ÁRVORES, FLORESTAS E SUA FUNÇÃO COMO PREDITORES: UMA APLICAÇÃO NA AVALIAÇÃO DO GRAU DE MATURIDADE DE EMPRESAS

TREES, FORESTS AND THEIR ROLE AS PREDICTORS: AN APPLICATION IN THE EVALUATION OF COMPANIES' DEGREE OF MATURITY

■ **LUIZ CAMPOS DE SÁ LUCAS**

GRADUAÇÃO EM ENGENHARIA PELA **PONTIFÍCIA UNIVERSIDADE CATÓLICA DO RIO DE JANEIRO (PUC-RJ)**; MESTRE EM PROGRAMAÇÃO MATEMÁTICA PELA **UNIVERSIDADE FEDERAL DO RIO DE JANEIRO (UFRJ-COPPE)**; PROFESSOR NA **ESCOLA SUPERIOR DE PROPAGANDA E MARKETING DO RIO DE JANEIRO (ESPME-RJ)**; DIRETOR DE ATENDIMENTO E PLANEJAMENTO DO **IBOPE INTELIGÊNCIA – RJ**; ATUAL REPRESENTANTE NO BRASIL DA **ESOMAR**.

E-MAIL: LUIZSALUCAS@IDSBR.COM.BR

RESUMO

Na área de modelagem matemática em marketing, tanto em pesquisa de mercado como em *data mining,* o chamado aprendizado mecânico supervisionado, isto é, os classificadores, são de extrema importância na estimativa de indicadores como satisfação global com um produto ou serviço, frequência de compra, resposta à mala direta etc. Existem modelos nas mais diversas áreas, seja na estatística mais tradicional, como modelos de regressão múltipla linear ou logística, ou em técnicas mais recentes como *fuzzy modeling,* redes neurais, programação evolucionária etc. Nesse conjunto menos convencional, sobressaem-se as árvores de decisão e, mais recentemente, os conjuntos de árvores de decisão denominadas como florestas aleatórias *(random forests).* Este artigo procurou ilustrar a utilização das árvores de decisão e como as florestas aleatórias são capazes de melhorar a precisão das estimativas dos classificadores resultantes.

PALAVRAS-CHAVE:

Classificadores, árvores de decisão, *random forests.*

ABSTRACT

In the area of marketing mathematical modeling, not only in marketing research but also in data mining, the so-called supervised machine learning techniques, i.e., classifiers, have extreme importance when estimating indicators such as overall satisfaction with a product or service, purchase frequency, response to direct mail, etc. There are models in the most diverse areas, whether in more traditional statistics, such as multiple linear or logistic regressions, or in more recent techniques, such as fuzzy modeling, neural networks, evolutionary programming, etc. In this less conventional set, decision trees stand out and, more recently, the sets of decision trees known as random forests. This work illustrates the use of decision trees and demonstrates how random forests are able to improve the accuracy of the resulting classifier estimates.

KEY WORDS:

Classifiers, decision trees, random forests.

1. INTRODUÇÃO

Na área de modelagem matemática em marketing, tanto em pesquisa de mercado como em *data mining*, o chamado aprendizado mecânico supervisionado, isto é, os classificadores, são de extrema importância na estimativa de indicadores como satisfação global com um produto ou serviço, frequência de compra, resposta à mala direta etc. Existem modelos nas mais diversas áreas, tanto na estatística mais tradicional, como regressão múltipla linear ou logística (VENABLES; RIPLEY, 2002), quanto em técnicas mais recentes como *fuzzy modeling* e redes neurais (COX, 2005), programação evolucionária (EIBEN; SMITH, 2007) etc. Nesse conjunto menos convencional, sobressaem-se as árvores de decisão e, mais recentemente, as florestas aleatórias *(random forests)*.

2. ÁRVORES DE DECISÃO

A importância das árvores de decisão na área de *classifiers* decorre, essencialmente, de duas características:

- Árvores de decisão são excelentes preditores, como será mostrado, adiante, através de um exemplo.
- Permitem não só uma visualização gráfica, mas também geram regras de classificação do tipo *IF* condição *THEN* resultado.

