EPI5717: Machine learning para predições em saúde

Aula 11

Prof. Dr. Alexandre Chiavegatto Filho

PRÉ-PROCESSAMENTO DOS DADOS

1 - Preditores plausíveis:

- Pré-selecionar variáveis que sejam preditoras plausíveis (bom senso do pesquisador).

- Coincidências acontecem em análises de big data e pode ser que o algoritmo dê muita importância para associações espúrias.

PRÉ-PROCESSAMENTO DOS DADOS

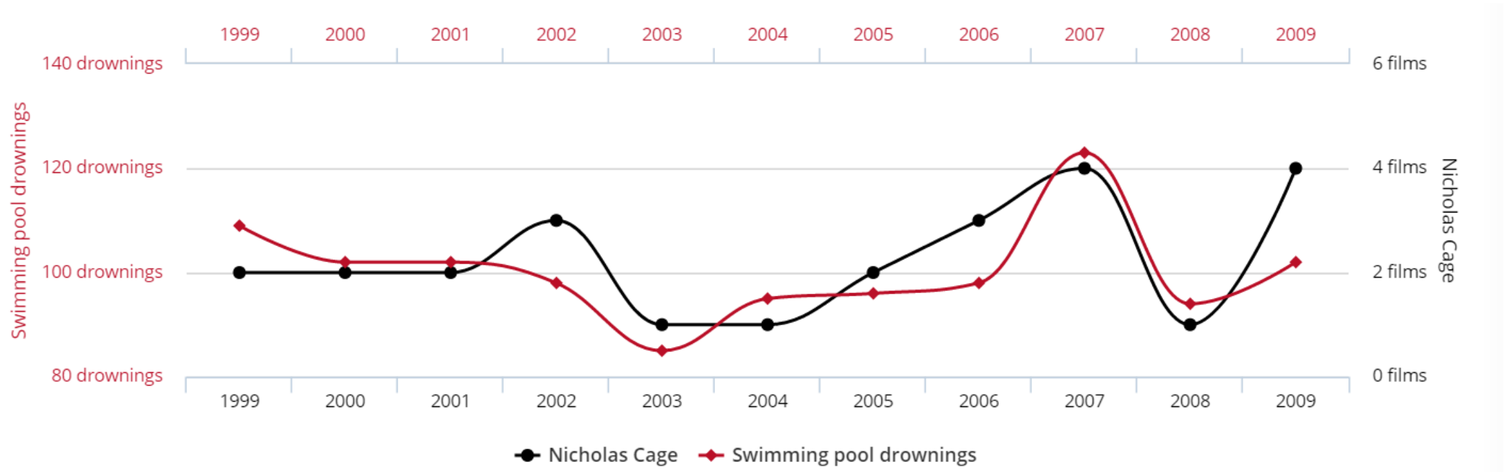
tylervigen.com

Number of people who drowned by falling into a pool

correlates with

Filmas Nicolas Cage appeared in

Correlation: 66,6% (r=0,666004)



PRÉ-PROCESSAMENTO DOS DADOS

Entretanto não precisam ser causais.

- Exemplo: na predição de risco de uma pessoa ir a óbito talvez seja interessante incluir o fato de ela ter sido internada em UTI recentemente.

- O fato de ela ter ido para a UTI **não é a causa** de ela ir a óbito no futuro, é apenas um preditor (ninguém cogita proibir UTI para diminuir óbito).

Preditores não precisam causar o desfecho, apenas consistentemente predizê-lo (podem ter mediadores, serem uma proxy...).

