ABRAHAM LAREDO SICSÚ (org.)
ANDRÉ SAMARTINI
NELSON LERNER BARTH

# TÉCNICAS DE MACHINE LEARNING

Blucher

Abraham Laredo Sicsú (org.)
André Samartini
Nelson Lerner Barth

# TÉCNICAS DE MACHINE LEARNING

*Técnicas de machine learning*

Editora Edgard Blücher Ltda.

*Publisher* Edgard Blücher
*Editor* Eduardo Blücher
*Coordenação editorial* Jonatas Eliakim
*Produção editorial* Ariana Corrêa
*Preparação de texto* Amanda Fabbro
*Diagramação* Roberta Pereira de Paula
*Revisão de texto* Maurício Katayama
*Capa* Leandro Cunha
*Imagem da capa* iStockphoto

Em caso de futuras correções técnicas, gráficas ou atualizações do software, o material ficará disponível na página do livro, no site da editora. Acesse pelo QRcode localizado na quarta capa.

Blucher

Rua Pedroso Alvarenga, 1245, 4º andar
04531-934 – São Paulo – SP – Brasil
Tel.: 55 11 3078-5366
**contato@blucher.com.br**
**www.blucher.com.br**

Segundo o Novo Acordo Ortográfico, conforme 6. ed. do *Vocabulário Ortográfico da Língua Portuguesa*, Academia Brasileira de Letras, julho de 2021.



Dados Internacionais de Catalogação na Publicação (CIP)
Angélica Ilacqua CRB-8/7057

Sicsú, Abraham Laredo

Técnicas de machine learning / organização de Abraham Laredo Sicsú ; André Samartini, Nelson Lerner Barth. – São Paulo : Blucher, 2023.

394 p. : il.

Bibliografia

ISBN 978-65-5506-396-7

1. Matemática e estatística – Processamento de dados 2. Algoritmos 3. Dados – Análise 4. Dados – Aglomeração – Análise 5. Modelos matemáticos I. Título II. Samartini, André III. Barth, Nelson Lerner

22-6713 CDD 519.5

Índice para catálogo sistemático:
1. Matemática e estatística – Processamento de dados

# CONTEÚDO

# CAPÍTULO 1
# Fundamentos

*Prof. Abraham Laredo Sicsú*

## 1.1 INTRODUÇÃO

Este livro foi escrito com o objetivo de permitir que analistas de empresas e pesquisadores entendam e apliquem diferentes algoritmos de *machine learning,* sem aprofundar-se nas teorias que os fundamentam. Preferimos apresentar os principais conceitos de forma intuitiva, para que o usuário compreenda a lógica de cada algoritmo, para que finalidade deve ser utilizado, quais os passos a seguir para aplicá-lo corretamente utilizando o software R e suas vantagens e limitações quando comparado com outros algoritmos que podem ser utilizados no mesmo problema.

Os algoritmos aqui apresentados podem ser utilizados nas mais diferentes áreas de atividade e conhecimento. Vejamos alguns exemplos:

- Previsão de salários: para poder estimar o salário de um operador em determinado segmento industrial, considerando sua formação, idade, experiência, especialidade, região onde trabalha etc., podemos construir um modelo a partir de uma amostra de operadores cujos salários sejam conhecidos. Estabelecendo a relação entre esses salários e as características já citadas, é possível não somente estimar o salário de um operador como também identificar os aspectos com maior peso na variação do salário.

- Um problema importante na área de crédito é prever se um cliente pagará ou não um empréstimo, ou seja, classificá-lo como potencial bom ou mau pagador. Partindo das variáveis que caracterizam os solicitantes do crédito (idade, profissão, estado civil etc.) podemos criar uma regra que permita obter tal classificação. Hoje, praticamente todas as empresas que concedem crédito utilizam essa metodologia para a tomada de decisão.
- Uma aplicação usual de *machine learning* em marketing é a previsão do desligamento voluntário (*churning*), ou seja, o cancelamento de um contrato de uso de um serviço por parte do cliente. Em especial, às operadoras de telefonia celular convém prever com bastante antecedência se um cliente irá cancelar seu contrato optando, provavelmente, por outra operadora. Se a probabilidade de *churning* for alta, a operadora poderá fazer uma série de ofertas ao cliente tentando evitar que ele cancele seu contrato. O mesmo se aplica a uma seguradora, prevendo a probabilidade de renovação ou não do contrato.
- Uma área em que *machine learning* poderia ser mais utilizada é em RH. Por exemplo, na contratação de funcionários em uma empresa. Admita que uma grande rede varejista deseja aumentar a efetividade da sua força de vendas. Comparando vendedores dessa empresa que tiveram bom desempenho no passado com os que tiveram mau desempenho, é possível construir um modelo que permita classificar um candidato a vendedor com potencial bom desempenho no futuro.
- Em medicina, as técnicas de classificação podem ser utilizadas no diagnóstico de doenças. Por exemplo, em um estudo bastante conhecido na área médica, com base em indicadores relativos à pressão arterial, hábito de fumo, diabetes, sedentarismo etc., aplicou-se uma técnica de classificação para discriminar um paciente em um de dois grupos: com ou sem alto risco de desenvolver doenças cardiológicas.
- Uma aplicação interessante para conhecer melhor os clientes de um banco e direcionar eficazmente campanhas de marketing é o agrupamento dos clientes em segmentos homogêneos, no que tange ao seu relacionamento ou uso dos serviços prestados pelo banco. Esse tipo de agrupamento é mais eficaz que a usual segmentação de mercado por faixa etária ou gênero.

## 1.2 ALGORITMOS E MODELOS

Vamos diferenciar dois conceitos neste texto: algoritmo e modelo.

*Algoritmo* é um conjunto de procedimentos a serem executados, em determinada sequência, a fim de transformar um conjunto de dados de entrada em um ou mais valores como saída.

*Modelo* é o produto que resulta quando aplicamos um algoritmo a uma base de dados. Em um sentido mais amplo, o modelo deve incluir não só os valores fornecidos pelo algoritmo, mas também, caso necessário, regras a serem seguidas para utilizá-lo nas previsões ou classificações dos dados da população. Um mesmo algoritmo, quando aplicado a diferentes conjuntos de dados, fornece diferentes modelos.

Um exemplo de algoritmo é o conjunto de passos para ajustar uma reta de mínimos quadrados a uma nuvem de pontos. O algoritmo recebe as coordenadas dos pontos, calcula os parâmetros *a* e *b* e fornece o modelo $y = a + bx$. Com esse modelo podemos fazer previsões para novos valores de x. Diferentes bases de dados, utilizando o mesmo algoritmo, fornecerão diferentes valores de *a* e *b,* o que define, portanto, diferentes modelos. O modelo representa o que foi aprendido pelo algoritmo ao utilizar a base de dados.

