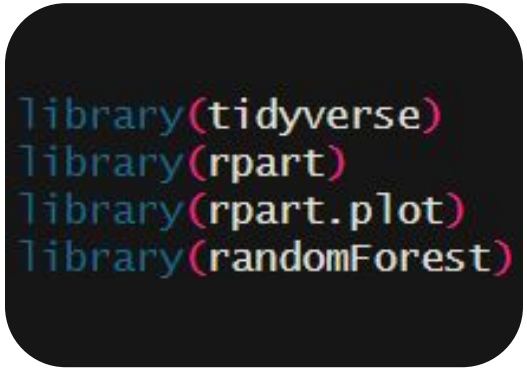
Experimentos com florestas & Importância de variáveis

Baseado em: “*An introduction to statistical learning with applications in R*”

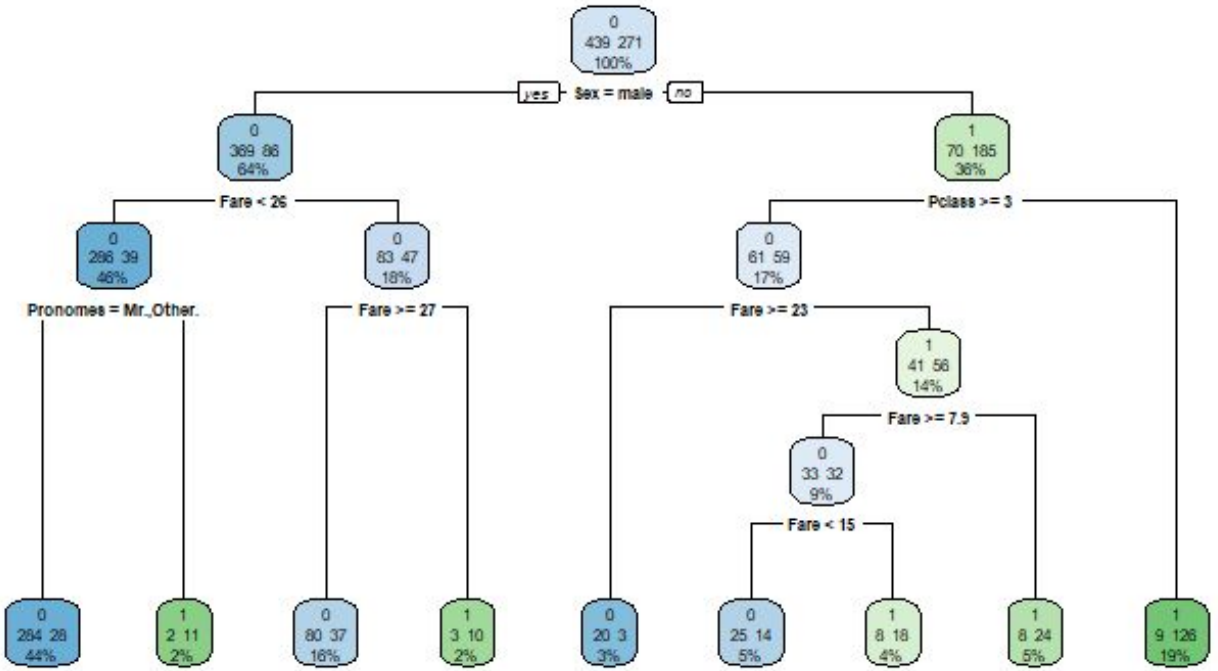
Considerações:

- Dataset usado: Titanic, Diabetes prediction e Abalone
- Recurso extra: ListenData
- Bibliotecas do R Studio usadas:



R studio

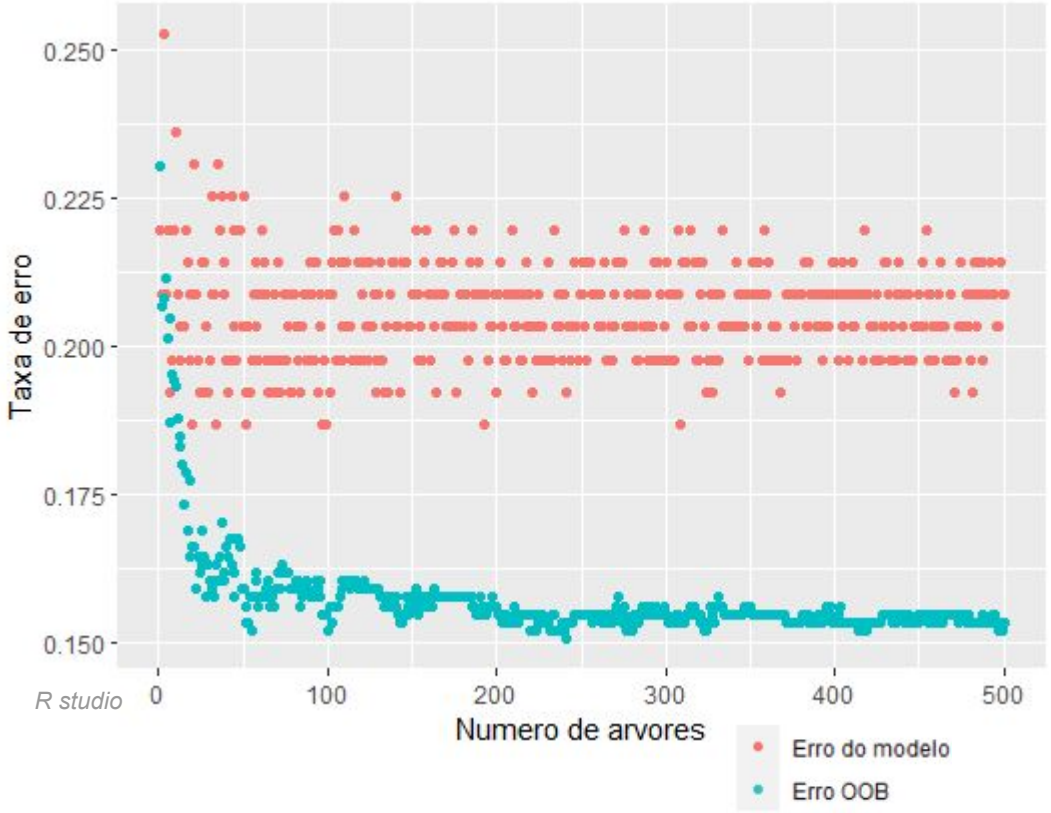
Experimentos com florestas: Árvore



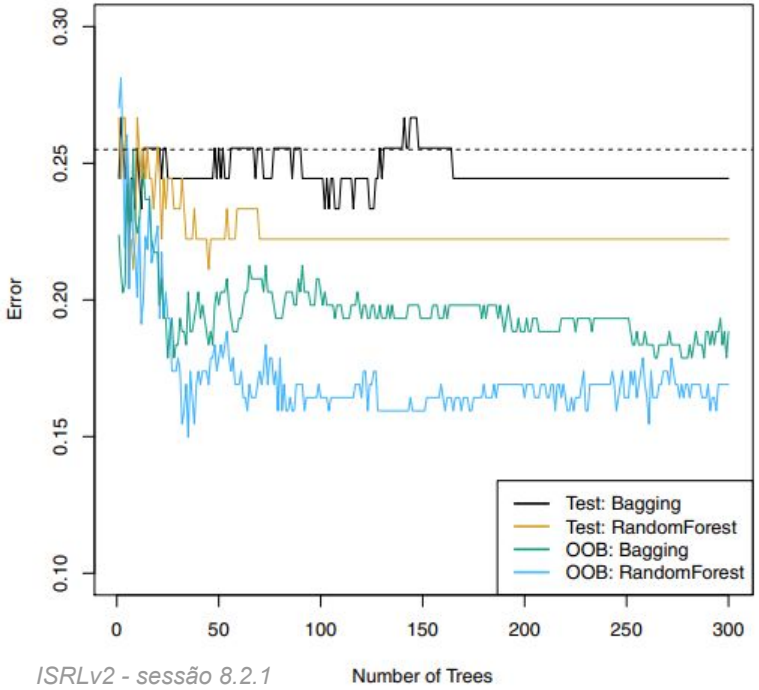
R studio

```
> 1-sum(predict
[1] 0.1758242
```