Para ilustrar essas características, será utilizado aquele que talvez seja o conjunto de dados mais famoso da Estatística: os dados de íris de Fisher. Esse conjunto de dados consiste em 150 observações de três tipos de íris: setosa, versicolor e virginica. Na base têm-se 50 observações de cada tipo. Cada observação consiste não só na classificação nos três tipos, mas também em quatro medidas:

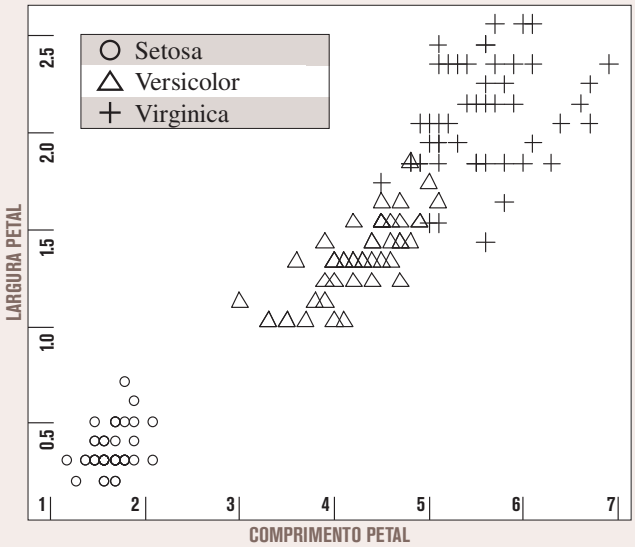

FIGURA 1
Dados de Íris.

- Comprimento da pétala.
- Largura da pétala.
- Comprimento da sétala.
- Largura da sétala.

É sabido que as duas variáveis que melhor se discriminam entre os tipos de íris são comprimento e largura da pétala.

Assim, a Figura 1 apresenta o diagrama de dispersão dos dados, indicando os tipos de íris.

Nota-se que as setosas formam um grupo claramente distinto à esquerda e abaixo na Figura 1. Já as versicolors e as virginicas apresentam uma "zona cinzenta" entre elas, o que dificulta o trabalho do classificador.

Aplicam-se aos dados dois tipos de árvores de decisão:

- CART – *Classification and Regression Tree* (BREIMAN et al., 1984).
- CIT – *Conditional Inference Tree* (HOTHORN; HORNIK; ZEILEIS, 2006).

A Figura 2 apresenta a árvore construída através da CART.

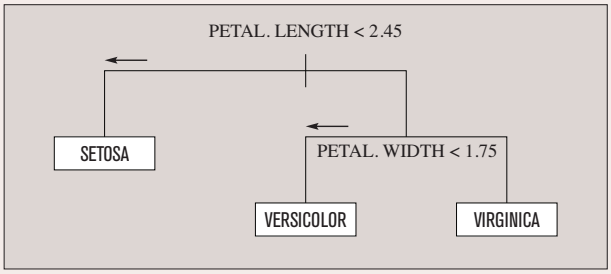

FIGURA 2
CART — íris data.

Já a Figura 3, na sequência, indica a árvore gerada pela CIT para os mesmos dados.

Cada procedimento tem seu método específico de escolher a variável que, em cada nó, irá formar a árvore. Os detalhes podem ser encontrados nas referências já citadas (BREIMAN et al., 1984; HOTHORN; HORNIK; ZEILEIS, 2006).

A ideia geral é escolher, em cada nó, a variável que mais se discrimina entre as respostas. Nesses dois casos, as árvores são sempre binárias: cada nó bifurca a árvore em dois ramos.

