Machine Learning e Credit Scoring: um estudo de caso.

Cinthia Becker

SOBRE MIM

1995

2015 - 2018

Graduação em Estatística

UFRGS

2017 - 2019

Analista de

Riscos Financeiros

Getnet

2019

Pós Graduação em

Data Science e Big Data

UniRitter

2019

Analista de

Modelagem de Crédito

Realize / Lojas Renner S.A.

MOTIVAÇÃO

“É só fazer uma Regressão Logística e tá tudo resolvido”

“Hoje em dia tudo é Inteligência Artificial e Machine Learning”

CENÁRIO

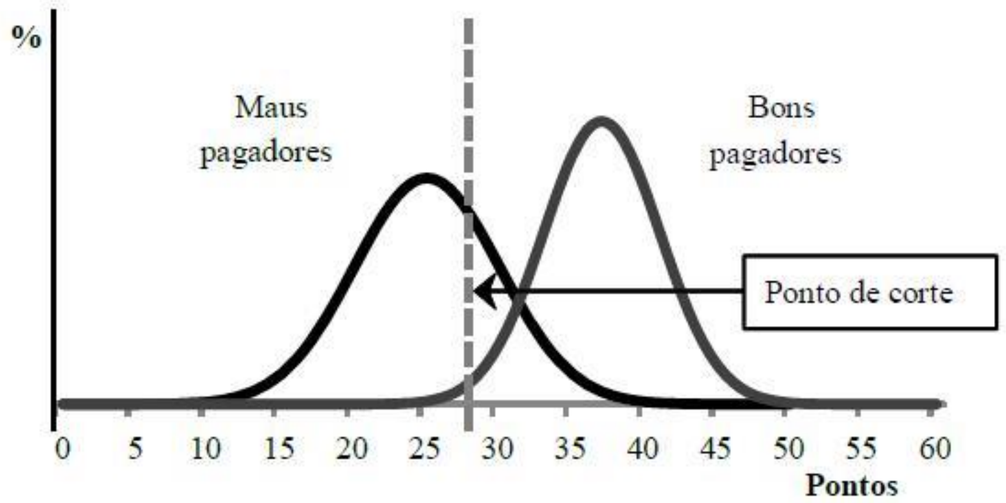
Concessão de crédito

Aprovados

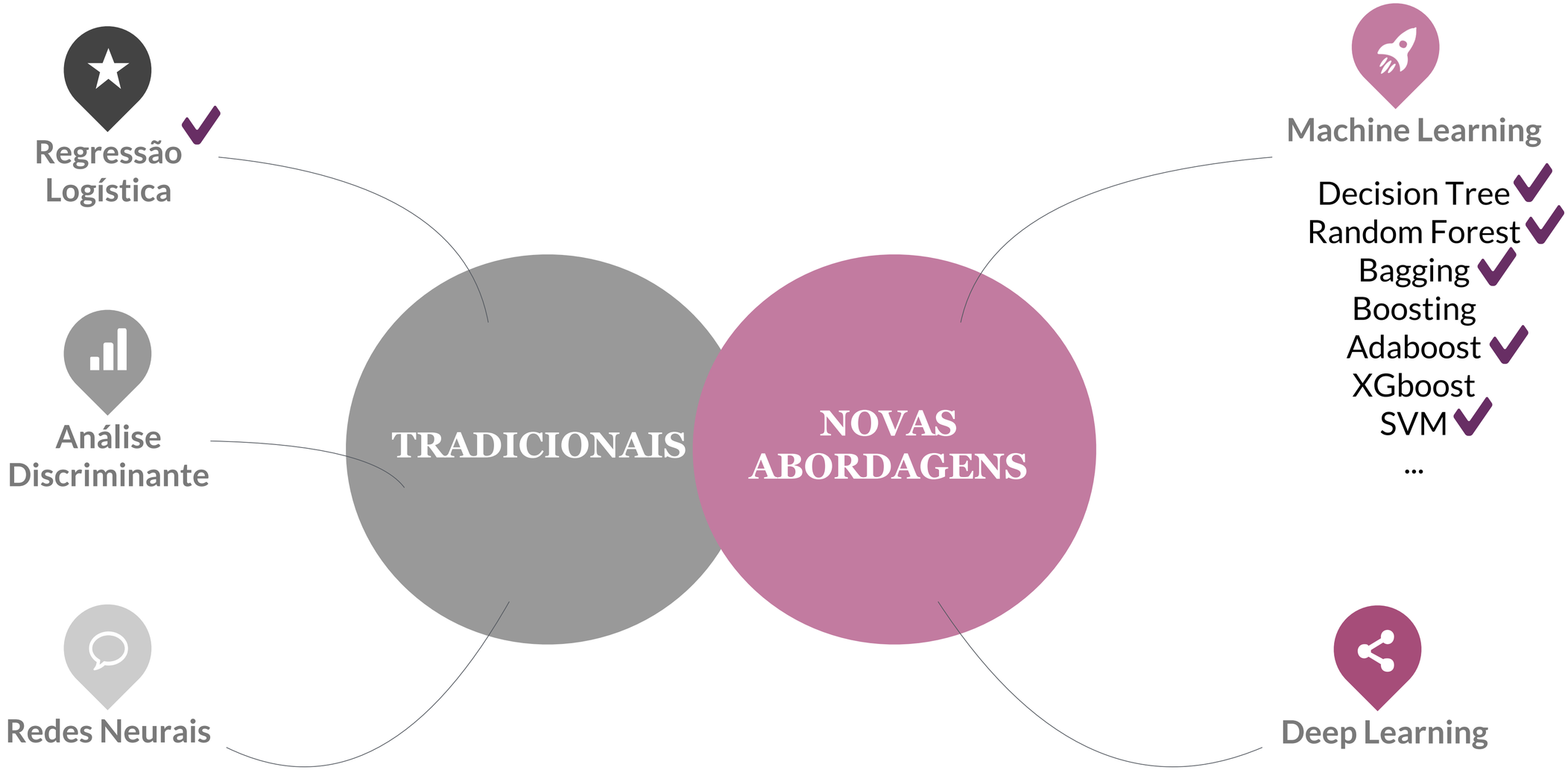
- Bons Clientes