Experimentos com florestas: *Bagging* & Erro *OOB*

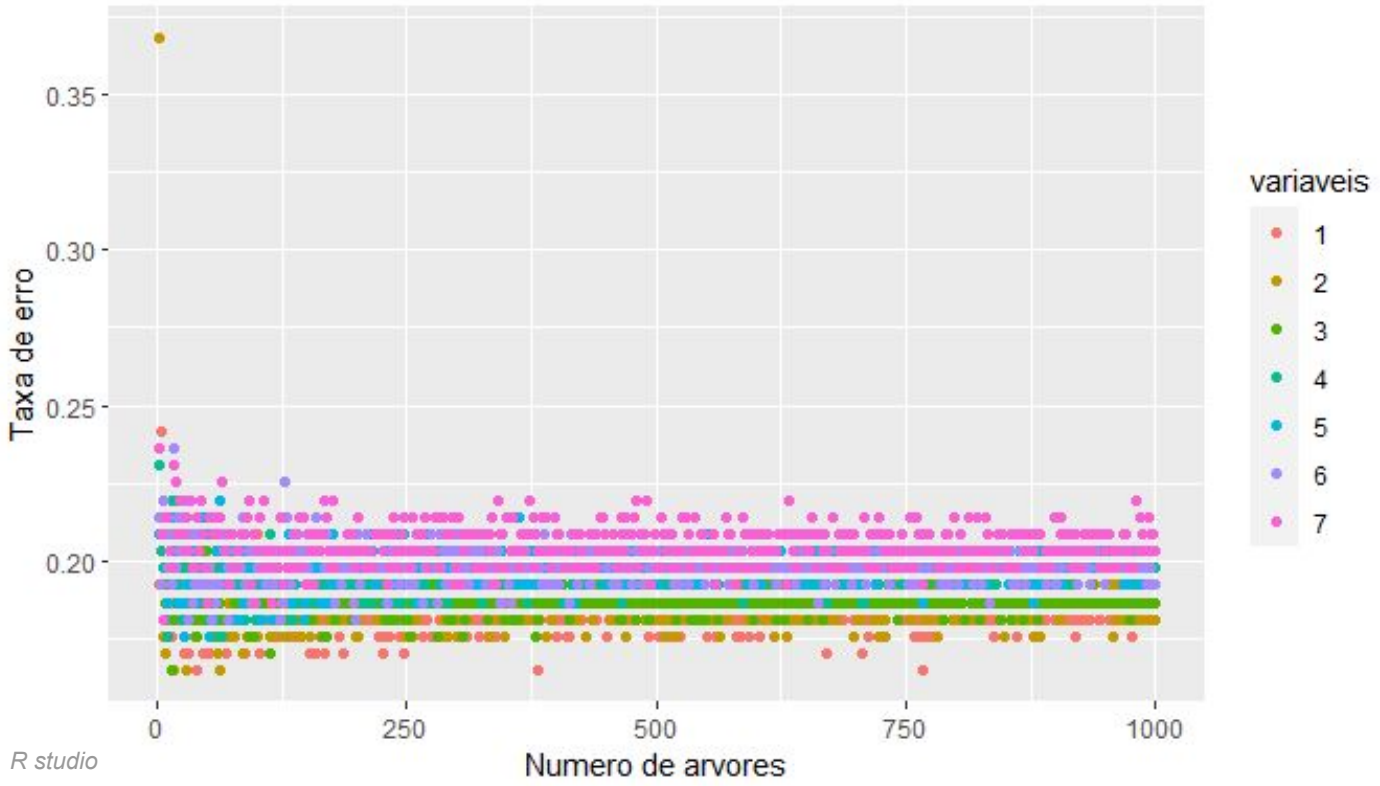


R studio



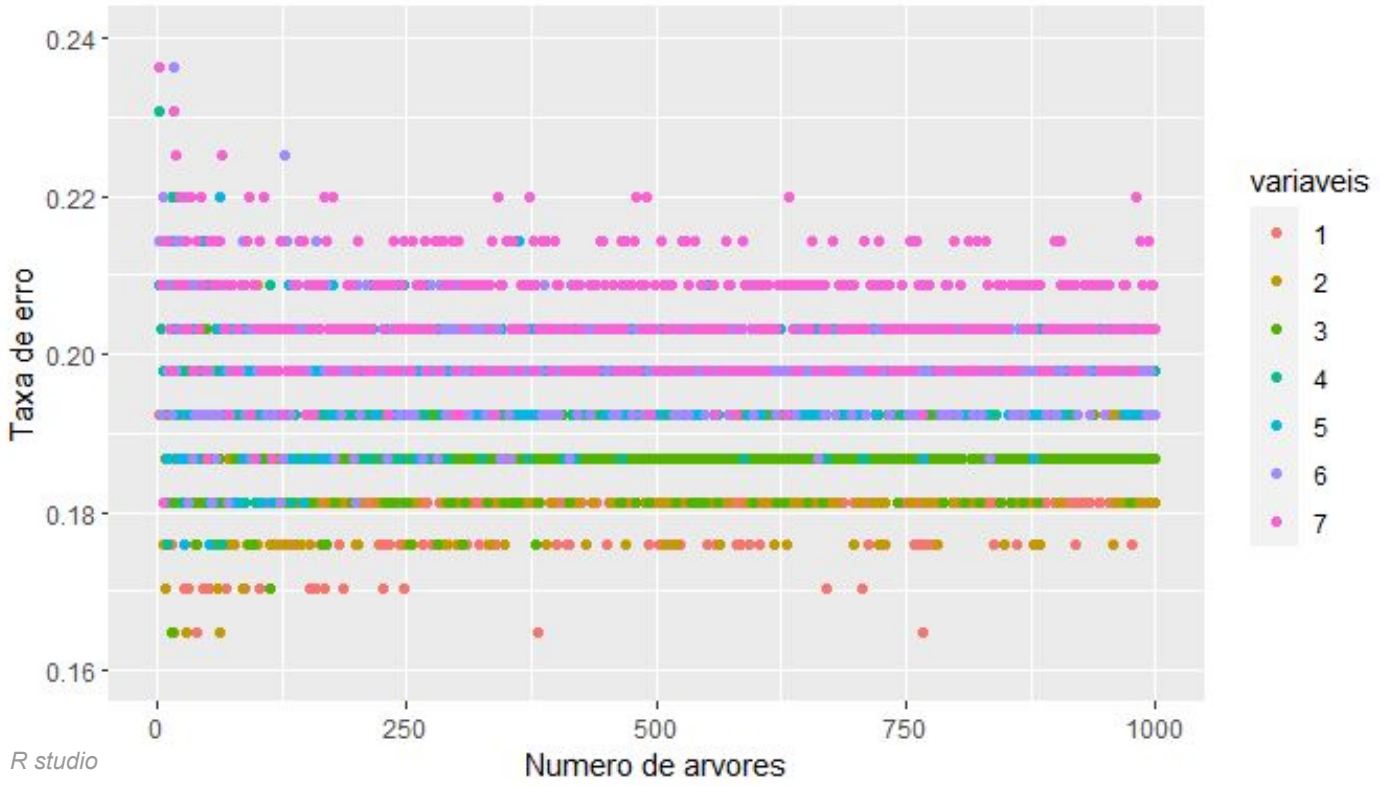
ISRLv2 - sessão 8.2.1

Experimentos com florestas: Floresta



