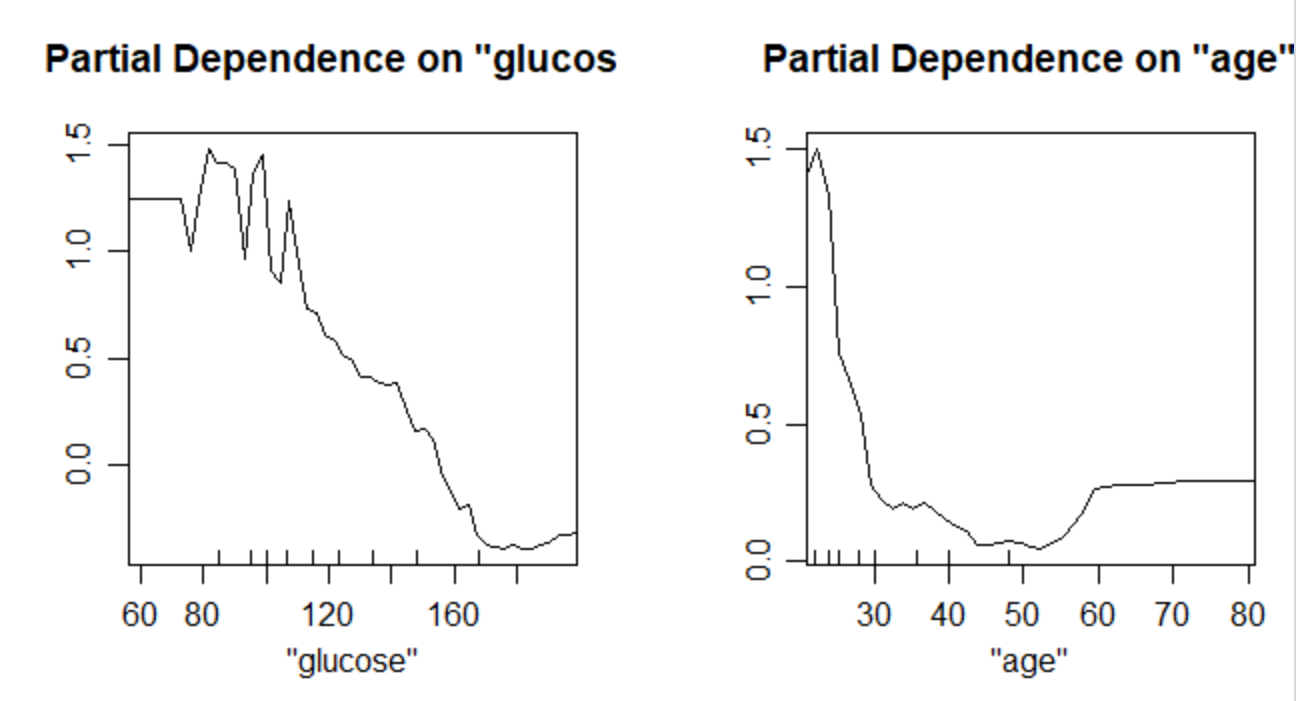
Random Forest model importance factor and test Accuracy