- Maus Clientes

Reprovados

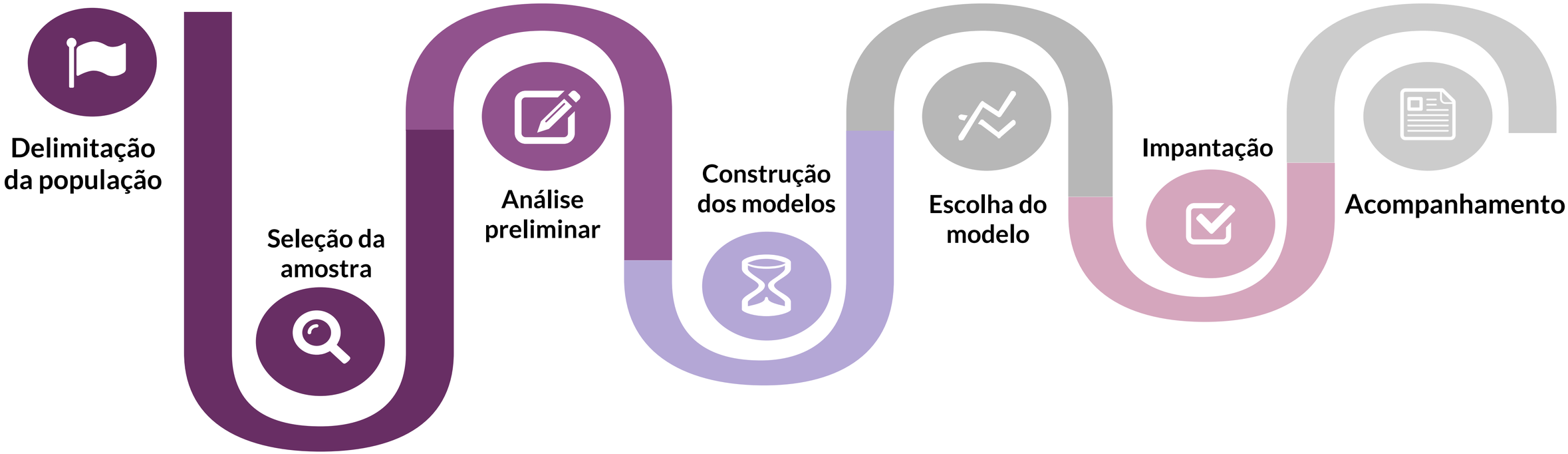
- Bons Clientes
- Maus Clientes



CENÁRIO



PROCESSO DE MODELAGEM



- Público alvo;
- Desempenho satisfatório;
- Horizonte de previsão.

- Variáveis disponíveis;
- Período e tamanho das amostras;
- Validação dos dados.

- Variáveis pro modelo;
- Agrupamento de atributos;
- Criação de *dummies*.

- Técnicas utilizadas;
- Software (R ou Py);❤
- Seleção da variável independente;
- Validação das suposições.

- Métricas:
  - KS;
  - ROC;
  - Acurácia.

- Preparação dos sistemas de informação;
- Determinação do ponto de corte.