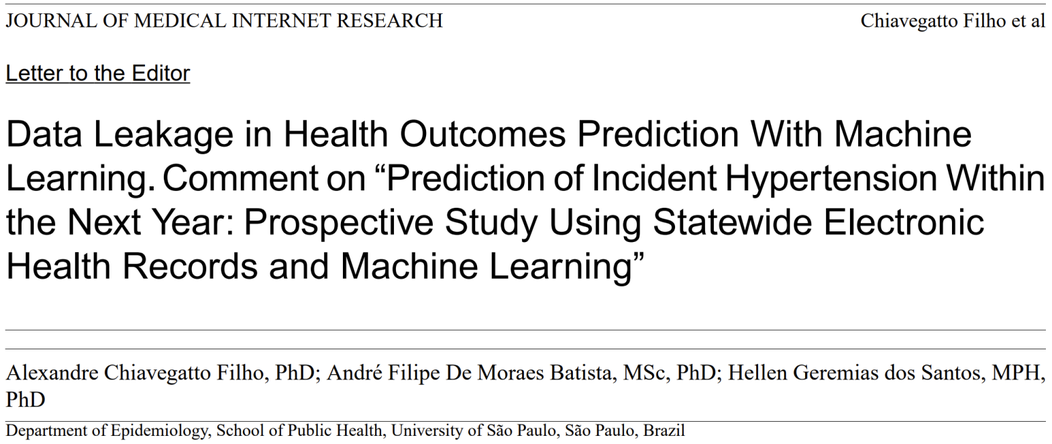
PRÉ-PROCESSAMENTO DOS DADOS

2 - Cuidado com vazamento de informação (“data leakage”).

- Acontece quando os dados apresentam informação escondida que faz com que o modelo aprenda padrões que não são do seu interesse.

- Uma variável preditora tem escondida o resultado certo:

  - Não é a variável que está predizendo o desfecho, mas o desfecho que está predizendo ela.



JOURNAL OF MEDICAL INTERNET RESEARCH Chiavegatto Filho et al

Letter to the Editor

Data Leakage in Health Outcomes Prediction With Machine Learning. Comment on “Prediction of Incident Hypertension Within the Next Year: Prospective Study Using Statewide Electronic Health Records and Machine Learning”

Alexandre Chiavegatto Filho, PhD; André Filipe De Moraes Batista, MSc, PhD; Hellen Geremias dos Santos, MPH, PhD

Department of Epidemiology, School of Public Health, University of São Paulo, São Paulo, Brazil

PRÉ-PROCESSAMENTO DOS DADOS

Exemplo

Incluir o número identificador do paciente como variável preditora

PRÉ-PROCESSAMENTO DOS DADOS

Exemplo

Incluir o número identificador do paciente como variável preditora

Problema

Se pacientes de hospital especializado em câncer tiverem números semelhantes.
Se o objetivo for predizer câncer, algoritmo irá dar maior probabilidade a esses pacientes.
Esse algoritmo aprendeu algo interessante para o sistema de saúde?

PRÉ-PROCESSAMENTO DOS DADOS

Exemplo	Problema	Motivo
Incluir o número identificador do paciente como variável preditora	Se pacientes de hospital especializado em câncer tiverem números semelhantes. Se o objetivo for predizer câncer, algoritmo irá dar maior probabilidade a esses pacientes. Esse algoritmo aprendeu algo interessante para o sistema de saúde?	Motivo pelo qual os dados e os algoritmos de machine learning precisam ser abertos. Sempre analisar importância preditora das variáveis (Shapley).

VARIÁVEIS COLINEARES

Uma forma de diminuir a presença de variáveis com alta correlação é excluí-las.

- Variáveis colineares trazem informação redundante (tempo perdido).
- Além disso, aumentam a instabilidade dos modelos.
- Estabelecer um limite de correlação com alguma outra variável (0,75 a 0,90).

PRÉ-PROCESSAMENTO DOS DADOS

Após seleção de preditores plausíveis e retirada de possíveis casos de vazamento de informação:

- Muitas vezes é importante reduzir o número de variáveis para:
  - Facilitar a coleta dos dados (principalmente para a utilização dos algoritmos em regiões remotas, com menor estrutura).
  - Diminuir o tempo de treinamento dos algoritmos.
  - Fazer uma segunda triagem de preditores plausíveis (diminuir o ruído).

EVALUATION OF VARIABLE SELECTION METHODS FOR RANDOM FORESTS AND OMICS DATA SETS

Authors: Frauke Degenhardt, Stephan Seifert and Silke Szymczak.