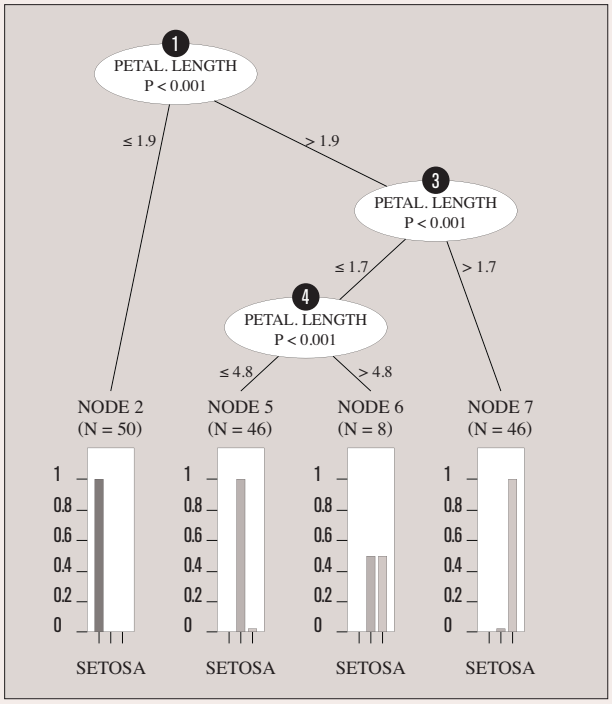

FIGURA 3
CIT — íris data.

Há uma hierarquia: primeiro se utiliza a variável mais discriminante; em seguida, em cada ramo, escolhe-se, para cada novo nó, outra variável mais discriminante e assim por diante.

Comparando a Figura 1 com as Figuras 2 e 3, é fácil verificar que o comprimento petal separa claramente as setosas do grupo (versicolor + virginica).

Pela Figura 2 (CART), se o comprimento petal é menor que 2.45, a observação é classificada como setosa. Em seguida, se o comprimento petal é maior ou igual a 2.45 e a largura petal é inferior a 1.75, a observação é classificada como versicolor, e assim por diante.

Um esquema bastante semelhante é definido pela CIT na Figura 3, porém nesse caso, os pontos de corte são diferentes e o método define quatro grupos em vez de três (comparado com a Figura 1, o quarto grupo corresponde às observações que se misturam naquilo que se denomina de "zona cinzenta").

Essa classificação não é isenta de erro. Em cada ponto terminal (folha), a classificação é dada pela classe dominante. É claro que nem sempre todas as observações da folha pertencerão à mesma classe.

As Tabelas 1 e 2 indicam a precisão da classificação nos dois casos.

TABELA 1
CART – íris data.

	SETOSA	VERSICOLOR	VIRGINICA
SETOSA	50	0	0
VERSICOLOR	0	49	1
VIRGINICA	0	5	45

As colunas indicam a previsão e cada linha soma 50 casos, que é o número real (base). Assim, as 50 setosas foram classificadas corretamente pela CART. Já das 50 versicolors, 49 foram classificadas corretamente, porém 1 foi, incorretamente, classificada como virginica. Finalmente, das 50 virginicas, 45 foram classificadas corretamente e 5, incorretamente, como versicolors. A taxa de acerto *(hit-rate)* foi então de (50 + 49 + 45) / 150 = 96%, o que não é ruim. As árvores de decisão são excelentes preditores, embora nem sempre sejam bons "explicadores" das respostas, ou seja, simulam bem a resposta, mas nem sempre explicam bem essa resposta.

A Tabela 2 indica erros na mesma faixa para a CIT.

TABELA 2
CIT – íris data.

	SETOSA	VERSICOLOR	VIRGINICA
SETOSA	50	0	0
VERSICOLOR	0	49	1
VIRGINICA	0	5	45

A taxa de acerto é a mesma. No entanto, Hothorn, Hornik e Zeileis (2006), afirmam que as CIT têm, em geral, um poder explanatório maior que o das árvores de decisão mais convencionais como as CART.

Outra medida interessante na avaliação da precisão de classificadores é o coeficiente Kappa (SÁ LUCAS, 2007; WITTEN; FRANK, 2005; BEN-DAVID, 2006).

Uma descrição detalhada desse coeficiente está além do escopo deste artigo, mas pode ser encontrada nas referências citadas. Imaginem-se os dados da Tabela 3, onde um classificador extremamente "ingênuo" classificou três grupos de tamanhos 100/800/100 no segundo grupo — o de maior incidência:

TABELA 3
Um classificador "ingênuo".