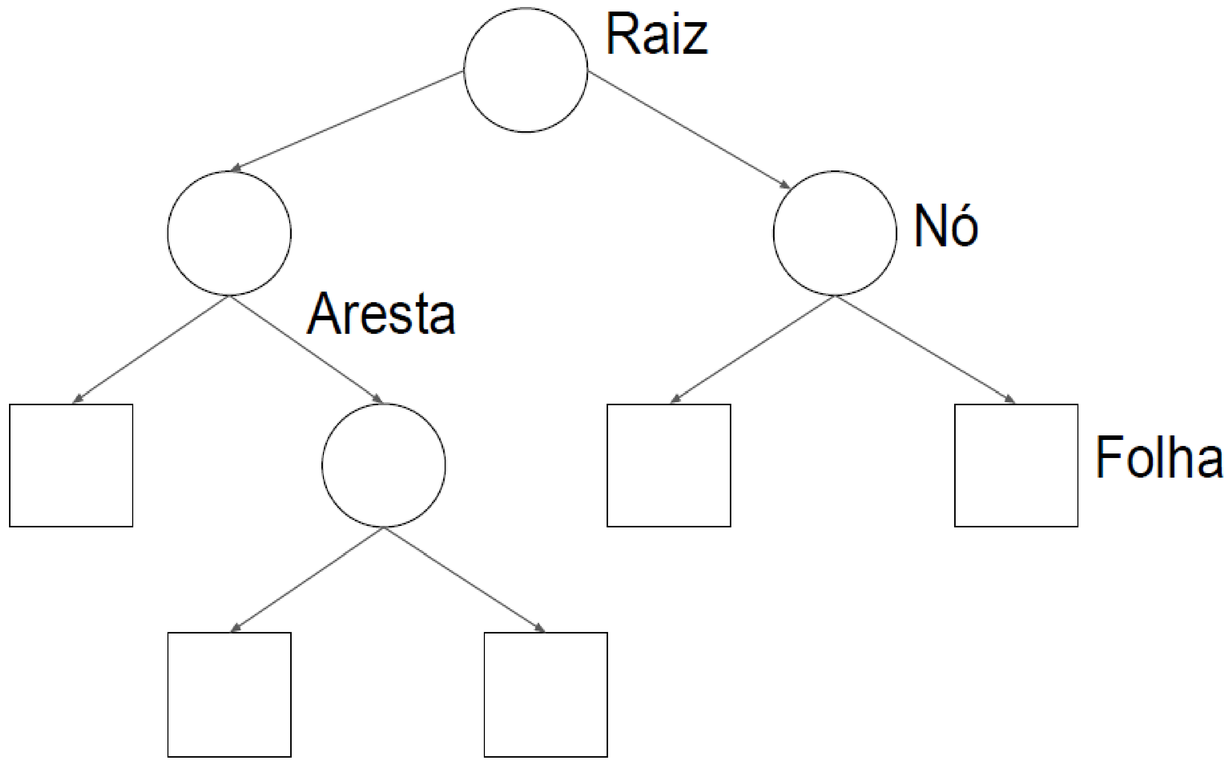
- Monitoramento do desempenho do modelo;
- Sinalização quando há necessidade de revisão.

TÉCNICAS UTILIZADAS

DECISION TREE

- Série de decisões lógicas.
- Os nós são ordenados por ganho de informação.
- Modelo para quando não há mais ganho de informação com a inclusão de uma nova variável.

<u>Desvantagem:</u>
Pode apresentar ***overffiting*** dos dados.



Fonte: REIS FILHO, 2006.

COMANDO NO R:
rpart()