Resumo

Comparação das abordagens de seleção de variáveis para alta dimensionalidade através de simulações utilizando o Random Forest (RF):

- **Boruta** : 1° Melhor/só ele pode ser usado com baixa dimensionalidade
- **Vita** : 2° Melhor/Mais rápido
- Importância variável relativa recorrente
- Permutação
- Altmann
- Eliminação recursiva de variáveis

Introdução

- Omics
- Nº de variáveis maior do que o N° de indivíduos
- Apenas um pequeno conjunto de variáveis está associado ao desfecho
- Padrões complexos de correlações
- RF pode ser utilizado para classificação ou regressão
- RF utilizado em vários estudos de omics
- Conjunto reduzido de variáveis com + informação e - ruído

Introdução

- Seleção via importância preditiva nem sempre é ideal (é o ranking usando todas as varíaveis, pode mudar em outras configurações).
- Simulação 1: estrutura de correlação simples
- Simulação 2: padrões mais complexos
- Experimento: metilação e expressão genética (Classificação)
- Avaliação: performance preditiva e estabilidade da seleção de variáveis
- Nas simulações também são avaliadas: sensibilidade, poder empírico e taxa de falsos positivos

Random Forest

- Cada árvore treinada com uma amostra de bootstrap
- Em cada nó são selecionados subconjuntos de variáveis
- Critérios de seleção:
- Classificação --> índice de Gini
- Regressão --> Redução de variância
- Predição Global = voto majoritário ou média
- Importância de permutação: calculada como a diferença de performance preditiva antes e depois de permutar os valores da variável em todas as árvores.

Métodos de Seleção de Variáveis

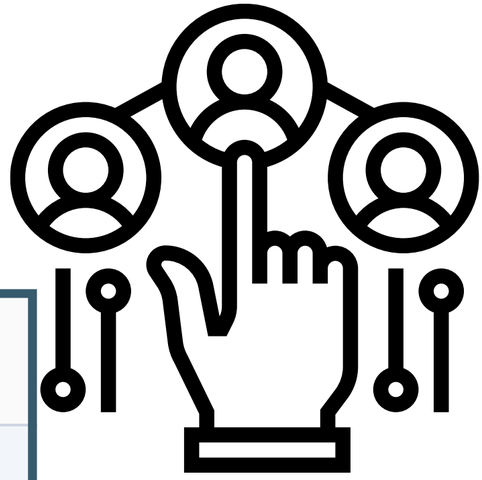


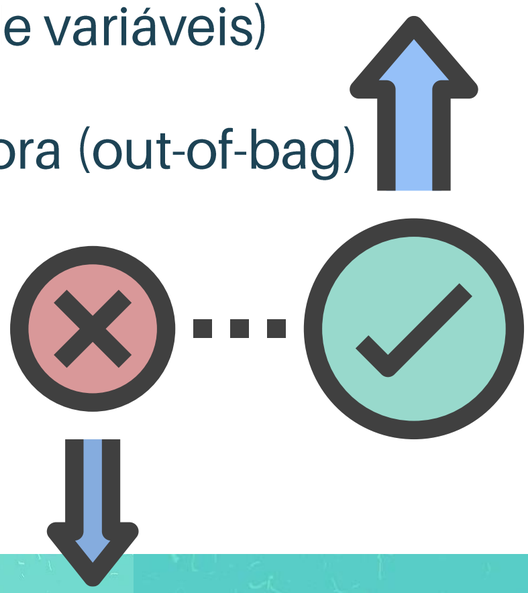
Table 1

Information about the different variable selection approaches that are compared

Abbreviation	Name	Goal	Approach
Altmann	Altmann	All relevant variables	Permutation of outcome; parametric *P*-value
Boruta	Boruta	All relevant variables	Importance significantly larger than those of shadow variables
Perm	Permutation	All relevant variables	Permutations of outcome; nonparametric *P*-value
r2VIM	Recurrent relative variable importance	All relevant variables	Relative importance based on minimal observed importance; several runs of RF
RFE	Recursive feature elimination	Minimal set	RF with smallest error based on iterative removal of least important variables
Vita	Vita	All relevant variables	*P*-values based on empirical null distribution based on non-positive importance scores calculated using hold-out approach

Eliminação Recursiva de Variáveis

- Objetivo: encontrar um conjunto mínimo de variáveis
- A redução das variáveis é realizada de forma recursiva (retirada das variáveis menos importantes):
  - Até restar apenas uma variável (ranking recalculado a cada novo conjunto de variáveis)
- A performance preditiva é sempre mensurada nas observações que ficam de fora (out-of-bag)
- Melhor conjunto de variáveis será o que proporcionar um menor erro
- Método popular, muito utilizado com dados moleculares



Boruta

- Compara as importâncias das preditoras reais com “sombras”.
- A cada execução é gerada uma cópia de cada variável
- Valores das sombras são permutações das variáveis originais
- Random Forest é treinado com essa extensão dos dados
- Comparar a importância de variáveis reais com o valor máximo das sombras:
    - Variáveis só são relevantes se forem mais importantes que as permutadas.