## 1.3 PARÂMETROS E HIPERPARÂMETROS

Parâmetros são características de um modelo a serem estimados pelo algoritmo. Por exemplo, retomando o exemplo da reta y = a + bx, *a* e *b* são parâmetros a serem estimados a partir dos dados.

Alguns algoritmos possuem caraterísticas que não podem ser estimadas a partir dos dados e devem ser definidas pelo analista antes de rodá-los. Vamos denominá-los hiperparâmetros. Por exemplo, quando rodamos uma rede neural o número de camadas intermediárias da rede e o número de neurônios de cada camada devem ser fixados *a priori*. O mesmo ocorre em procedimentos em que, por exemplo, a classificação de um indivíduo é determinada pelas classificações de seus vizinhos mais próximos na nuvem de pontos (*kNN: k nearest neighbor*). O tamanho da vizinhança, ou seja, o número k de vizinhos a serem considerados deve ser prefixado pelo analista.

A definição do valor mais adequado dos hiperparâmetros é feita por tentativa e erro pelo analista. Alterando o valor do hiperparâmetro ou do conjunto de hiperparâmetros necessários, ele pesquisará qual o valor ou combinação de valores que otimiza os resultados do algoritmo, quando aplicado àquela base de dados.

## 1.4 CLASSIFICAÇÃO DOS ALGORITMOS

Os algoritmos que serão apresentados neste texto podem ser classificados em duas grandes famílias: algoritmos supervisionados e algoritmos não supervisionados.

### 1.4.1 ALGORITMOS SUPERVISIONADOS

Os algoritmos supervisionados são aplicados quando temos uma base de dados em que a cada observação corresponde um conjunto de variáveis X1, X2, …, Xp, geralmente denominadas previsoras; e uma variável Y denominada variável alvo.[1] Alguns autores dizem que cada observação tem uma etiqueta (*label*) Y.

**Tabela 1.1 –** Base de dados para métodos supervisionados

| Previsoras | | | | Alvo |
|---|---|---|---|---|
| $X_1$ | $X_2$ | ... | $X_p$ | Y |
| 1 | 5 | 6 | -4 | 8 |
| 2 | 3 | 8 | 6 | 7 |
| 3 | 2 | 11 | -5 | -4 |

A variável Y funciona como um guia (supervisor) do algoritmo. Pode ser quantitativa ou qualitativa. A construção desses algoritmos fundamenta-se em detectar a relação entre as variáveis previsoras e a variável alvo. Uma vez identificada a relação, ela pode ser aplicada para prever (Y quantitativa) ou classificar (Y qualitativa) novos casos a partir das variáveis previsoras.

Os algoritmos supervisionados de previsão têm por objetivo prever o valor de uma variável alvo quantitativa Y em função das variáveis previsoras X1, X2, …, Xp. Por exemplo, consideremos uma amostra de apartamentos com valor Y conhecido e algumas caraterísticas, sendo X1 = área útil, X2 = número de suítes, X3 = andar, X4 = idade em anos e X5 = vagas etc. A partir dessa amostra, utiliza-se um algoritmo de previsão para construir um modelo que permita prever preços de novos apartamentos com base nessas mesmas caraterísticas.

[1] A variável Y é denominada *variável dependente ou variável resposta* ou, como é usual em *machine learning*, variável alvo (*target*). Utilizaremos indistintamente cada uma dessas denominações no decorrer do texto.

**Tabela 1.2 –** Dados para um problema de previsão

| Obs. | Area útil | Suítes | Andar | Idade | Vagas | .... | Y = valor (R$) |
|---|---|---|---|---|---|---|---|
| 1 | 140 | 1 | 3º | 5 | 2 | | 750.000,00 |
| 2 | 79 | 1 | 5º | 8 | 1 | | 425.000,00 |
| 3 | 160 | 2 | 14º | 2 | 3 | | 1.200.000,00 |
| | ... | ... | ... | ... | ... | | ... |

O algoritmo supervisionado de previsão mais conhecido é o de regressão linear múltipla. Outros exemplos são as árvores de regressão, *random forests*, *XGBoost* e redes neurais aplicadas à previsão.

Os algoritmos supervisionados de classificação são utilizados para classificar uma nova observação em uma das categorias da variável qualitativa Y, ou seja, para preverem em qual categoria deve ser classificada uma nova observação. A variável alvo Y pode ser binomial (duas categorias: bom/mau pagador, spam/não spam, renova/não renova o seguro etc.) ou multinomial (três ou mais categorias: funcionário com alto/médio/baixo potencial, cliente de uma livraria prefere romances/biografias/poesias/história/outros). Exemplos de algoritmos utilizados para classificação são a regressão logística, as árvores de classificação, as *random forests*, *XGBoost* e o SVM-*Supervised Vector Machine*.

Por exemplo, uma aplicação comum na área financeira é classificar, a partir de uma amostra de clientes, qual será o comportamento de um futuro tomador de crédito (Y: bom ou mau pagador) a partir de suas características sociodemográficas.

**Tabela 1.3 –** Dados para um problema de classificação

| Cliente | Idade | UF | Resid. | Est. civil | Instrução | renda | Y=status |
|---|---|---|---|---|---|---|---|
| 1 | 33 | SP | Própria | Casado | Fundamental | 1200 | Bom |
| 2 | 52 | PA | Própria | Solteiro | Fundamental | 6500 | Bom |
| 3 | 65 | SP | Própria | Solteiro | Superior | 7832 | Mau |
| . . . | | | | | | | |

### 1.4.2 ALGORITMOS NÃO SUPERVISIONADOS

Nos algoritmos não supervisionados não há uma variável alvo que sirva para direcionar os resultados. A função desses algoritmos é obter determinados padrões de comportamento entre as observações da amostra. O algoritmo deve ser programado para aprender a identificar tais padrões.

**Tabela 1.4 –** Dados para um método não supervisionado

| X1 | X2 | ... | Xp |
|---|---|---|---|
| 2,3 | 3 | ... | 100,51 |
| 3,8 | 10 | ... | 89,32 |
| 11,5 | 8 | ... | 27,65 |

A não existência de um alvo Y torna os algoritmos mais complexos e as saídas mais difíceis de interpretar. Ao contrário dos métodos supervisionados, nos quais a comparação da previsão como o valor conhecido Y permite aferir a qualidade dos resultados, a interpretação e validação dos resultados são, em geral, uma tarefa complexa.

Há vários tipos de algoritmos não supervisionados. Neste livro vamos tratar apenas dos algoritmos de agrupamento, provavelmente os mais utilizados. Seu objetivo é classificar[2] em grupos homogêneos as diferentes observações.