TÉCNICAS UTILIZADAS

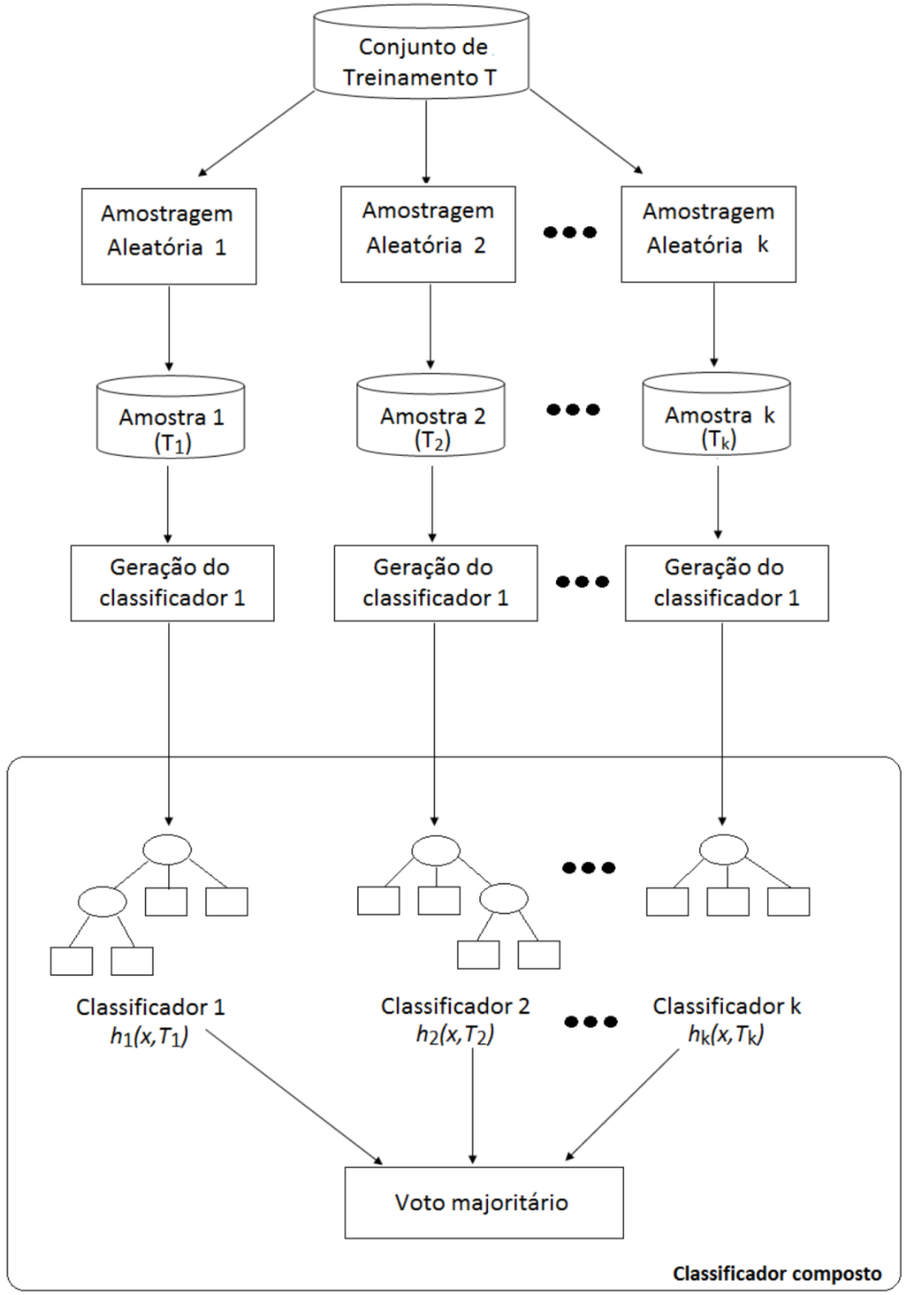
- Classificador ***Ensemble***:
  - Treinados de forma independente e com diferentes conjuntos de treinamento.

- A combinação de K funções de predição, resulta em um **estimador com variância menor**.

- Eficaz quando apresenta **classificadores instáveis**:
  - Pequenas mudanças no conjunto de treinamento podem causar grandes mudanças no classificador gerado.

BAGGING

COMANDO NO R:
bagging()



Fonte: OSHIRO, 2013.