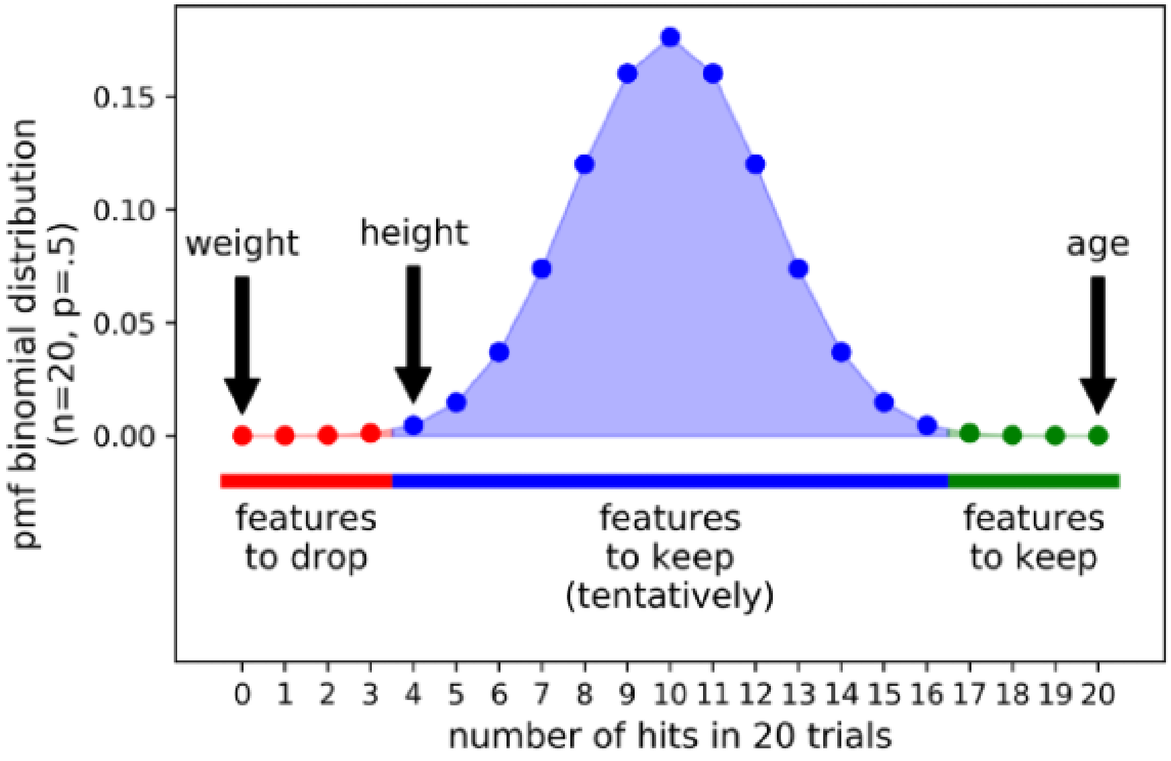
Boruta

Boruta explained the way I wish someone explained it to me

Looking under the hood of Boruta, one of the most effective feature selection algorithms

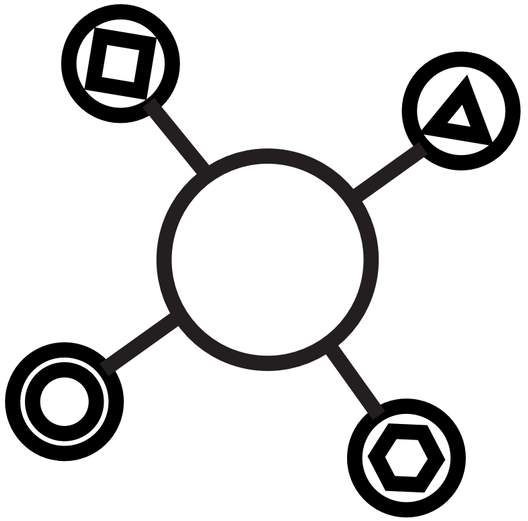
Medium / Samuele Mazzanti / 17 de mar. de 2020

- Repetir o processo várias vezes e transformar em uma distribuição binomial
- Hit: quando a variável é mais importante que a sombra mais importante.
- Sombras e variáveis menos importantes são retiradas e o processo é repetido até haver uma classificação para cada variável.