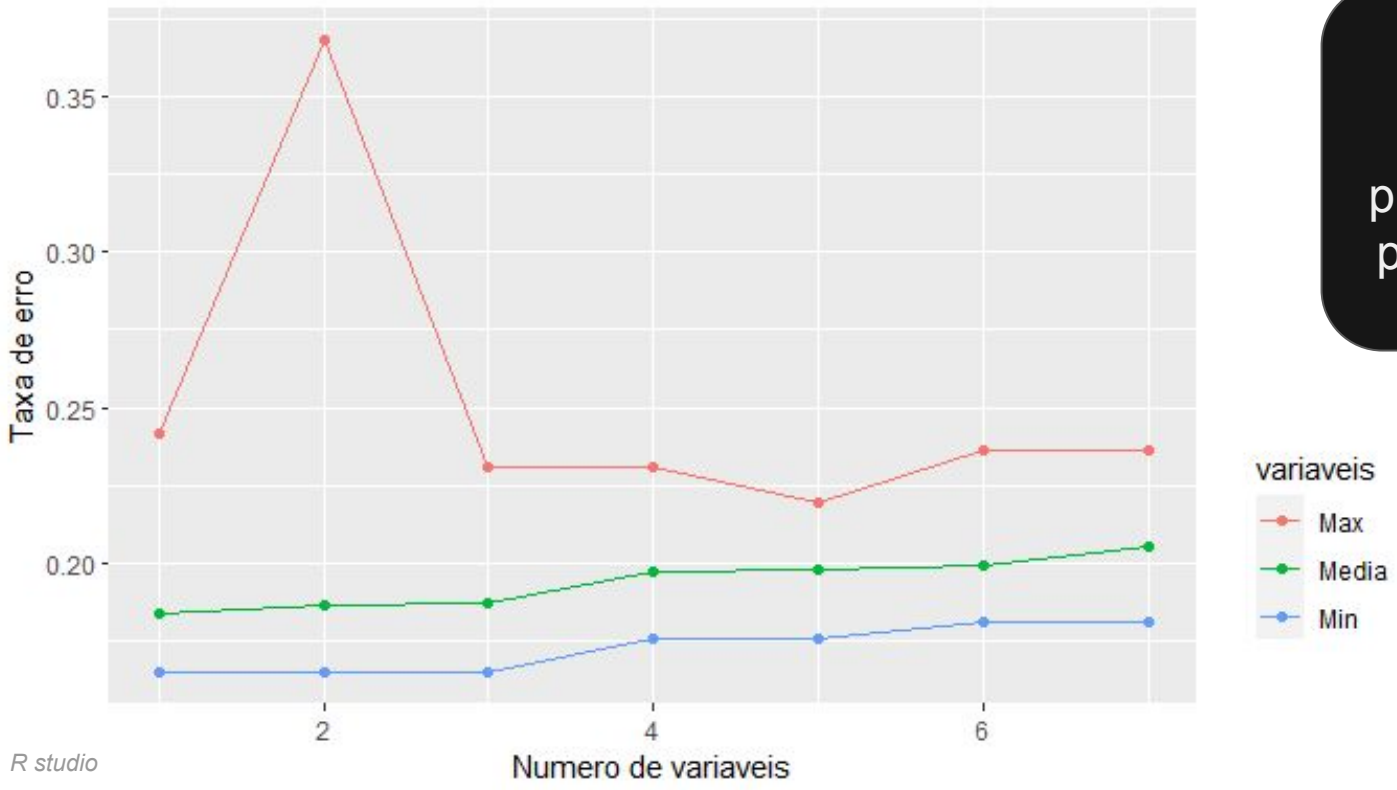
R studio

Experimentos com florestas: Acerto x Variáveis



R studio

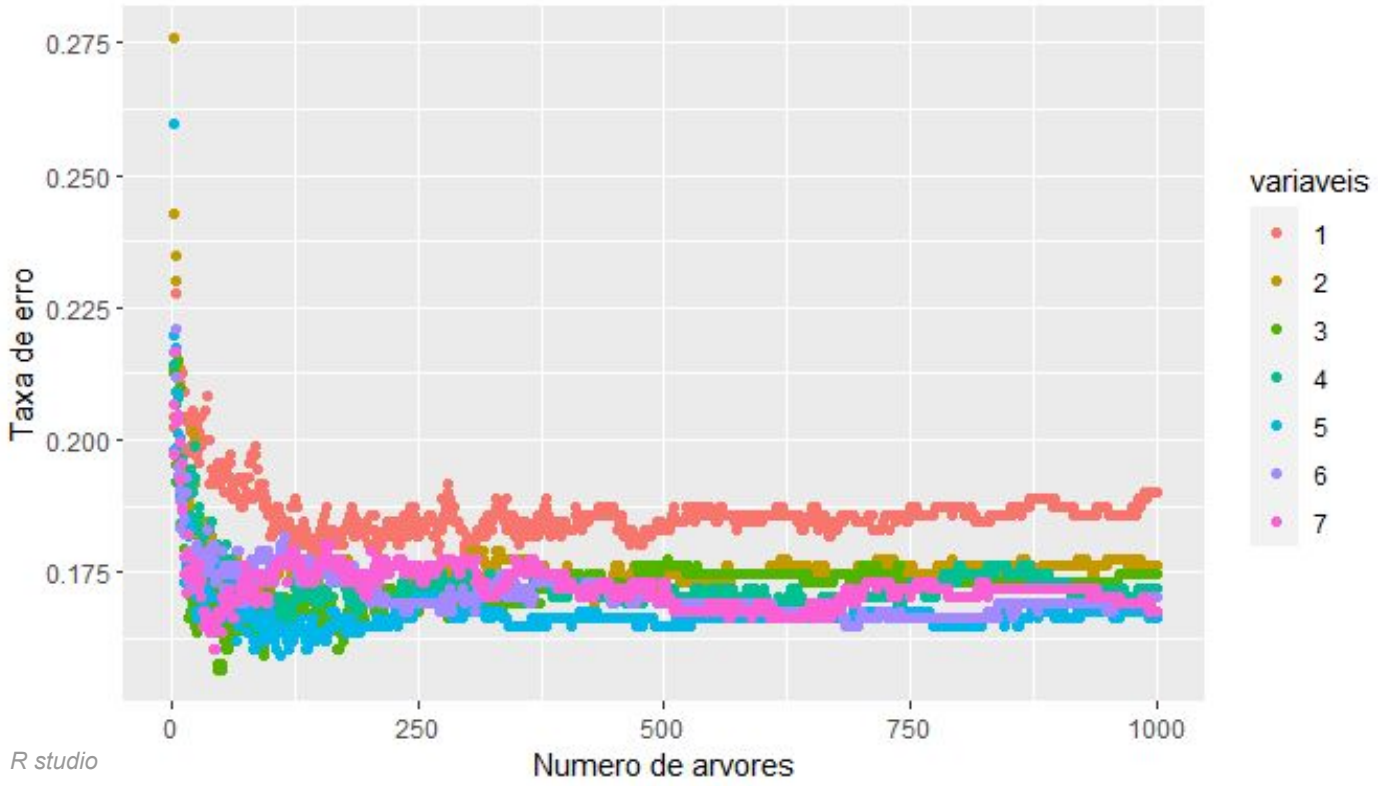
Experimentos com florestas: Acerto x Variáveis