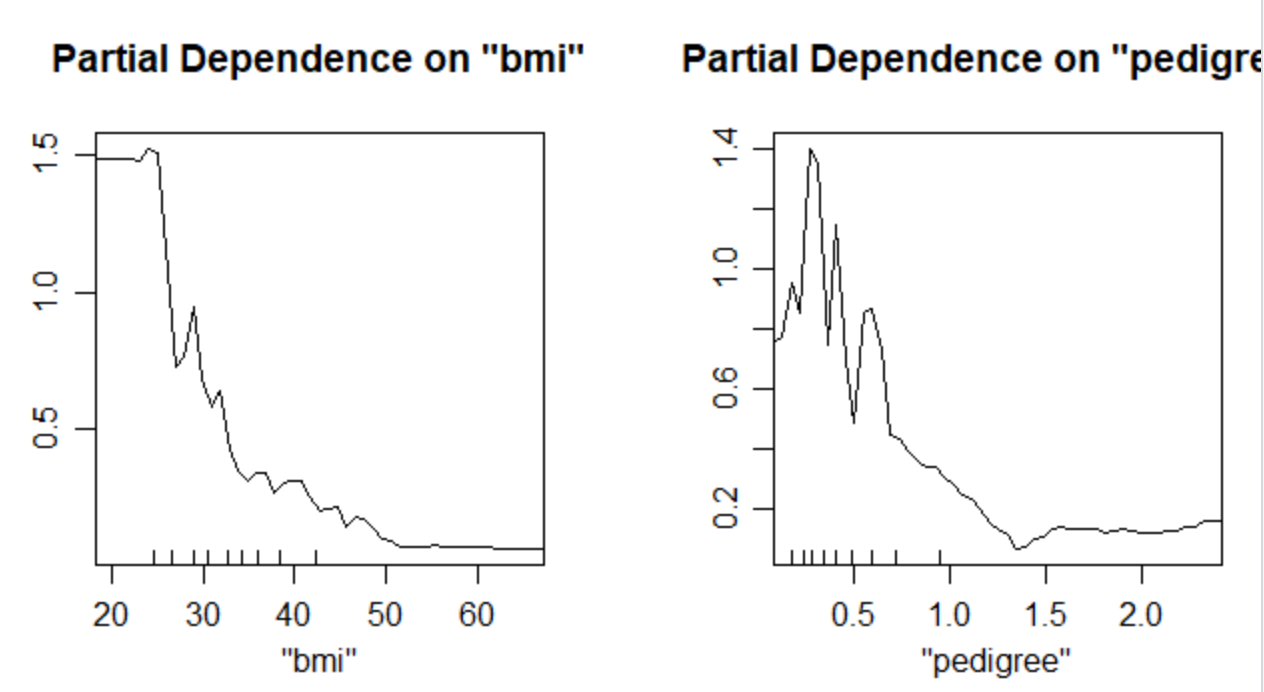


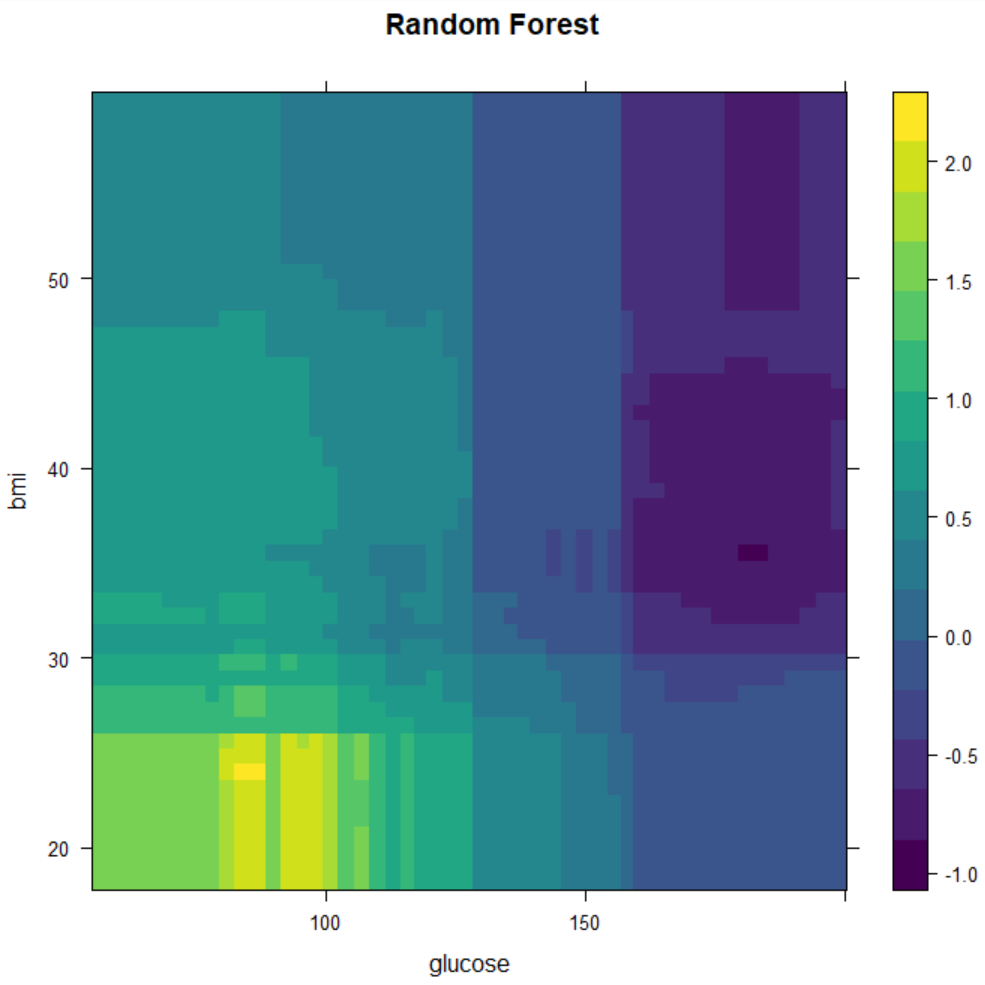




Partial Dependence plot







Disscusion:

***Accuracy***

***TREE***

Accuracy (Full tree test prediction): 0.65

Accuracy (Pruned tree test prediction): 0.76875

The Full tree model may be overfitted by the training set. Therefore, the pruned tree model’s test prediction shows better performance than the full tree model.

***Random Forest***

Accuracy (test error test prediction): 0.80625

As an ensemble model, the Random Forest model shows good performance compared with the tree model.

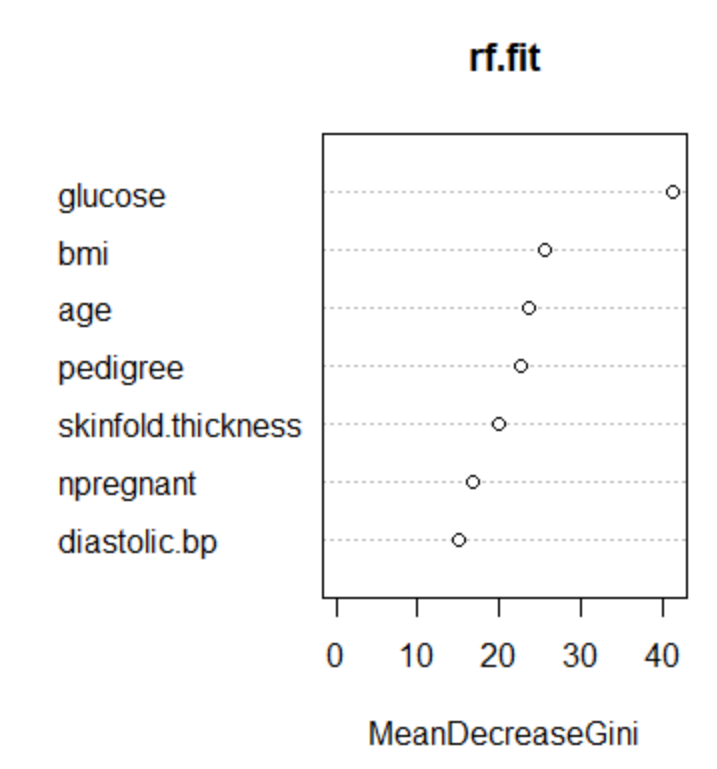
***Boosting***

Accuracy (test error test prediction): 0.7625

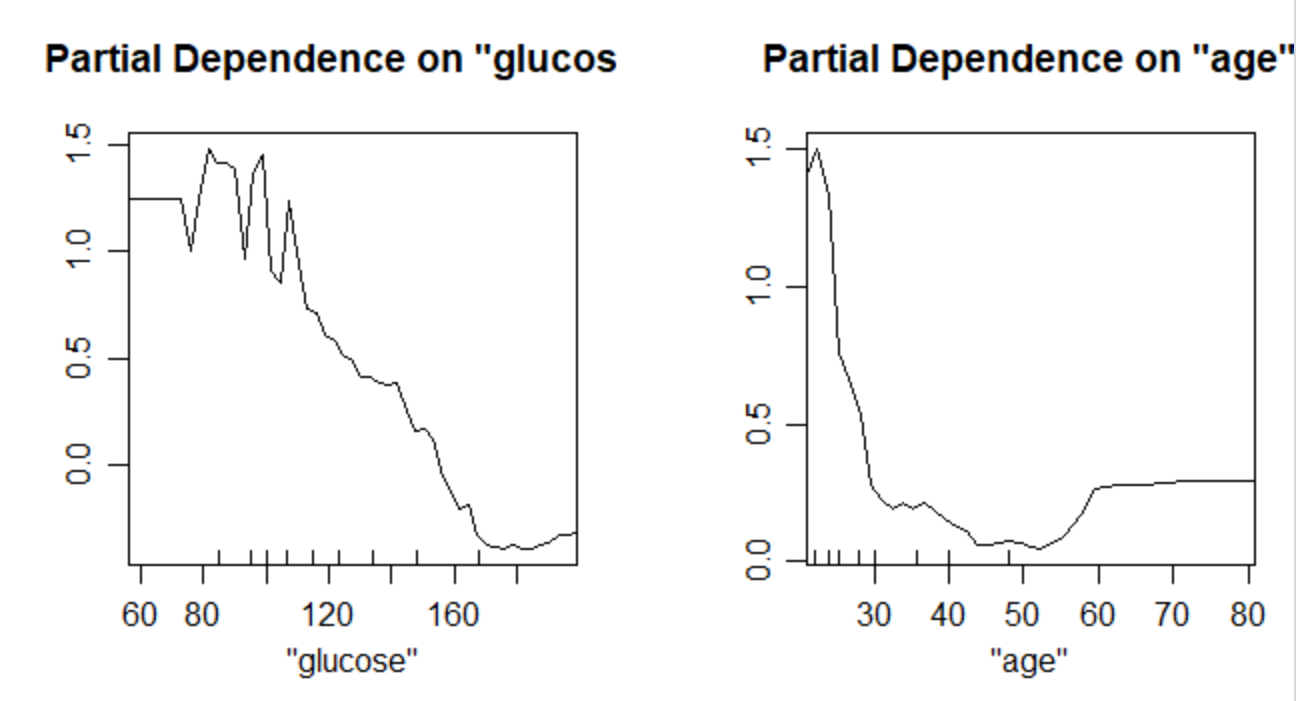
I found the best boosting model by manipulating shrinkage values. The 0.6 was the best shrinkage value for the model which n.trees value is 1000.