Uma aplicação clássica dos algoritmos de agrupamento, conhecida como segmentação de mercado, é classificar os diferentes clientes de uma empresa em grupos homogêneos de acordo com seus perfis de consumo. Esse agrupamento permite alavancar as vendas oferecendo propostas diferenciadas para cada um dos segmentos.

Outra aplicação interessante pode ser o agrupamento de uma série de marcas e tipos de cervejas em função de variáveis que medem diferentes percepções dos consumidores. Produtos que pertencem a um mesmo grupo são vistos como parecidos pelos clientes e como concorrentes pelos fabricantes.

Outros algoritmos não supervisionados permitem, por exemplo, identificar associações do tipo "*se X* ➔ *então Y*" entre produtos. As regras de associação, muito utilizadas no *e-commerce*, são um bom exemplo dessa aplicação. Ao escolher um produto, o site sugere outros produtos que são frequentemente comprados simultaneamente. Por exemplo, se compra cachorro-quente então, provavelmente, compra também mostarda, pãezinhos, cerveja etc.

[2] Neste contexto, utilizamos classificar como sinônimo de agrupar.

## 1.5 O DILEMA VIÉS – VARIABILIDADE (BIAS – *VARIANCE TRADE OFF*)

Ao rodar um algoritmo para obter a relação entre a variável alvo Y e as variáveis previsoras (X1, X2, …), os valores ajustados de Y, em geral, não são iguais ou próximos dos valores observados. As diferenças entre esses valores são denominadas erros. O objetivo do analista é obter um modelo com erros pequenos, que possa ser *generalizado*, isto é, aplicado a outras amostras da população com bons resultados.

Quando tentarmos representar uma relação complexa entre a variável alvo e as variáveis previsoras por meio de um modelo simples, este provavelmente não refletirá corretamente essa relação. Essa situação é conhecida como *underfitting* (*subajuste*). A diferença entre os valores ajustados e os valores observados é grande. Diremos que o modelo apresenta um *viés* alto. Por exemplo, consideremos os pontos no gráfico seguinte, gerados artificialmente a partir de um polinômio de quarto grau.[3] Os pontos amostrais gerados (*x, ylrn*)[4] estão representados em cinza-claro.

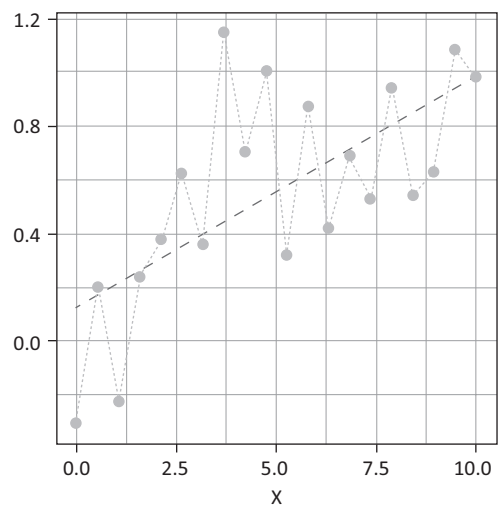


**Figura 1.1 –** Gráfico mostrando *underfitting*.

Ao ajustar a reta *yhat* (tracejada), notamos que os pontos amostrais estão bem distantes da reta. Ou seja, o ajuste não é bom.

Por outro lado, quando optamos por um modelo complexo para representar a relação entre as variáveis, de forma que os erros sejam praticamente nulos, podemos

[3] Adicionamos pequenos erros aleatórios (ruído) para simular uma situação real.

[4] Ylrn: o sufixo lrn (de learn) indica a amostra com a qual o algoritmo ajustou os dados. São os valores observados.

incorrer em outro problema. O modelo provavelmente apresentará grande *variância*, ou seja, será muito sensível a pequenas mudanças nos dados amostrais. O modelo complexo "decora" a amostra a partir do qual foi construído, considerando até o ruído no seu ajuste. Caímos na situação conhecida como *overfitting* (superajuste). O modelo funciona muito bem com a amostra utilizada na sua obtenção, mas não pode ser generalizado para aplicação em outras amostras da população, pois apresentará má performance. Isso torna o modelo inútil.

Consideremos o mesmo conjunto de dados anteriores e, a partir dele, ajustemos um polinômio de grau 19. Como temos 20 pontos, o ajuste será perfeito. Os erros serão todos nulos! A Figura 1.2 mostra uma amostra adicional de pontos (*x, ytest*)[5] (cinza escuro) extraídos da mesma população, e a curva polinomial (pontilhada) ajustada à amostra (*x, ylrn*) (cinza claro). Notamos que a curva polinomial se ajusta perfeitamente à amostra (*x, ylrn*). Notamos também que os pontos da amostra adicional estão bem distantes da curva polinomial. A média dos erros, em valores absolutos, é 0,24. Considerando a magnitude dos valores de Y, trata-se de um erro médio bastante grande.

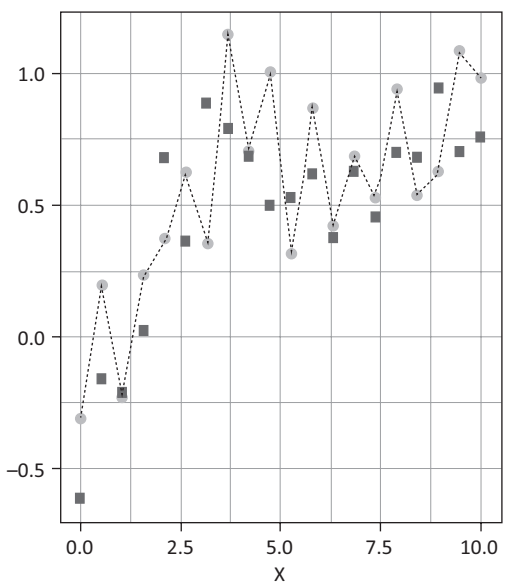


**Figura 1.2 –** Gráfico mostrando *overfitting*.

Em suma, temos dois problemas em modelagem: viés e variância. A utilização de um modelo muito simples (*underfitting*) para o problema conduz a sério viés; à medida

[5] *ytest*: o sufixo test indica a amostra utilizada para testar o modelo e verificar se ele pode ser generalizado para outros pontos da população.

que a complexidade do modelo aumenta, a variância aumenta podendo, eventualmente conduzir ao *overfitting*. O analista deverá encontrar um modelo que apresente um bom *trade-off* entre viés e variância. Em geral, preferimos trabalhar com um algoritmo que gere um pequeno viés e evitar um modelo superajustado que não possa ser generalizado para o restante da população.

Uma forma de medir a variabilidade e o viés de um algoritmo, quando aplicado aos dados de uma população, é rodando esse algoritmo a partir de diferentes amostras aleatórias de mesmo tamanho extraídas dessa população. Outras formas interessantes, que serão descritas adiante contemplam a utilização dos métodos de reamostragem e a troca dos valores de inicialização requeridos pelo algoritmo.