	SETOSA	VERSICOLOR	VIRGINICA
SETOSA	0	100	0
VERSICOLOR	0	800	0
VIRGINICA	0	100	0

O *hit-rate* é de 80%, o que indicaria um excelente poder de predição; no entanto, o coeficiente Kappa, nesse caso, é 0%, o que indica que o classificador não está prevendo nada.

Nos casos dos dois classificadores indicados anteriormente (CART e CIT) nos dados de íris, o *hit-rate* foi de 96% e o Kappa de 94%. A pequena diferença entre os dois se deve ao fato de que o número de casos nos três grupos é de 50, logo, não existe efeito "número de grupos" a ser descontado. O Kappa não é de grande valor aqui.

3. FLORESTAS ALEATÓRIAS

As árvores de decisão são excelentes preditores. No entanto, podem apresentar alguma dificuldade, pois nem sempre generalizam bem. É essa a razão pela qual em CART, por exemplo, se executa a poda *(pruning)* da árvore, que fica menor e mais generalizável, "aderindo" menos a possíveis idiossincrasias da amostra (um fenômeno muitas vezes descrito como *overfitting)*. Uma descrição detalhada de *pruning* pode ser obtida em Venables e Ripley (2002) ou em Maindonald e Braun (2003).

Um passo além, nesse processo, é o uso de métodos que utilizem técnicas como:

- *Boosting:* Utilização de um conjunto de técnicas menos precisas *(weak learners)* para se obter um resultado de consenso que tenha boa qualidade, já que uma solução complementa a outra.
- *Bagging:* Uso de uma pluralidade de votos (ou mesmo de amostras) como base para uma boa solução que seja mais generalizável.
- *Randomizing:* Em cada método/técnica empregada, utiliza-se um conjunto diferente de variáveis.

Um bom exemplo é a técnica de *Random Forests* (BREIMAN, 2001), que apresenta excelentes características de precisão, generalização para outras amostras que não aquelas em que o classificador foi treinado e capacidade de bom desempenho em pequenas amostras. Para maiores detalhes da aplicação da técnica na determinação da Importância Derivada, veja, por exemplo, estudos de satisfação em pesquisa de mercado em Soares e Esteves (2008).

Essencialmente, a técnica de *Random Forests* consiste em:

- *Bagging:* Gerar, através de reamostragem *(bootstraping)*, um conjunto de, por exemplo, 500 amostras retiradas da amostra original através de seleção aleatória com reposição.
- *Boosting:* Desenvolver em cada amostra, uma CART onde, para cada observação incorretamente classificada numa determinada amostra do *bootstrapping*, sua ponderação seja aumentada quando houver a criação de uma nova árvore em outra amostra do mesmo *bootstrapping*.
- *Randomizing:* Desenvolver a CART utilizando, em cada nó, um subconjunto de, por exemplo, apenas cinco variáveis, selecionadas aleatoriamente, para executar a divisão *(split)*. Os subconjuntos em cada nó são diferentes numa mesma CART.

Os modelos preditivos derivados de *Random Forests*, em geral são tão bons ou melhores do que os obtidos com uma CART individual.

A Tabela 4 indica a aplicação da técnica aos dados de íris, onde a taxa de acerto agora é de 100%.

TABELA 4
Random Forests – íris data.

	SETOSA	VERSICOLOR	VIRGINICA
SETOSA	50	0	0
VERSICOLOR	0	50	0
VIRGINICA	0	0	50

O *hit-rate* e o Kappa, nesse caso, são iguais, no valor de 100%.

Outra característica extremamente interessante é a capacidade do método em indicar a importância de cada variável de entrada na saída estimada. Esse aspecto é descrito de forma mais detalhada em Soares e Esteves (2008).

Essa importância muitas vezes é denominada de Importância Derivada e, no caso dos dados de íris, tomam os valores indicados na Tabela 5, onde se notam claramente como os dados relativos à pétala dominam a classificação.

TABELA 5
Importância Derivada *Random Forests* — íris data.