Binomial distribution and positioning of the features

Permutação



- Aplica-se um teste de permutação padrão
- Permuta valores do desfecho, ao contrário do Boruta
- Padrões de correlação ficam intocados
- Repetições das permutações
- Se a importância da variável no desfecho real for maior do que em todas as permutações, ela é selecionada como importante

Importância variável relativa recorrente

- Gera-se várias RF com o mesmo conjunto de dados e valores de parâmetros que diferem apenas na semente aleatória
- Cada RF é usada para calcular valores de importância preditora
- Esses são divididos pela importância mínima absoluta de cada execução
- São importantes aquelas com a importância maior ou igual a um valor especificado

Vita

- Semelhante ao anterior
- Não utiliza permutações
- Dados divididos em dois subconjuntos de mesmo tamanho
- Duas RF são treinadas usando cada um dos conjuntos
- Importância variável estimada com base no conjunto não utilizado no treino
- Importância final calculada pela média dos dois escores estimados por variável

Simulação 1

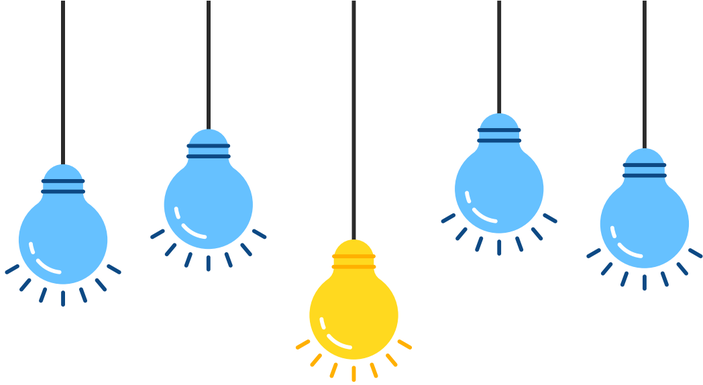
Table 2

Parameters used for RF and variable selection methods

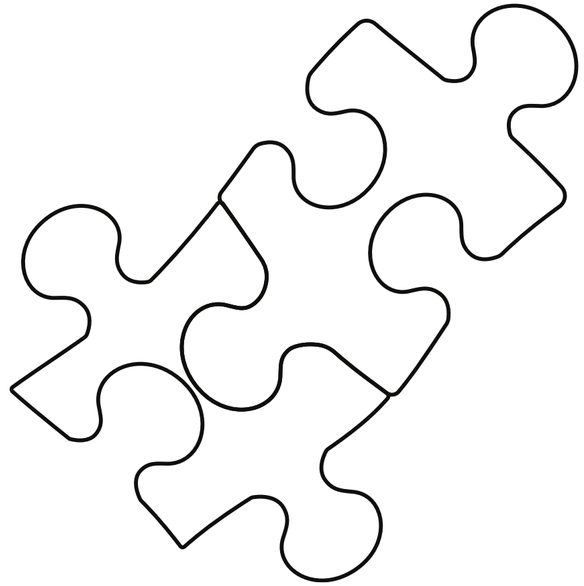
Approach	Parameter	Description	Value
RF	ntree	Number of trees	10 000
	mtry	Number of variables selected at each split	33% of number of variables
	nodesize	Minimal number of individuals in terminal node	10% of sample size
Altmann	no.perm	Number of permutations	50
	p.t	Threshold for *P*-values	0
Boruta[a]	pValue	Confidence level	0.01
Perm	no.perm	Number of permutations	500
	p.t	Threshold for *P*-values	0
r2VIM	no.runs	Number of RFs to be generated	20
	factor	Minimal relative importance score for a variable to be selected	3
RFE	prop.rm	Proportion of variables removed at each step	0.1
	tol	Acceptable difference in optimal performance (in %)	10
Vita	p.t	Threshold for *P*-values	0