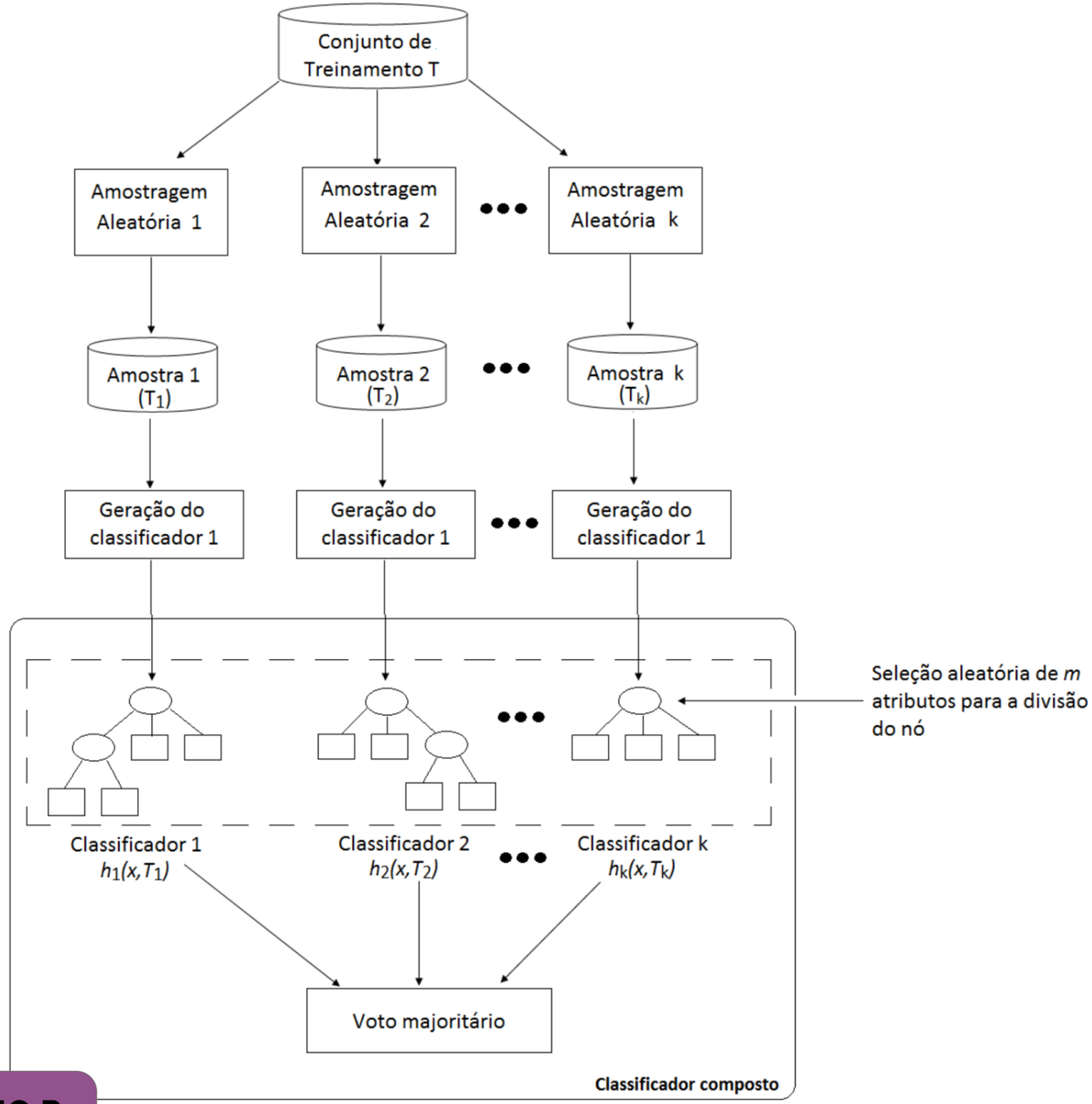
TÉCNICAS UTILIZADAS

- Trabalha muito bem em **grande conjuntos de dados:**
  - Não falha no que diz respeito a “maldição da dimensionalidade”.

- Utiliza uma **parte das variáveis** independentes disponíveis:
  - Seleção aleatória para a construção de cada árvore.

- A **aleatorização das covariáveis reduz correlação** entre as predições, **aumentando o viés** em troca de uma **diminuição da variância do estimador.**

RANDOM FOREST



COMANDO NO R:
randomForest()

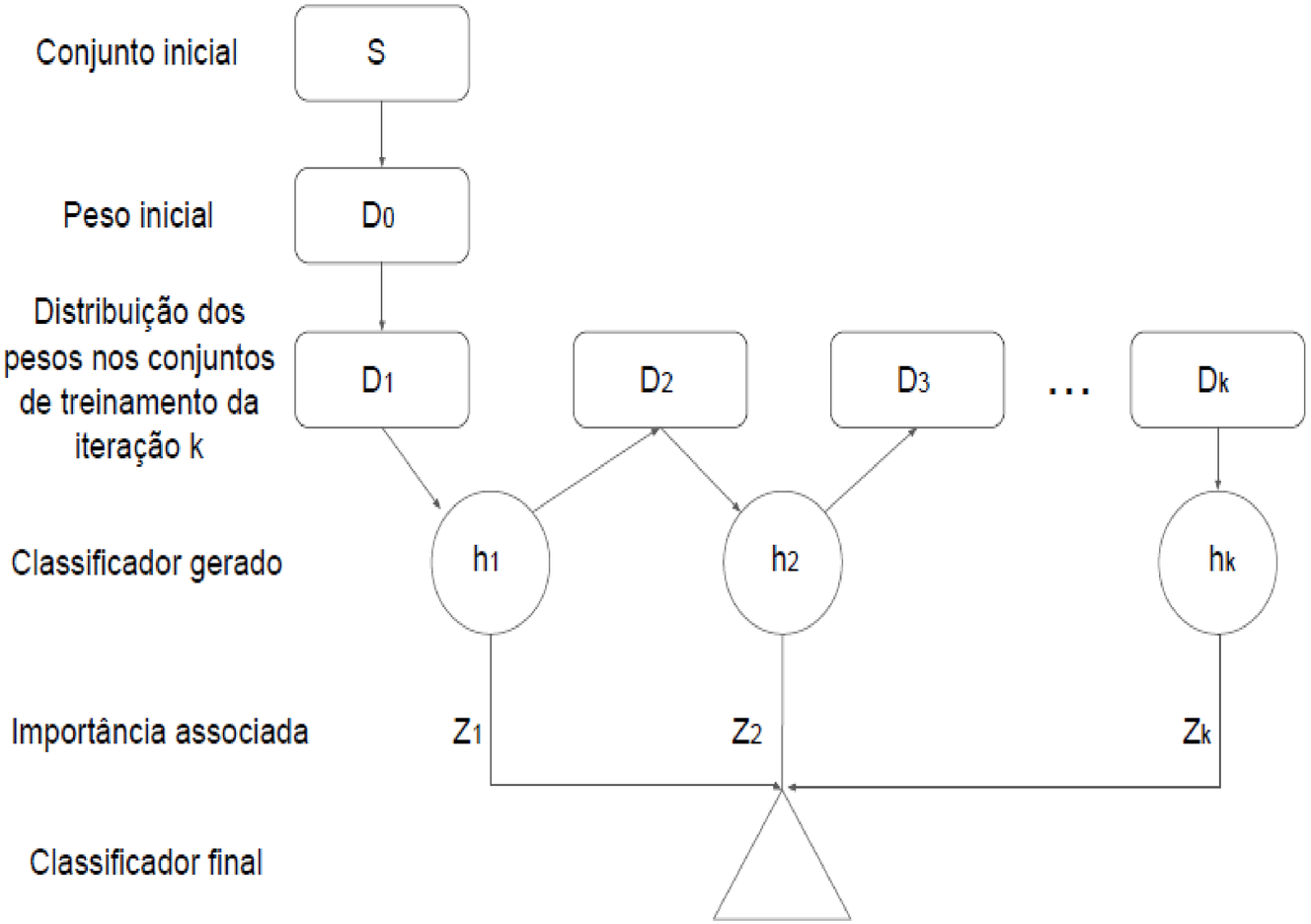
Fonte: OSHIRO, 2013.