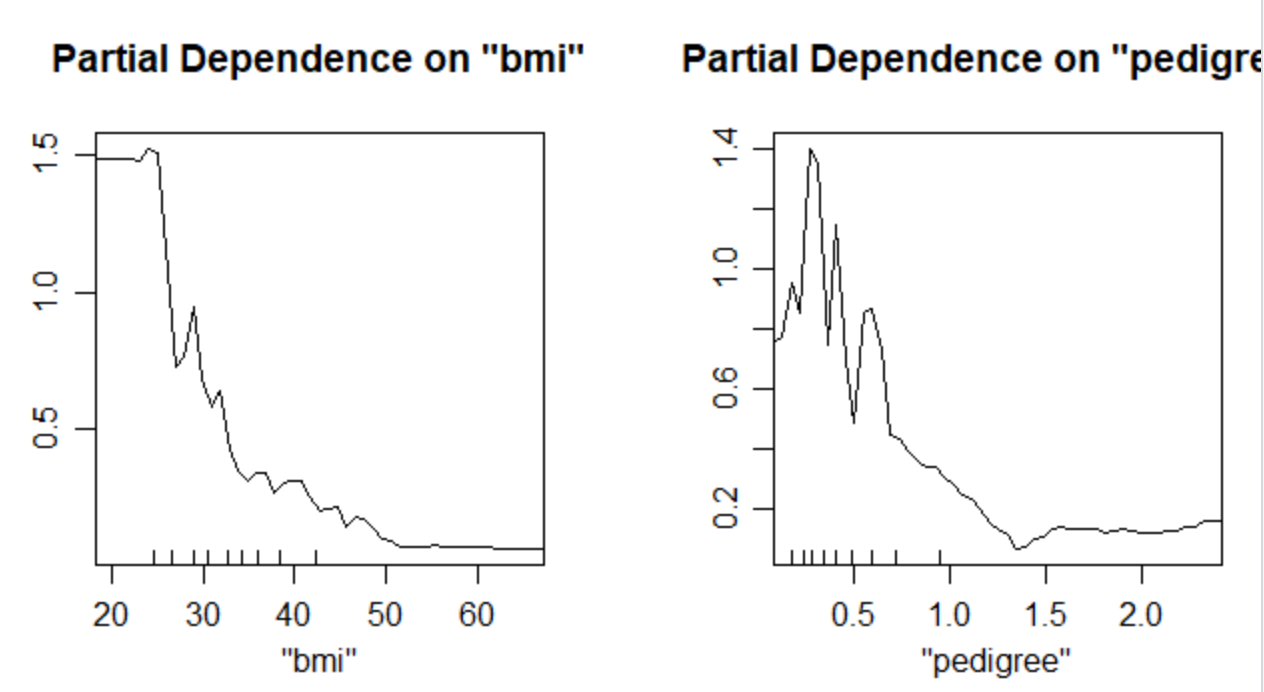
The Random Forest model got 0.80625 accuracy on test data prediction. The Boosting test accuracy is 0.75625, which shows a similar performance to the pruned tree models. The full tree model test accuracy is 0. 65, which shows the worst performance among these models.