VARIÁVEL	IMPORTÂNCIA DERIVADA
COMPRIMENTO DA SÉPALA	11%
LARGURA DA SÉPALA	2%
COMPRIMENTO DA PÉTALA	43%
LARGURA DA PÉTALA	44%

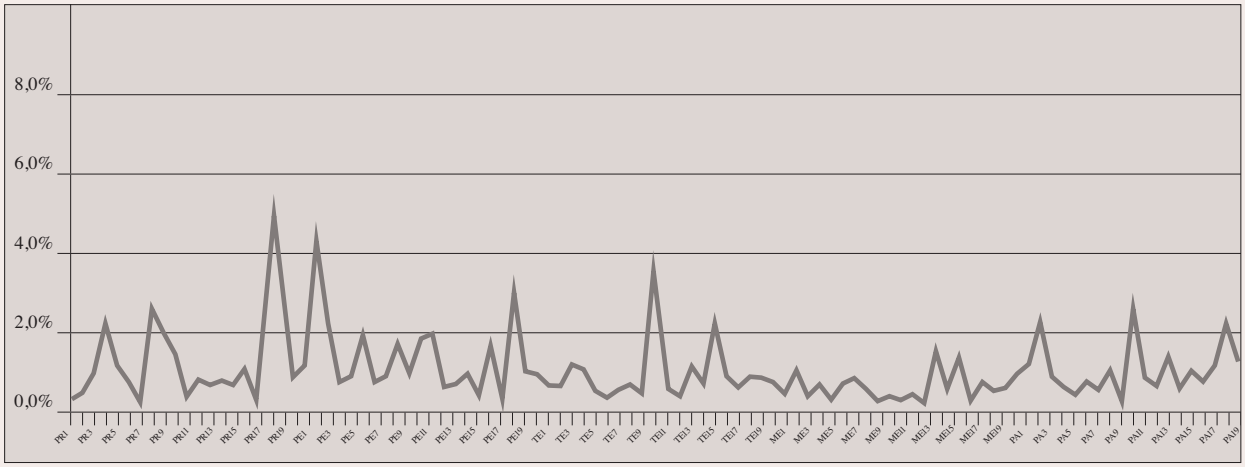

GRÁFICO 1
Importância Derivada — *Random Forests* — Excelência Competitiva.

4. EXCELÊNCIA COMPETITIVA

Magalhães (2010), em sua tese de doutorado, apresenta uma técnica de avaliação do desempenho de empresas, denominado **Excelência Competitiva,** que se baseia em cinco dimensões: Processos, Pessoas, Tecnologia, Mercado e Parcerias. Em cada uma dessas dimensões, Magalhães definiu dez processos típicos de cada dimensão. A avaliação de cada processo ia de um valor mínimo de 1, característico de cada processo, a um máximo de 5, variável também conforme o processo. Havia também uma avaliação global de 1 a 5 em cada dimensão e uma avaliação global final de 1 a 5: avaliação global da empresa. Os detalhes do método podem ser obtidos em Magalhães (2010).

A **Excelência Competitiva** foi aplicada a uma amostra de cerca de 50 empresas participantes do Programa Parceiros para Excelência (Paex), da Fundação Dom Cabral e das empresas parceiras do programa e dos professores e coordenadores do B.I. International.

O B.I. International é a mais global das escolas de negócios do Brasil e parceira de *business schools* nos EUA, Ásia e Oceania, referências mundiais em suas áreas de conhecimento, e que, há 20 anos, gerencia e contribui para a formação de mais de 30 mil empresários, executivos e profissionais liberais no Brasil.

A vantagem aqui da aplicação do *Random Forests* deriva inclusive do fato de que a técnica é bastante robusta para pequenas amostras. As importâncias derivada obtidas estão ilustradas no Gráfico 1.

A cada ponto da escala horizontal corresponde uma das 50 práticas. Evidentemente o espaço disponível não permite representar todas as legendas. No entanto, como se nota, a maioria das importâncias ficou entre 1% e 2%, com algumas poucas práticas indo até 4%, o que indica que todos os processos contribuem, efetivamente, para a avaliação global.

Os resultados detalhados são apresentados em Magalhães (2010), mas cabe ressaltar que, ao final do processo, foi definido um grau de maturidade para cada empresa igual a:

$$GM = \sum_{i=1}^{n} \alpha_i X_i$$

Onde:

GM = grau de maturidade.
n = número de práticas (igual a 100).
α_i = peso dado pelos *Random Forests* a i-ésima prática.
X_i = desempenho na i-ésima prática da empresa.