Quando o modelo matemático a ser obedecido pelo algoritmo é fornecido pelo analista, por exemplo, em regressão múltipla ou regressão logística, a tendência é um maior viés e menor variância. Por outro lado, algoritmos menos restritivos, que não forçam o ajuste de uma determinada relação matemática entre as variáveis, e dependem da detecção, pela máquina, da relação entre as variáveis, apresentam maior variância.

Algumas alternativas para reduzir a possibilidade de ocorrência de *overfitting* são:

- Trabalhar com maiores bases de dados para treinar os modelos.
- Controlar os hiperparâmetros do algoritmo.
- Trabalhar com *ensemble methods*. São combinações de algoritmos que serão explicados em capítulo adiante.

## 1.6 PREMISSA FUNDAMENTAL EM MODELAGEM

Quando se desenvolve um modelo é utilizada uma ou mais bases de dados. Esses dados representam uma situação anterior à data de desenvolvimento. Ou seja, são dados históricos, ainda que recentes. O objetivo ao desenvolver um modelo é poder aplicá-lo no futuro para poder prever ou classificar novos indivíduos. Porém, se a realidade no futuro for muito diferente do passado, o modelo não terá utilidade; seria o equivalente a dirigir um automóvel olhando pelo retrovisor.

Por exemplo, utilizar as vendas mensais de uma empresa para prever vendas futuras, considerando como histórico (amostra) as vendas de 2020 e 2021, anos atípicos em razão do impacto da pandemia na economia, será de pouca valia (esperando que no futuro voltemos ao normal, sem esses problemas). O mesmo valeria para classificar clientes que solicitam crédito utilizando modelos desenvolvidos com dados anteriores a 2020. O desemprego e a queda do poder aquisitivo da população, decorrente da Covid-19, certamente afetaram a capacidade dos tomadores de crédito de honrar os compromissos financeiros assumidos. O perfil do novo inadimplente provavelmente será diferente dos perfis dos inadimplentes utilizados para desenvolver o modelo. Em outras palavras, o modelo de classificação deixa de ser confiável. Anos atrás, quando houve um confisco de bens no governo do Presidente Collor, muitos modelos financeiros deixaram de valer da noite para o dia. Simplesmente tornaram-se inúteis.

Nessas situações utilizar um modelo histórico é arriscado. O que muitas empresas fazem em situações em que há uma pequena mudança no comportamento de mercado é utilizar o modelo para ter uma ideia, ainda que grosseira, da previsão ou classificação e, posteriormente, ajustar os resultados utilizando argumentos subjetivos, com base nos sentimentos dos analistas da área. Não é o ideal, aumenta o risco, mas não haveria outra saída. Nenhuma empresa desejaria esperar o término da pandemia mais um período de pelo menos 12 meses para coletar dados e desenvolver um novo modelo. Já houve casos em que ficamos na dúvida se não seria mais interessante, ou mais simples, voltar à tomada de decisões sem a utilização de modelos.

## 1.7 PRIMEIROS PASSOS

Em geral, o desenvolvimento de modelos de *machine learning* segue um roteiro bem definido. As principais etapas são dadas a seguir:

1) Entender o problema, definir objetivos e avaliar a viabilidade.
2) Identificar as variáveis.
3) Coletar as amostras.
4) Analisar e tratar os dados para adequá-los ao desenvolvimento do modelo.
5) Aplicar o(s) algoritmo(s) selecionados e fazer ajustes para melhorar sua performance.
6) Analisar e interpretar os resultados.
7) Validar o modelo.

A seguir, vamos discutir os passos 1, 2 e 3. Os demais passos serão discutidos adiante, em capítulos separados.

A compreensão e formalização dos objetivos, muitas vezes não respeitadas na ânsia de "rodar o algoritmo", são fundamentais para o sucesso do projeto. É importante que o analista de *machine learning* converse com os usuários do modelo (que denominaremos clientes – internos ou externos) para esclarecer e definir, entre outros pontos:

- Para que querem o modelo (que resultados esperam)?
- O modelo é necessário?
- Como e onde será aplicado o algoritmo?
- Qual(is) o(s) critério(s) para que o usuário considere o modelo confiável e útil?
- A partir de que bases de dados o modelo poderá ser desenvolvido? Essa base é suficiente?
- Poderão ser adquiridos dados externos para obter modelos mais confiáveis?
- Quais indivíduos da empresa participarão da elaboração do modelo?

- Qual o prazo desejado para elaboração e o tempo previsto para implantação do modelo? (Perguntas difíceis de responder a esta altura do planejamento, mas é interessante ter ideia das expectativas.).

Nossa recomendação é conversar muito com os usuários dos modelos antes de começar qualquer atividade, para extrair deles as respostas às perguntas acima. O mais importante é verificar se o modelo é realmente necessário e se é viável. Não é raro que um cliente solicite um modelo sem saber exatamente o que deseja. Suas demandas, às vezes, são muito vagas. Em certas ocasiões, ao tomar conhecimento de que um concorrente desenvolveu determinada aplicação, o cliente solicita a elaboração de um modelo similar, mas que não se aplica à sua realidade. Em outras situações, mais raras, o cliente nem precisa de um novo modelo!

Em grande parte dos casos o cliente não tem nem mesmo uma base de dados adequada para elaborar o modelo. Nesses casos precisamos verificar sua disponibilidade em adquirir dados de bureaux de informações. Isso é um problema sério, especialmente no caso de novas empresas, como as *fintechs*.

O conhecimento do contexto do problema (*problem domain*) é fundamental em todas as etapas do desenvolvimento. Dificilmente um analista de dados, por mais que conheça os algoritmos de *machine learning*, poderá obter sucesso sem conhecer as peculiaridades do negócio ou da empresa. É importante conseguir a participação de funcionários da empresa solicitante, ou de consultorias externas envolvidas com a empresa, para validar, por exemplo, algumas transformações dos dados ou para avaliar a ocorrência de dados atípicos. Em particular, devemos contar com pessoas da área de informática que conheçam as bases de dados e fiquem a par das especificidades do modelo necessárias para sua implantação em produção.

## 1.8 IDENTIFICAÇÃO DAS VARIÁVEIS PREVISORAS

A identificação das variáveis previsoras (ou atributos), é um passo fundamental na elaboração das regras de classificação. A qualidade de um modelo é função das variáveis utilizadas em seu desenvolvimento. A omissão de variáveis relevantes comprometerá seriamente a qualidade do modelo resultante. Identificar as variáveis adequadas é um misto de experiência, conhecimento e arte. A identificação correta das variáveis só será possível se os objetivos estabelecidos para o modelo estiverem definidos de forma muito precisa.