TÉCNICAS UTILIZADAS

- **Boosting:**
  - Dados reamostrados construídos a fim de gerar **aprendizados complementares;**
  - A importância do voto é ponderada conforme o desempenho de cada modelo.

- O **voto final é ponderado** com base na importância associada a cada classificador.

ADABOOST



Fonte: Adaptado de CHAVES, 2012.

COMANDO NO R:
boosting()

TÉCNICAS UTILIZADAS

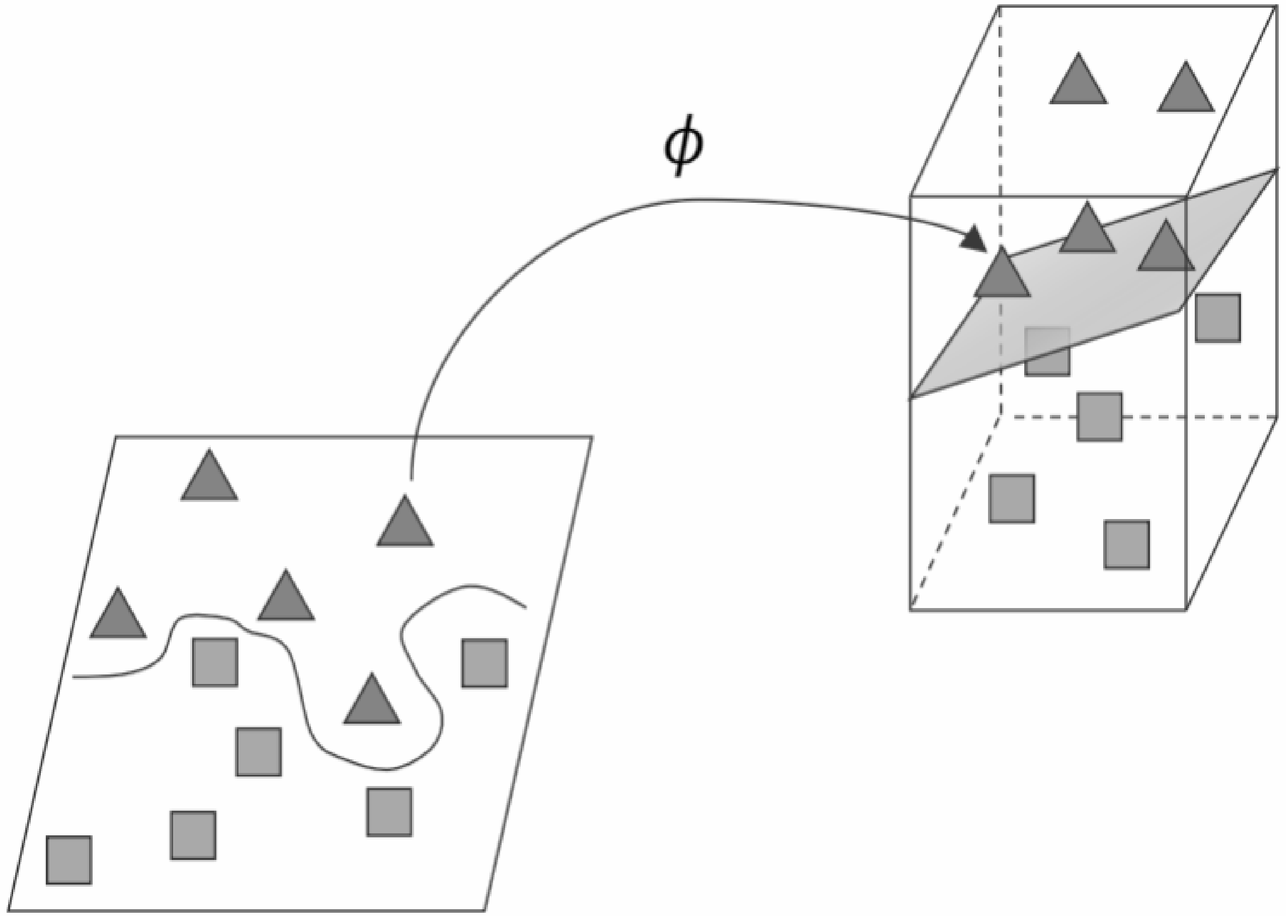
SUPPORT VECTOR MACHINE

- Objetivo:
  - Construção de um **hiperplano** para separar os dados;
  - Baseado na relação a **maior distância possível** entre os **vetores de suporte**.

- Capaz de **separar linearmente** dados não-lineares:
  - Função *kernel* (ø): mapeia os vetores para uma dimensão de ordem maior.

- Produz classificadores com uma **boa capacidade de predição** em dados não presentes na amostra de treinamento.

<u>Desvantagem:</u>
Estimação de parâmetros é **não probabilística**, sendo feita por meio de medidas de distâncias (lineares ou não lineares).



Fonte: BECKER, 2017.

COMANDO NO R:
svm()

APLICAÇÃO

BANCO DE DADOS REAL

de concessão de crédito à Pessoas Físicas.

Variável Resposta

Atraso superior a 60 dias

Variáveis Preditoras

Idade, Sexo, Escolaridade, Tipo de residência, Nº de filhos, Profissão, etc.

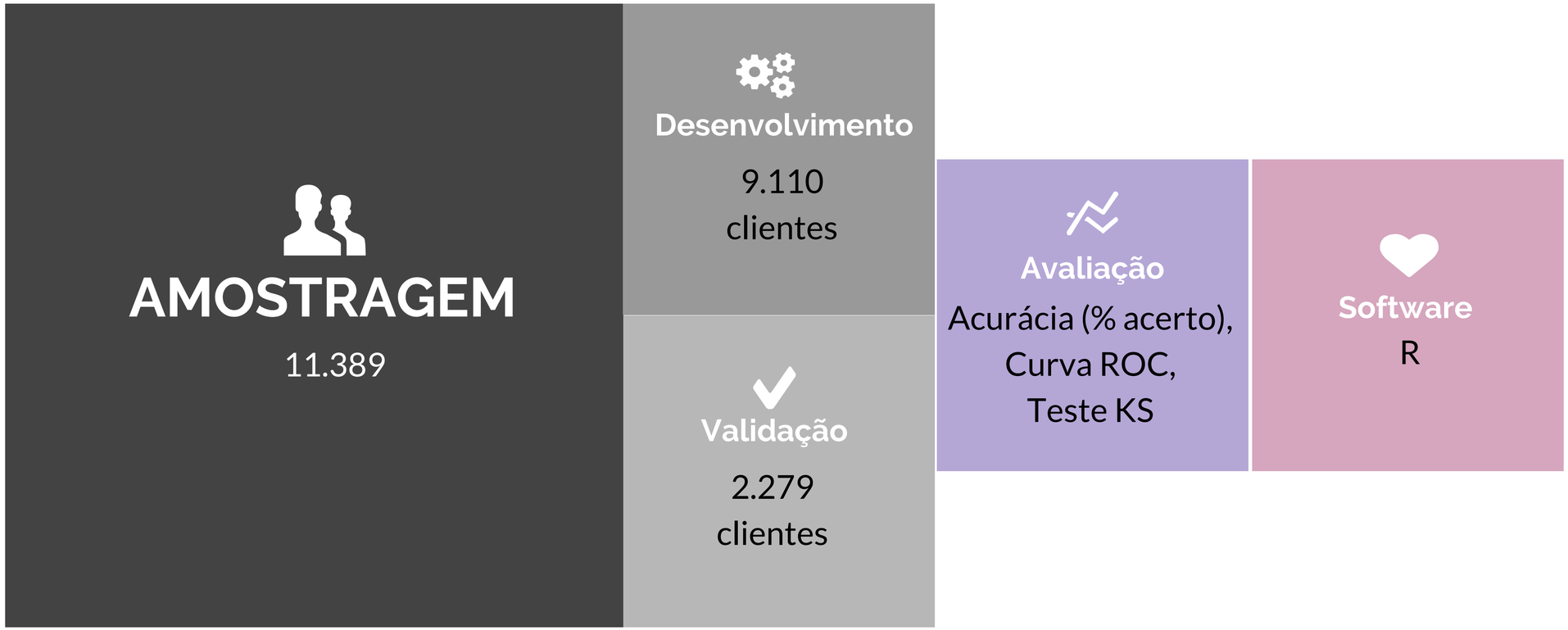
Agrupamento

baseado no Risco Relativo de cada atributo.