Com base nesse Índice de Maturidade, as empresas foram classificadas, conforme mostra a Tabela 6.

TABELA 6
Índice de Maturidade das empresas.

GRAU DE MATURIDADE	%
PERENE	6,3%
VENCEDORA	34,4%
COMPETITIVA	50,0%
SOBREVIVENTE	6,3%
FRÁGIL	3,1%
TOTAL	**100,0%**

Onde:
- Frágeis: GM igual a 1.
- Sobreviventes: GM de 1,1 a 2.
- Competitivas: GM de 2,1 a 3.
- Vencedoras: GM de 3,1 a 4.
- Perenes: GM acima de 4.

A grande maioria das empresas foi classificada na categoria de Vencedoras ou Competitivas, o que concorda com a natureza da amostra.

Obviamente aqui não cabe calcular *hit-rates* ou Kappas. Não se dispõe do coeficiente de correlação nesse caso, mas em vários outros estudos em que o autor participou o coeficiente de correlação entre valor estimado para avaliação global e valor real, foi bastante elevado, o que concorda com a literatura a respeito, com relação à precisão do estimador.

5. CONCLUSÃO

Pelo exposto, fica claro o poder e precisão das florestas aleatórias nos modelos preditivos.

Finalmente, cabe uma observação quanto à utilização das árvores *versus* a utilização das florestas. Muitas vezes, esses modelos preditivos deverão ser implementados em enormes bases de dados com, por exemplo, milhões de registros. Em geral, esses registros são acessados por sistemas computacionais com capacidade limitada de procedimentos matemáticos. Nesse caso, é natural a opção de se utilizar as árvores de decisão, que podem ser programadas facilmente, mesmo em linguagens como SQL e similares, através de regras simples como:

- *IF* condição *THEN* resultado.

Uma comparação detalhada da qualidade dos resultados derivados através de CART, CIT, *Random Forests* com outros métodos complexos, como *Adaboost* e *Multiboost,* pode ser obtida em Sá Lucas (2007).

7. REFERÊNCIAS BIBLIOGRÁFICAS

BEM-DAVID, A. *What's wrong with Hit ratio?* IEEE Intelligent Systems, v. 21, n. 6, 2006.

BREIMAN, L. Random Forest. *Machine Learning,* 45 (l):5-32, 2001.

BREIMAN, L.; FRIEDMAN, J. H.; OLSHEN, R. A.; STONE, C. J. *Classification and regression trees.* New York: CRC Press, 1984.

COX, E. *Fuzzy modeling and genetic algorithms for data mining and exploration.* San Francisco: Morgan Kaufmann, 2005.

EIBEN, A. E.; SMITH, J. E. *Introduction to evolutionary computing.* Berlin: Springer Verlag, 2007.

HOTHORN, T.; HORNIK, K. E.; ZEILEIS, A. Unbiased recursive partitioning: a conditional inference framework. American Statistical Association. *Journal of Computational and Graphical Statistics,* v. 15, n. 3, p. 651-674, 2006.

MAINDONALD, J.; BRAUN, J. *Data analysis and graphics using* R. New York: Cambridge University Press, 2003.

MAGALHÃES, M. F. *Excelência competitiva — a execução da estratégia nas empresas que visam durar.* Tese de doutorado. Universidade Federal do Rio de Janeiro — COPPE, 2010.

SÁ LUCAS, L. *Joint Segmenting Consumers using Both Behavioral and Attitudinal Data. Sawtooth Software Proceedings,* p. 199-220, 2007.

SOARES, L.; ESTEVES, W. Consumer drivers: estimando a importância derivada em pesquisa de mercado a partir de *Random Forests. 3º Congresso Brasileiro de Pesquisa — Mercado, Opinião e Mídia,* ABEP, 2008.

VENABLES, W. N.; RIPLEY, B. D. *Modern applied statistics with* S. New York: Springer Verlag, 2002.

WITTEN, I.; FRANK, E. *Data mining: practical machine learning tools and techniques.* Elsevier, 2005.