Para a identificação das variáveis é fundamental que especialistas da área contribuam apontando fatores que possam influir nas previsões e classificações. Desprezar essa experiência é um erro grave, frequentemente cometido por analistas de métodos quantitativos. Por exemplo, ao identificar as variáveis a serem utilizadas em modelos para classificar solicitantes de empréstimos como bons ou maus pagadores, devemos ouvir analistas de crédito que conheçam profundamente a área e, com base em sua

experiência, sugiram variáveis que acreditam diferenciar os dois tipos de clientes. Isso não significa que esta deva ser a única fonte de sugestão de variáveis potenciais.

O ideal é coletar essas informações com auxílio de um *brainstorming* ou com simples e numerosas entrevistas, sem se preocupar, inicialmente, com um excesso de preciosismo na definição das variáveis. Muitas pessoas, apesar de conhecerem a fundo a área de atividade e o ambiente no qual será aplicado o modelo de *machine learning*, não têm conhecimento de métodos quantitativos e não conseguem expressar-se de forma precisa. Caberá ao analista de dados, em um trabalho paciente, ir transformando todas as informações colhidas em variáveis que possam ser convenientemente utilizadas no projeto.

A identificação das variáveis pode ser facilitada construindo grandes "*famílias de variáveis*" e depois fazendo o desdobramento dentro de cada uma delas. Por exemplo, se o objetivo é agrupar em segmentos homogêneos, empresas clientes de um atacadista, podemos pensar inicialmente em famílias de variáveis, como (a) características sociodemográficas dos clientes (porte, localização, número de sócios etc.), (b) caraterísticas dos produtos adquiridos (higiene, limpeza, alimentos etc.), (c) informações sobre o relacionamento comercial (número médio de pedidos por mês, valor médio dos pedidos, formas de pagamento etc.). Dentro de cada família poderemos, então, com mais facilidade, identificar variáveis que acreditamos afetar o agrupamento.

Além de ouvir pessoas envolvidas com o dia a dia da área na qual será aplicado o modelo, o analista deve investigar se outros estudos similares foram realizados e que variáveis foram utilizadas. Uma busca em jornais científicos ou na Internet pode facilitar tal tarefa.

Em um primeiro momento não devemos omitir informações, ainda que pareçam pouco relevantes. Na identificação das variáveis é melhor pecar pelo excesso. É muito comum que certas variáveis não sejam consideradas em determinado estudo, pelo simples fato que o analista da área de atividade, ainda que muito experiente, não acredita serem importantes. Ou pior, assegura que não tem utilidade. Quando utilizadas posteriormente, podem eventualmente mostrar-se poderosas informações. Outrossim, certas variáveis, historicamente consideradas importantes pelos envolvidos na empresa, mostram-se pouco ou nada relevantes para a melhoria dos resultados do modelo. O folclore empresarial é rico em variáveis desse tipo. Um bom analista deve fugir dessas armadilhas e só tirar conclusões quanto à relevância de uma variável como previsora após analisar os dados. Cabe aqui lembrar a célebre frase do grande estatístico W. Deming: "*Só acredito em Deus. Os demais, apresentem os dados*".

## 1.9 DEFINIÇÃO OPERACIONAL DE UMA VARIÁVEL

Ao trabalhar com a base de dados disponível em uma empresa, a partir da qual o modelo deverá ser desenvolvido, devemos estar atentos à forma como eles foram imputados. Não havendo uma definição clara da variável, distintas pessoas podem interpretar de formas diferentes uma mesma informação, ou seja, os valores de algumas

variáveis na base de dados dependem de quem os imputou. Consideremos alguns exemplos:

- Ao utilizar a variável *renda*, os dados imputados podem corresponder à renda mensal do indivíduo, ou à renda anual, ou à renda familiar etc.
- *Experiência* é uma variável difícil de medir. Alguns declaram o tempo de formado e outros declaram há quanto tempo exercem a atividade atual. Na realidade, quando é o cliente que informa, o nome correto da variável seria *experiência declarada*.
- *Profissão* é outro problema. Suponha que uma pessoa é médica, professora e empresária. Qual a profissão que deverá ser considerada? O ideal em certos modelos é que seja a atividade que gera maior renda, mas isso dificilmente é especificado nos questionários. Então, o que estamos utilizando é a *profissão declarada*.
- *Área do apartamento* não deixa claro se é a área total ou a área útil.
- *Próximo ao metrô*, informação utilizada por alguns analistas para prever o valor de um imóvel, é inútil, pois não fica claro o que é próximo, variando de pessoa para pessoa.

No caso de utilizarmos dados adquiridos de fontes externas, o problema se agrava. Em geral, nesses casos, não temos a definição clara de como as variáveis foram imputadas (se é que uma definição formal existe). Ademais, as definições podem ser bastante diferentes entre duas empresas fornecedoras dessas informações.

Infelizmente, essa preocupação com a definição de uma variável não é usual. Quando recebemos a base de dados disponível na empresa, não há a possibilidade de sanar esse problema a curto prazo, pois teríamos que gerar uma nova base de dados, o que, apesar de ideal, é inviável. Incorreria em alto custo para a empresa e poderia atrasar o projeto por muitos meses, até mesmo anos. Nossa sugestão é começar a trabalhar com a base de dados disponível considerando as variáveis que lá se encontram,[6] cientes que os dados foram imputados, quer pelo cliente quer pela empresa, de acordo com a intepretação dada por cada um (já julgando como *informações imputadas*). No entanto, é sempre preferível remover as variáveis que considerarmos suscetíveis de diferentes interpretações, que podem comprometer a aplicação do modelo a ser obtido.

Faz parte das atribuições de um analista de dados orientar a empresa para contornar essas imperfeições na coleta de novos dados. Isso pode ser conseguido por meio da *definição operacional da variável*, ou seja, a definição não ambígua do que ela significa, como deve ser medida, em que unidades deve ser registrada, como tratar dados em branco (não informados) ou o uso de abreviações, entre outros cuidados. Isso garante que as diferentes pessoas que manipularem esses dados terão uniformidade na

[6] Em poucas palavras, a base "é o que temos"!

interpretação e na sua digitação. Tal iniciativa simplificará a tarefa de combinar diferentes bases de dados e a análise e tratamento das variáveis, incluindo o procedimento com dados omissos ou anomalias. Damos a seguir alguns exemplos:

- Ao coletar a variável *renda*, devemos especificar se é a renda do respondente, a renda familiar, se é mensal ou anual, as unidades monetárias em que devem ser expressas etc. Também deve-se definir como agir nos casos em que a renda não é informada (por exemplo, digitando NI ou -999), e evitar deixar algum campo em branco.
- Ao utilizarmos qualquer índice financeiro, é importante que todos apliquem a mesma fórmula de cálculo. Como agir quando o índice não pode ser calculado com as informações contábeis disponíveis?
- Ao definir a variável *próximo ao metrô,* especificar em metros o que significa "próximo".
- Ao utilizar a variável qualitativa *porte da empresa* (pequena, média ou grande), estabelecer claramente o critério estipulado para essa classificação.