Transformação

das variáveis em *dummies*.

APLICAÇÃO



RESULTADOS

Método	Desenvolvimento			Validação		
	% acerto	AUC	KS	% acerto	AUC	KS
Árvore de Decisão	66,14%	64,14%	25,30%	65,38%	63,26%	22,27%
Random Forest	80,03%	80,05%	37,91%	68,80%	72,75%	35,55%
Bagging	65,03%	66,81%	28,47%	63,01%	66,16%	25,61%
Adaboost	70,60%	78,12%	37,05%	68,63%	73,04%	36,72%
SVM	68,65%	72,85%	29,89%	67,99%	64,78%	29,57%
Regressão Logística	68,43%	72,89%	34,98%	68,27%	72,74%	34,86%

Random Forest, seguido pelo *Adaboost* e Regressão Logística foram as técnicas que apresentaram **maior percentual de acerto** nas classificações em ambos os cenários.

As técnicas de Árvore de Decisão e *Bagging* obtiveram **desempenho inferior** à abordagem tradicional nas duas amostras.

O SVM apresenta **indicadores bem próximos** dos encontrados na Regressão Logística na amostra de desenvolvimento. Mas no conjunto de validação, a diferença entre os dois aumenta bastante.

RESULTADOS

Método	Desenvolvimento			Validação		
	% acerto	AUC	KS	% acerto	AUC	KS
Árvore de Decisão	66,14%	64,14%	25,30%	65,38%	63,26%	22,27%
Random Forest	80,03%	80,05%	37,91%	68,80%	72,75%	35,55%
Bagging	65,03%	66,81%	28,47%	63,01%	66,16%	25,61%
Adaboost	70,60%	78,12%	37,05%	68,63%	73,04%	36,72%
SVM	68,65%	72,85%	29,89%	67,99%	64,78%	29,57%
Regressão Logística	68,43%	72,89%	34,98%	68,27%	72,74%	34,86%

AUC é um **critério de precisão**, que indica a discriminação entre as classes estudadas.

Na amostra de desenvolvimento, o *Random Forest* e o *Adaboost* apresentaram **desempenho superior** à Regressão Logística, o que se mantém na amostra de validação mas com uma diferença menor.

Com relação ao KS, novamente o desempenho do *Random Forest* e *Adaboost* se mostrou superior a Regressão Logística em ambos os cenários.

RESULTADOS

Método	Desenvolvimento			Validação		
	% acerto	AUC	KS	% acerto	AUC	KS
Árvore de Decisão	66,14%	64,14%	25,30%	65,38%	63,26%	22,27%
Random Forest	80,03%	80,05%	37,91%	68,80%	72,75%	35,55%
Bagging	65,03%	66,81%	28,47%	63,01%	66,16%	25,61%
Adaboost	70,60%	78,12%	37,05%	68,63%	73,04%	36,72%
SVM	68,65%	72,85%	29,89%	67,99%	64,78%	29,57%
Regressão Logística	68,43%	72,89%	34,98%	68,27%	72,74%	34,86%

No geral, a técnica que apresentou **melhor desempenho** na amostra de desenvolvimento foi a *Random Forest*.

Já no conjunto de validação, quem tem melhor performance é o *Adaboost*.

A queda de desempenho do Random Forest pode ser justificada por um **overfitting** nos dados de desenvolvimento, o que é comum em algoritmos que utilizam árvore de decisão.

CONSIDERAÇÕES FINAIS

- A utilização de algumas técnicas de *Machine Learning* em modelos de *Credit Scoring* pode gerar um **maior poder de predição** quando comparado às técnicas tradicionais.

- Uma desvantagem dos métodos de *Machine Learning* é a **interpretação dos parâmetros** do modelo, sendo questionável o seu uso quando não há um ganho de performance significativamente superior às abordagens tradicionais.

REFERÊNCIAS

SELAU, L. P. R. (2008). “Construção de modelos de previsão de risco de crédito”. Master’s thesis, Universidade Federal do Rio Grande do Sul.

REIS FILHO, J. (2006). Sistema inteligente baseado em árvore de decisão, para apoio ao combate às perdas comerciais na distribuição de energia elétrica. Master’s thesis, Universidade Federal de Uberlândia.

OSHIRO, T. M. (2013). Uma abordagem para a construção de uma única árvore a partir de uma Random Forest para classificação de bases de expressão gênica. PhD thesis, Universidade de São Paulo.

CHAVES, B. B. (2012). Estudo do algoritmo AdaBoost de aprendizagem de máquina aplicado a sensores e sistemas embarcados. Master’s thesis, Universidade de São Paulo.

BECKER, W. E. (2017). Uma abordagem de redes neurais convolucionais para análise de sentimento multi-lingual. Master’s thesis, Pontifícia Universidade Católica do Rio Grande do Sul.

OBRIGADA!

Cinthia Becker
cinthia.becker@gmail.com