Alternativa: repetir todo o processo anterior para o erro OOB

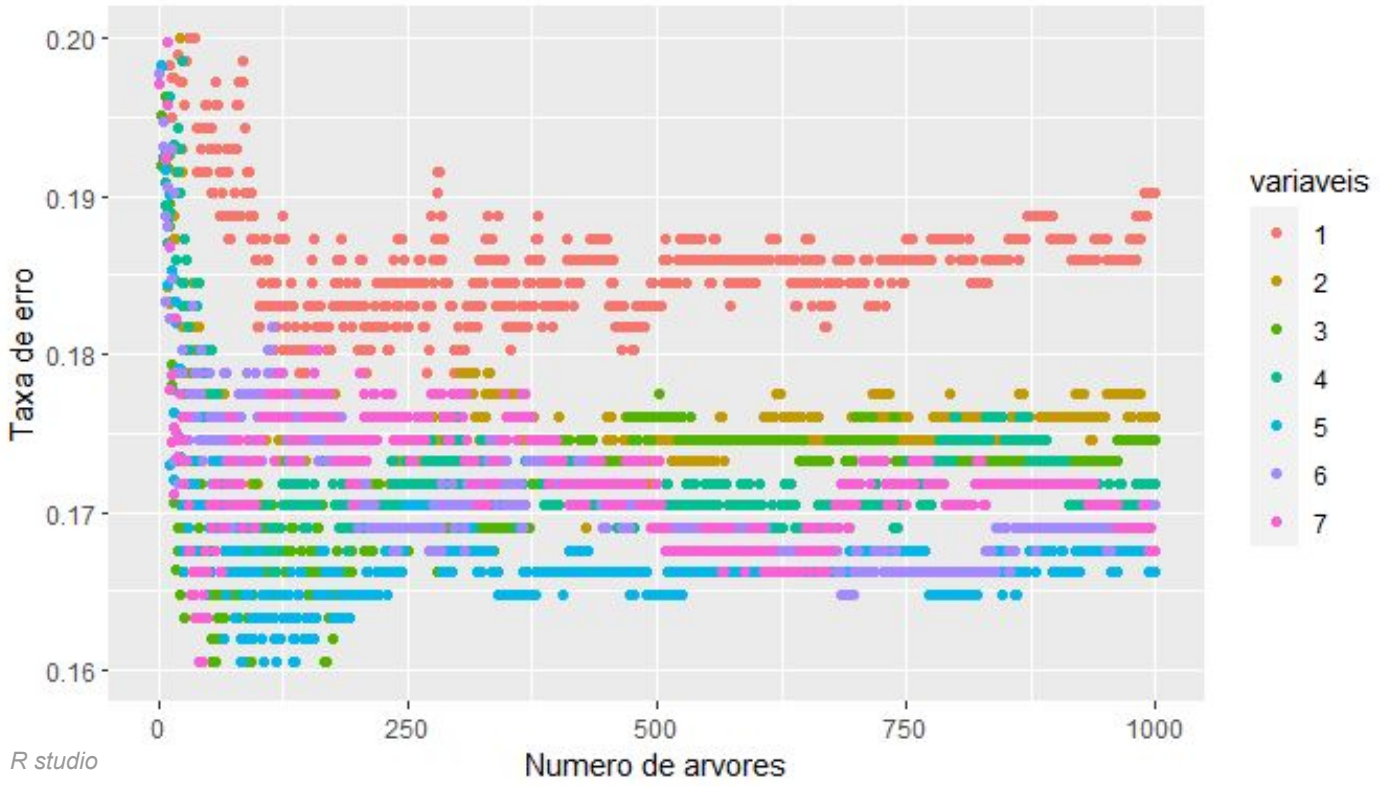
R studio

Experimentos com florestas: OOB x Variáveis



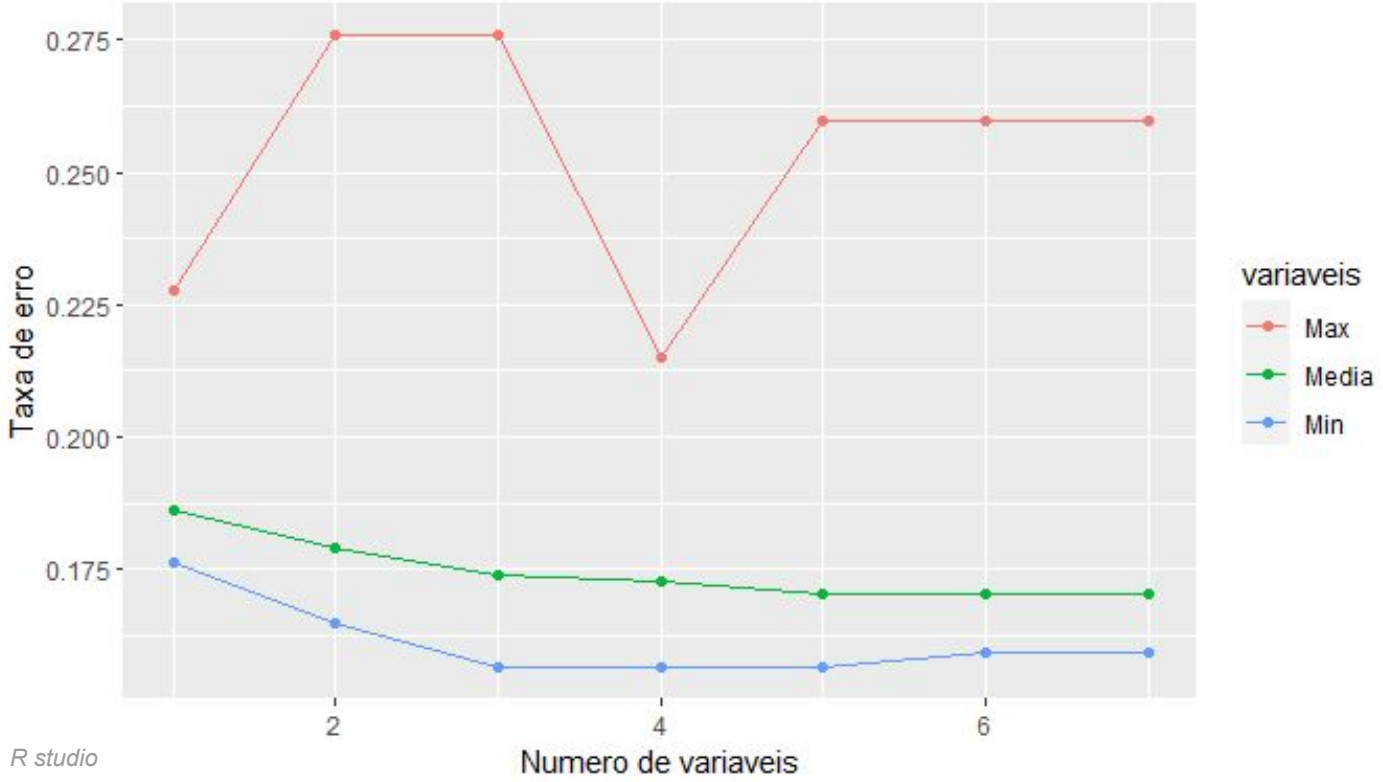
R studio

Experimentos com florestas: OOB x Variáveis



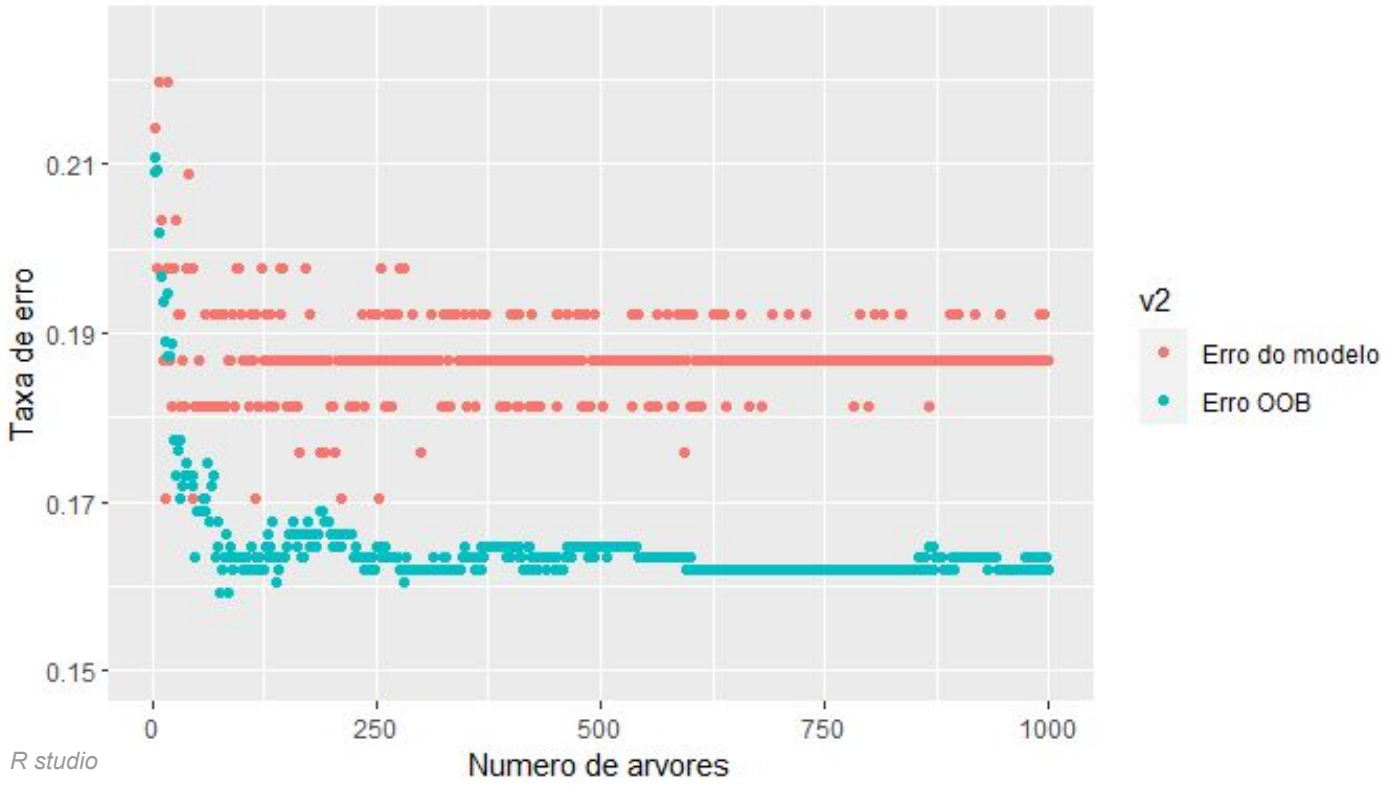
R studio

Experimentos com florestas: OOB x Variáveis



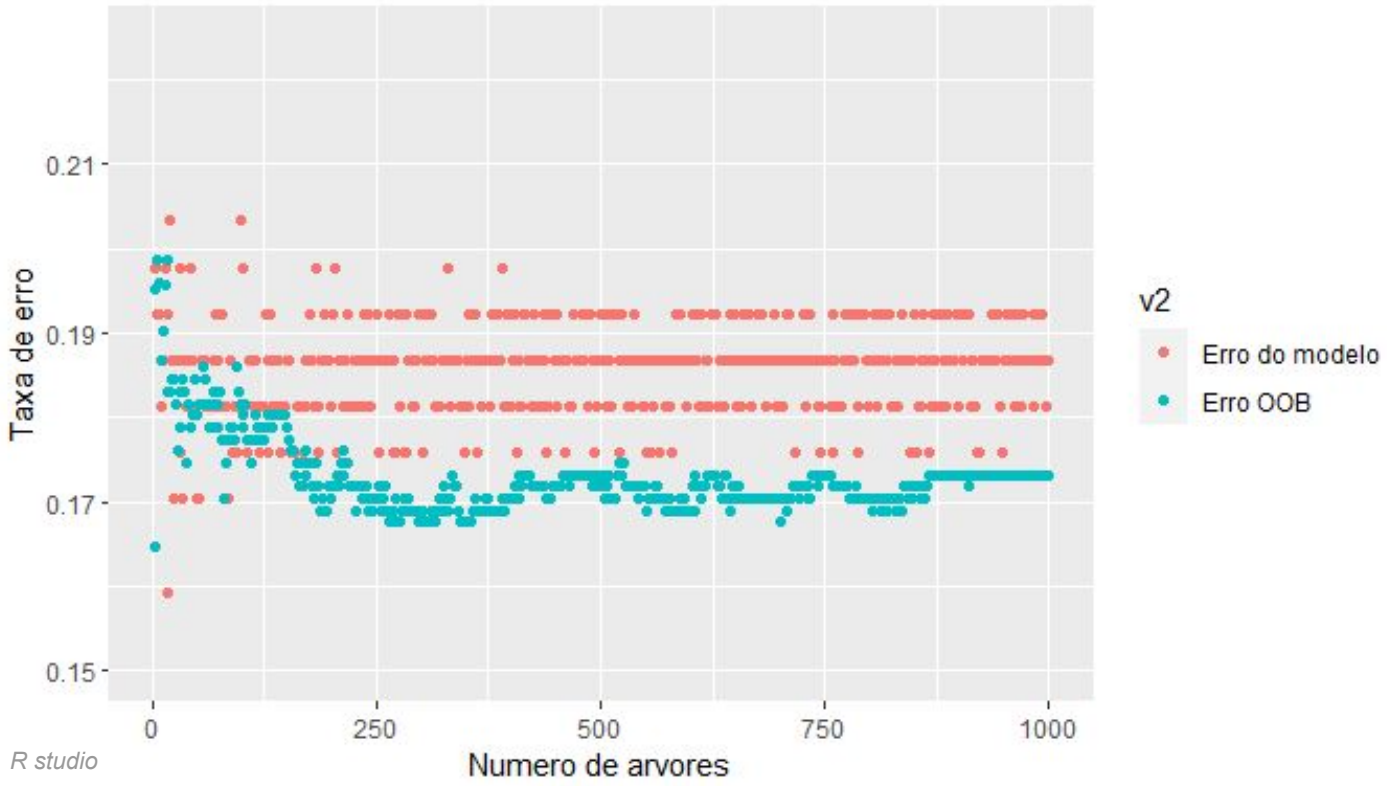
R studio

Experimentos com florestas: Floresta - 3 variáveis



R studio

Experimentos com florestas: Floresta - 2 variáveis



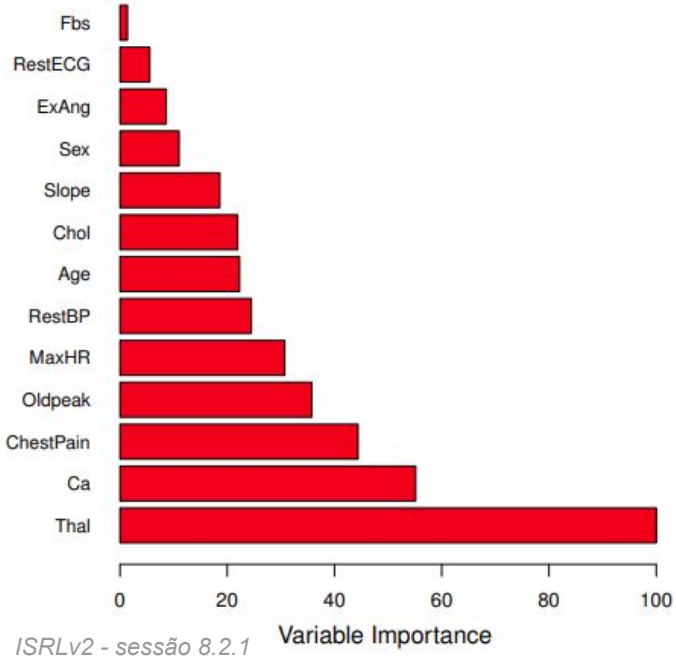
R studio

Importância de Variável:

A importância de uma variável é dada pela soma do ganho de pureza ((redução do gini)) de todas partições realizadas com a variável sobre a quantidade total de partições

$$fi_i = \frac{\sum_{j:node\ j\ splits\ on\ feature\ i} ni_j}{\sum_{k \in all\ nodes} ni_k}$$

[Towards Data Science]



ISRLv2 - sessão 8.2.1

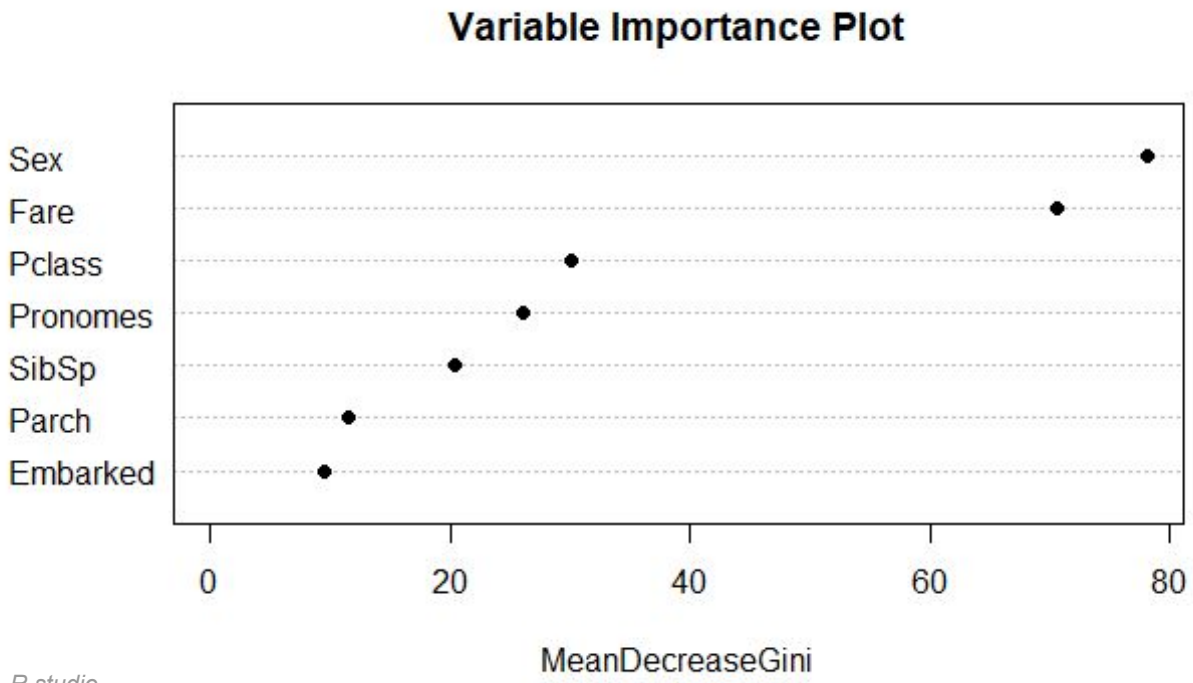
Importância de Variável: método alternativo

Calculation : How Variable Importance works

1. For each tree grown in a random forest, calculate number of votes for the correct class in out-of-bag data.
2. Now perform random permutation of a predictor's values (let's say variable-k) in the oob data and then check the number of votes for correct class. By "random permutation of a predictor's values", it means changing the order of values (shuffling).
3. Subtract the number of votes for the correct class in the variable-k-permuted data from the number of votes for the correct class in the original oob data.
4. The average of this number over all trees in the forest is the raw importance score for variable k. The score is normalized by taking the standard deviation.
5. Variables having large values for this score are ranked as more important. It is because if building a current model without original values of a variable gives worse prediction, it means the variable is important.