A título de exercício, sugerimos ao leitor que defina o que é um "mau pagador".[7]

## 1.10 AMOSTRAGEM

Ao desenvolver uma regra de classificação, podemos trabalhar com toda a base de dados disponível na empresa ou, quando não for viável, selecionamos uma amostra aleatória dessa base de dados. A seleção da técnica de amostragem adequada é importante. Mesmo quando trabalhamos com toda a base de dados disponível, devemos ter em mente que, em geral, ela é uma amostra da população-alvo (mercado). Por exemplo, supondo que estamos trabalhando com os dados de todos os clientes de um grande banco. Por maior que seja essa base, ela não contempla todos os indivíduos do mercado, potenciais clientes, mas que ainda não possuem conta nesse banco.

Em geral, prefere-se a amostragem aleatória simples (AAS), na qual cada indivíduo da população tem a mesma chance de ser selecionado para compô-la. No entanto, especialmente em problemas de classificação, essa técnica não é indicada. Admitamos que uma das categorias da variável alvo apresenta pequena frequência na população. Por exemplo, a porcentagem de indivíduos portadores de uma determinada doença na população será muito pequena, digamos 5%. Nesse caso, se utilizarmos amostragem aleatória simples teremos aproximadamente 95% dos indivíduos da amostra de uma categoria (não portadores) e apenas aproximadamente 5% da outra categoria (portadores da doença). Esse não balanceamento da amostra pode conduzir as regras de classificação não adequadas.

---

[7] A diretoria de crédito de um grande banco passou várias horas para chegar a um consenso sobre essa definição!

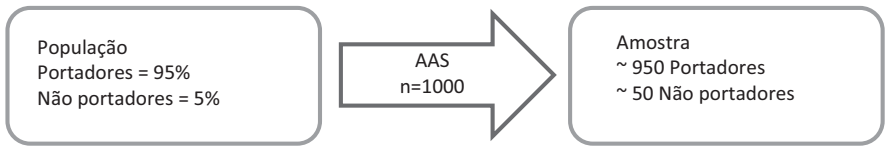


**Figura 1.3 –** Amostragem simples.

Em casos como esses, prefere-se a amostragem estratificada, usualmente denominada em trabalhos de classificação como "amostragem separada". Amostras aleatórias simples de cada categoria de Y são selecionadas separadamente, de forma a garantir um mínimo de indivíduos de cada uma. As amostras separadas não precisam ser do mesmo tamanho. No caso do exemplo anterior, dos portadores da doença, preferimos selecionar uma amostra aleatória simples dentre os não portadores e outra dentre os portadores. As duas suficientemente grandes e de tamanhos similares para garantir a obtenção de um modelo de classificação confiável.

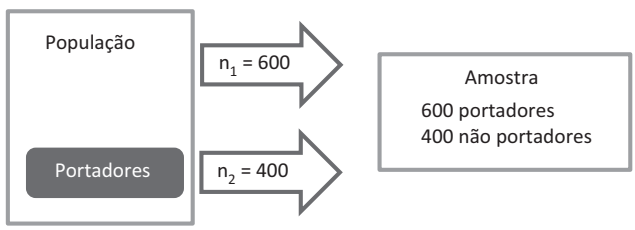


**Figura 1.4 –** Amostragem estratificada.

As proporções de cada categoria da variável alvo na população são denominadas *probabilidades a priori*. No caso, teremos $\pi_{portadores} = 0{,}95$ e $\pi_{naoportadores} = 0{,}05$. Após construir a regra de classificação com as amostras separadas e balanceadas, precisamos fazer certas correções para que os resultados sejam coerentes com a distribuição das categorias na população.[8] Essas correções são função das probabilidades *a priori*. Portanto, essas probabilidades precisam ser conhecidas ou estimadas para adequar a regra de classificação.

O tamanho da amostra depende do tipo de algoritmo a ser utilizado e, em particular, do número de variáveis envolvidas. Quanto maior o número de variáveis maior deverá ser a amostra. Algumas "*regras práticas*" encontradas na literatura não se aplicam a todos os algoritmos. Por exemplo, recomenda-se que para regressão linear múltipla o tamanho da amostra seja pelo menos igual a dez vezes o número de variáveis. Pode até ser interessante para esse algoritmo, mas não se aplica a outros. Os livros de *machine learning* recomendam, frequentemente, o uso de amostras muito grandes, da

[8] Essas correções serão vistas adiante.

ordem de milhares de observações. No entanto, bons resultados podem eventualmente ser obtidos aplicando as técnicas em amostras bem menores, da ordem de centenas de casos. O problema com amostras pequenas é que alguns algoritmos acabam decorando o perfil da amostra fornecendo um excelente modelo que só se aplica à amostra utilizada no desenvolvimento. É o problema de *overfitting* que foi discutido anteriormente. Se pudermos dar uma regra para definir o tamanho de uma amostra ela será: quanto maior a amostra, melhor, desde que não comprometa exageradamente o tempo de processamento.

## 1.11 CASOS PARA ANÁLISE

Nos capítulos que seguem vamos apresentar diferentes técnicas de previsão e classificação. O leitor poderá aplicar essas técnicas para prever ou classificar os indivíduos das bases de dados correspondentes aos casos seguintes.

### 1.11.1 DIRTYSHOP

Esta planilha apresenta os dados de clientes de um magazine. É utilizada no Capítulo 2 para ilustrar a análise exploratória de dados e a correção de não conformidades na base de dados. Contém as variáveis seguintes:

**Tabela 1.5 –** Variáveis do caso DIRTYSHOP

| Variável | Descrição |
|---|---|
| CLIENTE | Código simples de 1 a 1.400 identifica o cliente |
| STATUS | Bom = bom cliente; mau = mau cliente |
| IDADE | Idade em anos completos |
| UNIFED | Unidade da federação em que reside |
| RESID | Tipo de residência em que reside |
| TMPRSD | Tempo de residência em anos completos |
| FONE | 0 = em branco; 1 = sim; 2 = não |
| ECIV | Estado civil |
| INSTRU | Nível educacional |
| RNDTOT | Salários + outros rendimentos (GV$) |
| RST | Restrições creditícias? Sim ou Não |