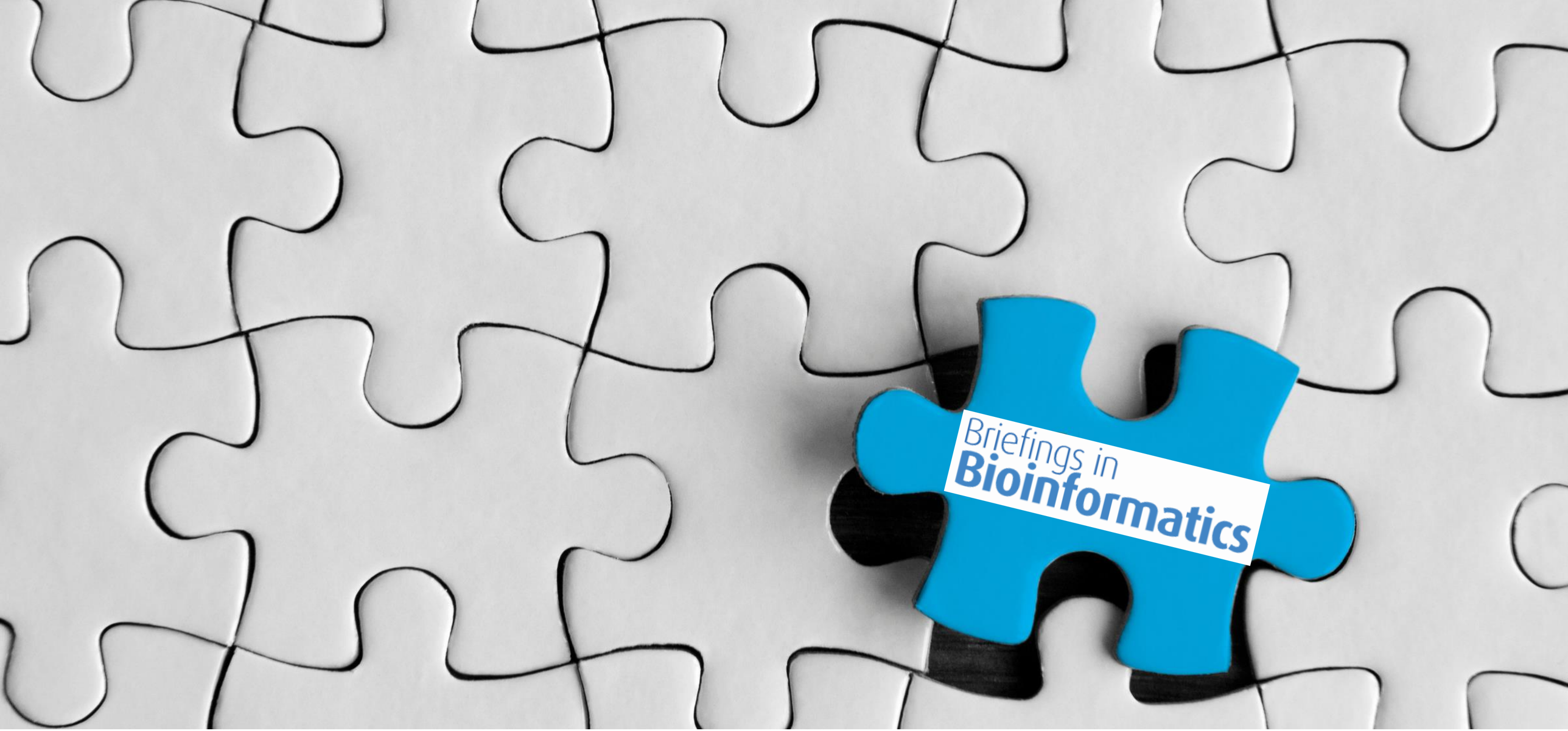
[a]Using default parameters (mtry = square root of number of variables; nodesize = 1 for classification, 5 for regression).

Discussão e Conclusão



- Ambos estudos de simulação identificaram o algoritmo Boruta como melhor
- Seguido do Método Vita
- Ambos ideais para trabalhar com grandes conjuntos de dados
- Boruta ideal também para trabalhar com baixa dimensionalidade
- Os resultados dos estudos de simulação são semelhantes tanto para as configurações de regressão quanto para classificação





Referência:

Degenhardt, F., Seifert, S., & Szymczak, S Evaluation of variable selection methods for random forests and omics data sets. *Briefings in bioinformatics*, 20(2), 492-503. (2019).