Importância de Variável:

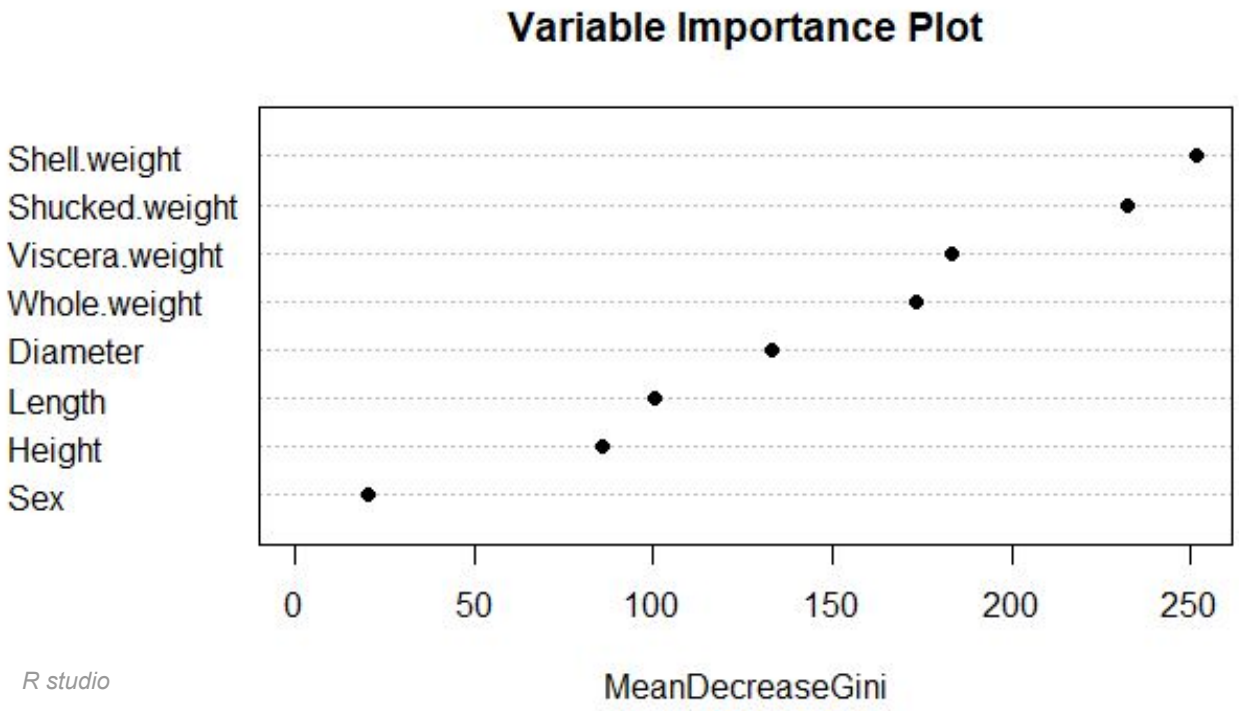
Conjunto Titanic
3 variáveis
500 árvores



R studio

Importância de Variável: Outros exemplos

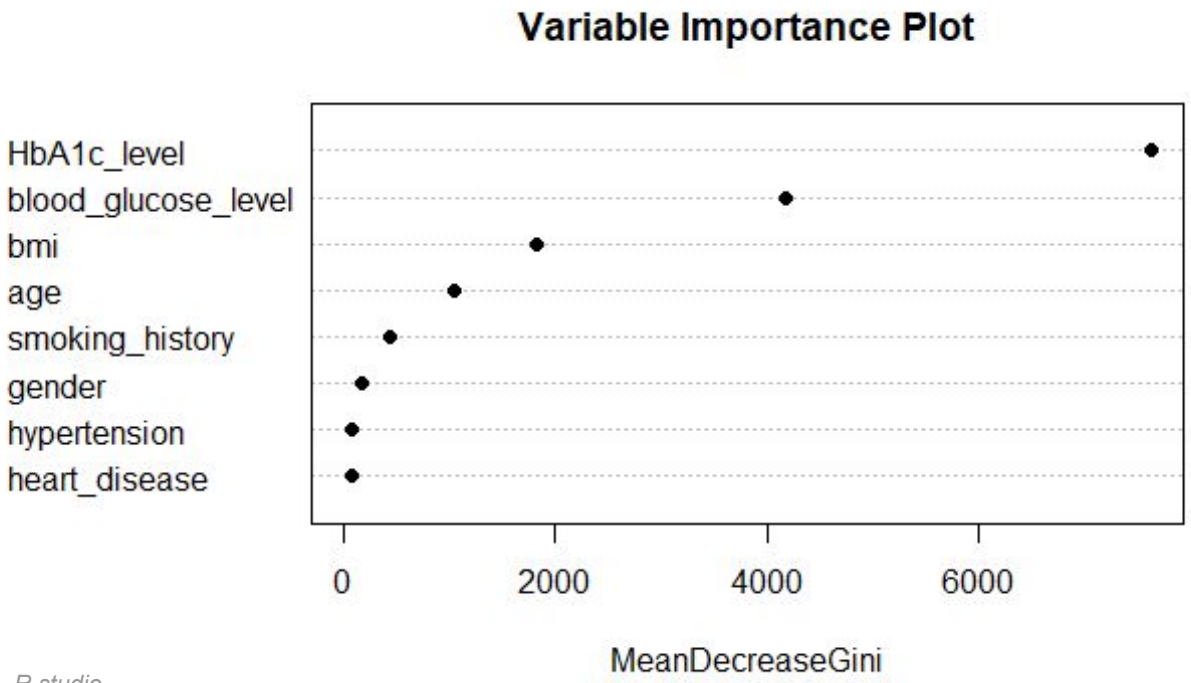
Conjunto Abalone
Todas as variáveis
((bagging))
500 árvores



R studio

Importância de Variável: Outros exemplos

Conjunto Diabetes
Todas as variáveis
((bagging))
500 árvores



R studio

FIM.