### 1.11.2 AVALIAÇÕES

As avaliações dos professores de uma escola superior, relativas a diferentes aspectos didáticos e de relacionamento, são realizadas pelos alunos todo fim de semestre. As notas variam de 0 a 10. A planilha AVALIAÇÕES apresenta as médias das avaliações de diferentes professores. Os itens avaliados são:

**Tabela 1.6 –** Variáveis do caso AVALIAÇÕES

| Variável | Descrição |
|---|---|
| AV1 | Uso de recursos audiovisuais |
| AV2 | Disponibilidade de acesso fora da aula |
| AV3 | Didática |
| AV4 | Qualidade do material didático |
| AV5 | Pontualidade |
| AV6 | Relacionamento com os alunos |
| AVGLOB | Avaliação global do curso |

### 1.11.3 XZCALL

A empresa XZCALL é um *Call Center* com aproximadamente nove mil atendentes. Um dos problemas que preocupa o diretor de RH é o processo de recrutamento de novos funcionários. Boa parte dos atendentes contratados permanece pouco tempo na empresa, não compensando o investimento em seu treinamento. Alguns são dispensados pouco tempo depois da contratação decorrente do baixo desempenho.

A empresa dispõe de uma base de dados de funcionários contratados em passado recente, classificados como "bom", quando permaneceram na empresa por 12 meses completos ou mais, ou como "mau", quando permaneceram menos de 12 meses, quer por pedir afastamento, quer por serem demitidos por mau desempenho. Deseja-se construir um modelo que permita classificar candidatos ao emprego de atendente como prováveis "bom" ou "mau". Na realidade, mais que classificar os candidatos, quer-se criar um escore para os candidatos de forma que possam ser classificados em diferentes faixas, de acordo com a probabilidade de se tornarem bons atendentes.

As informações disponíveis são:

**Tabela 1.7 –** Variáveis do caso XZCALL

| Variável | Definição |
|---|---|
| FUNCIONÁRIO | Identificação na amostra |
| STATUS | Bom ou mau |
| UF | Local de nascimento |
| ECIV | Estado civil |
| DIST_EMP | Distância residência – empresa |
| TIPORESID | Tipo de residência |
| PRIM_EMP | Primeiro emprego como atendente? |
| TESTE | Nota em teste psicotécnico |
| EDUC | Nível educacional |

## 1.12 TECAL

A TECAL é uma operadora de telefonia celular que, com a recente entrada de outras operadoras no mercado, tem perdido muitos assinantes por conta da migração.

A direção da TECAL pretende implantar um plano de ação para reduzir essa migração, oferecendo, com suficiente antecedência, benefícios e planos especiais àqueles assinantes que apresentarem alta probabilidade de cancelar sua assinatura. Seu problema é identificar esses clientes a tempo de retê-los e evitar o *churning*. A TECAL acredita que uma antecedência de seis meses é suficiente.

Com base em uma amostra aleatória de dois mil clientes, selecionada de forma a poder prever o cancelamento com seis meses de antecedência e considerando uma série de informações sobre eles, deseja-se construir um modelo para prever com seis meses de antecedência se um assinante se desligará da TECAL. As informações disponíveis para construção do modelo de classificação são as seguintes:

**Tabela 1.8 –** Variáveis do caso TECAL

| Variável | Descrição |
|---|---|
| id | Identificação do assinante |
| idade | Idade em anos completos do assinante |
| linhas | Número de linhas do assinante |
| temp_cli | Tempo como assinante em meses |
| renda | Renda familiar do assinante em reais |
| fatura | despesa média mensal do assinante em reais |
| temp_rsd | Tempo na residência atual do assinante, em anos |
| local | Região onde reside o assinante (A, B, C e D) |
| tvcabo | Assinante possui TV a cabo? |
| debtaut | Pagamento em débito automático? |
| cancel | Assinante cancelou contrato? (variável alvo) |

### 1.12.1 BETABANK

Uma amostra aleatória de clientes que receberam financiamento do BETABANK para compra de automóveis foi selecionada aleatoriamente. Clientes que atrasaram alguma parcela do pagamento por mais que 90 dias, ou que em um período de seis meses atrasaram dois ou mais pagamentos pelo menos 30 dias, são considerados maus pagadores. Aproximadamente, 65% dos clientes foram classificados como bons pagadores. A partir dessa amostra, deseja-se obter um modelo de *credit scoring* para prever a probabilidade de um novo solicitante de crédito ser mau pagador, caso o crédito seja concedido. Os dados foram coletados na data do financiamento. As variáveis disponíveis são:

**Tabela 1.9 –** Variáveis do caso BETABANK

| Variável | Descrição |
|---|---|
| Cliente | Identificação do cliente na amostra |
| ECIV | Estado civil |
| ESCOLARIDADE | Nível de escolaridade do cliente |
| IDADE | Idade do cliente ao contratar o empréstimo |
| NATUREZA | Natureza da ocupação do cliente (vide tabela seguinte) |
| PROFISSAO | Código na declaração do imposto de renda |
| SEXO | Sexo do cliente |
| RENDA | Renda mensal do cliente em reais |
| UF | Unidade da federação onde reside |
| STATUS | Bom ou mau cliente |

**Tabela 1.10 –** Categorias da variável Natureza da Ocupação

| Natureza da ocupação | |
|---|---|
| 1 | Funcionário de empresa privada |
| 2 | Sócio de empresa |
| 3 | Vive de renda |
| 4 | Funcionário público |
| 5 | Autônomo/Profissional liberal |
| 6 | Aposentado |
| 9 | Outros |

### 1.12.2 KIMSHOP

A KIMSHOP é uma loja especializada em rações para pets que financia suas vendas. A partir de sua base de dados deseja desenvolver um modelo de *credit scoring*, para avaliar o risco de novos solicitantes de crédito. As variáveis da base de dados são as seguintes:

**Tabela 1.11 –** Variáveis do caso KIMSHOP

| Variável | Descrição |
|---|---|
| TIPO | 1 bom pagador; 2 mau pagador |
| IDADE | Idade em anos completos |
| REGIAO | Região em que reside |
| RESID | Tipo de residência: 0 – não informado; 1 – própria; 2 – alugada; 3 – outros |
| DEPEND | número de dependentes |
| ECIV | Estado civil: 1 – casado; 2 – solteiro; 3 – divorciado; 4 – viúvo; 5 – outros |
| INSTRU | Nível educacional: 0 – não informado; 1 – fundamental; 2 – médio; 3 – superior |
| RNDTOT | Salário e outros rendimentos mensais em reais |
| DSB | Possui desabonos (SPC, Serasa, Boa Vista...) |

### 1.12.3 BUXI

Livrarias BUXI é uma cadeia de livrarias que tem quiosques nos principais supermercados das grandes capitais brasileiras. Em janeiro de 2017, começou a financiar as compras de livros em até quatro parcelas. Ademais, nesse mesmo mês a BUXI implantou seu site de vendas pela Internet.