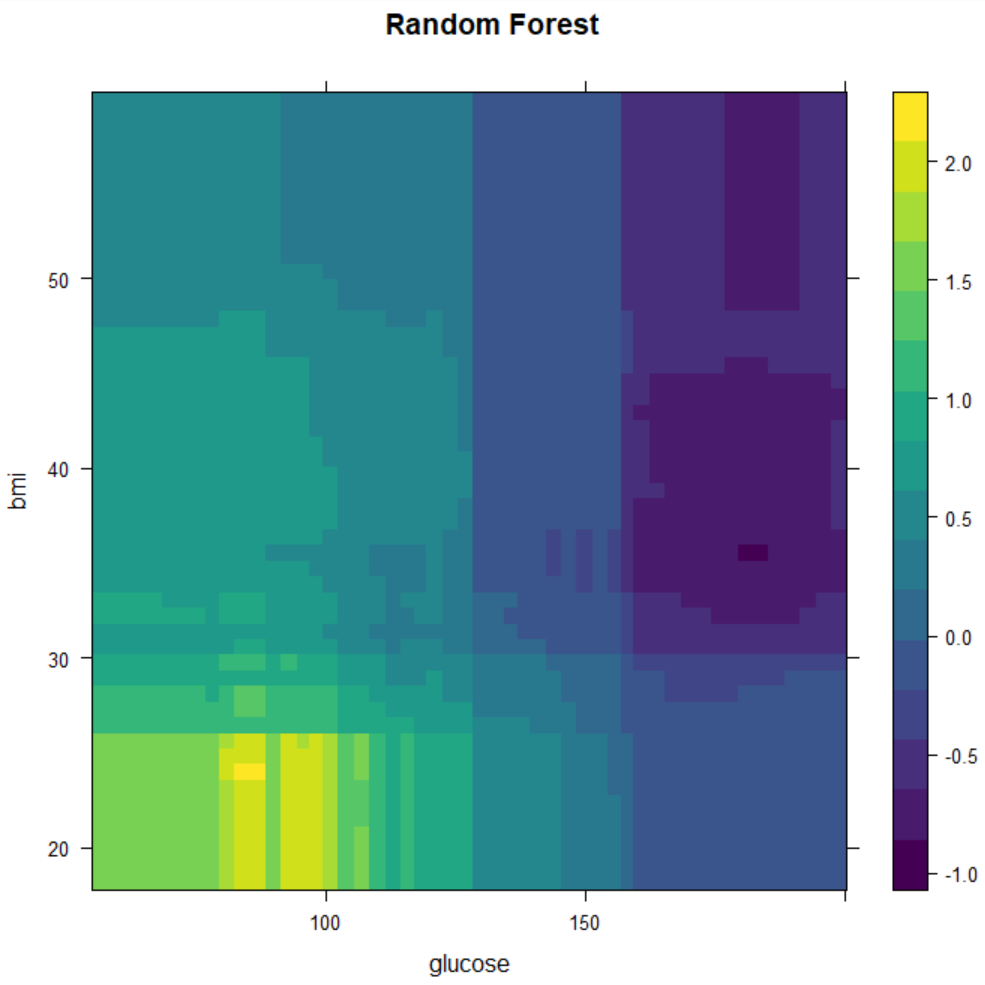
***Partial dependence plot***



I got the importance of factor and test Accuracy evaluated by Gini values. I decided to see the partial dependence plot of the top 4 variables in the data.





According to the partial dependence on ‘glucose’ which is the most important feature in the Random Forest model, under the value of ‘glucose’ 140 have a positive influence on the classification. As the number of glucose values is small, it tends to give a more positive influence on the classification. Likewise, other variables show a similar tendency to the model.

In this matrix, we can see the influence of two variables on the model at the same time. Since ‘bmi’ and ‘glucose’ got a similar distribution for each partial dependence plot(small num large positive influence), this combined matrix shows a similar tendency. In the matrix, we can see a small number of ‘glucose’ and ‘bmi’ range area shows are colored by yellow or yellow-green which means a positive influence on the classification. We can guess that all the other combined matrix with these 4 variables are similar to the matrix above.