A empresa decidiu construir um modelo de scoring de crédito para melhorar seus resultados na concessão de crédito. Além da inadimplência, cerca de 20% das solicitações de crédito não eram aprovadas, o que em muitos casos implicava na perda da venda.

Para desenvolver o modelo, coletou-se uma amostra aleatória de 2.600 clientes, cujo financiamento foi efetivado no período de janeiro a dezembro de 2018. A amostra mostrou grande desbalanceamento entre os dois tipos de clientes: 2.400 eram bons

pagadores e 200 maus pagadores. Se um cliente teve mais de um financiamento aprovado nesse período, considerou-se apenas as informações relativas ao último financiamento. A data do último financiamento é denotada como DF.

As informações disponíveis na base de dados da BUXI são as seguintes:

**Tabela 1.12 –** Variáveis do caso BUXI

| Variável | Descrição |
|---|---|
| STATUS | Caracterização do cliente (bom ou mau pagador) |
| IDADE | Idade do cliente em anos completos na data DF |
| UNIFED | UF em que reside o cliente na data DF |
| RESID | Tipo de residência do cliente na data DF |
| PRIM | Primeira compra do cliente na BUXI? |
| INSTRU | Grau de instrução do cliente na data DF |
| CARTAO | Cliente possuía cartão de primeira linha na data DF |
| RESTR | Possuía desabonos financeiros na data DF? (Informação fornecida por bureau externo.) |
| QUANTI | Cliente comprou mais de dois livros nessa operação? |
| NET | A compra foi realizada pela internet? |

### 1.12.4 PASSEBEM

Os principais produtos da empresa de turismo PASSEBEM são os pacotes AA, BB e CC. Uma amostra de clientes que compraram esses pacotes nos últimos dois anos será utilizada para desenvolver um modelo que permita prever que tipo de pacote oferecer a um novo cliente, bem como para melhor direcionar campanhas de marketing. A amostra aleatória de dez mil clientes contém as seguintes informações:

**Tabela 1.13 –** Variáveis do caso PASSEBEM

| Variável | Descrição |
|---|---|
| PACOTE | AA, BB, CC (último pacote adquirido pelo cliente) |
| IDADE | Idade do cliente em anos completos |
| UNIFED | Unidade da federação em que reside o cliente |
| RESID | Tipo de residência em que reside o cliente |
| ECIV | Estado civil do cliente |
| INSTRU | Grau de instrução do cliente |
| RNDTOT | Renda familiar mensal do cliente em reais |
| PAG_VISTA | Se cliente pagou ou não à vista |

### 1.12.5 CAR UCI[9]

Na planilha de dados CAR UCI, extraída do site da *UC Irvine Machine Learning Repository,* vários autos são avaliados quanto a diferentes características. A variável alvo é Classificação. Objetiva-se classificar outros autos em função de suas caraterísticas. A amostra é fortemente desbalanceada.

[9] Dua, D. and Graff, C. (2019). *UCI Machine Learning Repository* [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science at https://archive.ics.uci.edu/ml/datasets/car+evaluation.

**Tabela 1.14 –** Variáveis do caso CAR UCI

| Variável | Descrição |
|---|---|
| Classificação | Aceitabilidade: unacc, acc, good, vgood |
| Buying | Preço de compra: vhigh, high, medium, low |
| Maint | Custo de manutenção: vhigh, high, medium, low |
| Doors | Número de portas |
| Persons | Número de passageiros |
| Lug_boot | Tamanho do porta-malas: small, medium, big |
| Safety | Segurança: low, medium, high |

## 1.12.6 SUPERMERCADO IMPÉRIO

Uma amostra de 80 funcionários do SUPERMERCADO IMPÉRIO foi selecionada aleatoriamente. A partir dos dados disponíveis, descritos na tabela seguinte, deseja-se construir um modelo para prever o salário de um funcionário.

**Tabela 1.15 –** Variáveis do caso SUPERMERCADO IMPÉRIO

| Variável | Descrição |
|---|---|
| ID | Identidade do funcionário na amostra |
| EDUCAÇÃO | Nível educacional do funcionário |
| CARGO | Cargo do funcionário |
| LOCAL | Local onde atua o funcionário |
| IDADE | Idade em anos completos do funcionário |
| TEMPOCASA | Tempo de casa do funcionário, em anos completos |
| SALARIO | Salário mensal do funcionário em G$ |

### 1.12.7 SPENDX

Uma amostra aleatória de clientes do Cartão SPENDX, emitido pelo Banco SPENDX, foi selecionada para construir um modelo que permite prever o valor médio da fatura mensal em um período de doze meses. O modelo de previsão será desenvolvido utilizando as informações seguintes:

**Tabela 1.16 –** Variáveis do caso SPENDX

| Variável | Descrição |
|---|---|
| ID | Código do cliente na amostra |
| renda | Renda familiar mensal do cliente em reais |
| tempo | Há quanto tempo possui o cartão |
| classe | Classificação do cliente em função de seu relacionamento com o banco |
| cartões | Quantos cartões possui (titular e dependentes) |
| idade | Idade do cliente em anos completos |
| sexo | Sexo do cliente |
| propria | Cliente possui casa própria? |
| superior | Cliente tem curso superior? |
| UF | UF onde reside o cliente |
| fatura | Média do valor das faturas durante os últimos 12 meses |

### 1.12.8 2005 CAR DATA

Estes dados correspondem a centenas de carros GM de 2005, utilizados em artigo publicado por *Shonda Kuiper* no *Journal of Statistical Education,* em 2008.[10] Todos os carros dessa base tinham menos que um ano de uso. Um modelo deve ser construído para prever o preço a partir de caraterísticas técnicas dos automóveis, conforme especificadas a seguir. Sugerimos a não utilização das variáveis *Model* e *Trim* para simplificar o trabalho.

[10] Uso autorizado pela autora. O artigo pode ser encontrado em https://www.tandfonline.com/doi/full/10.1080/10691898.2008.11889579?needAccess=true.

**Tabela 1.17 –** Variáveis do caso 2005 CAR DATA

| Variável | Descrição |
|---|---|
| Price | Preço sugerido para o carro em excelentes condições |
| Mileage | Milhagem do carro |
| Make | Fabricante (Saturn, Pontiac e Chevrolet) |
| Model | Modelo do carro |
| Trim | Tipo do carro |
| Type | Tipo da estrutura do carro (sedan, cupê, ...) |
| Cylinder | Cilindrada do carro |
| Liter | Especificação do motor do carro |
| Doors | Número de portas |
| Cruise | Indica se tem piloto automático (1 = sim) |
| Sound | O carro tem sistema de alto-falantes especiais (1 = sim) |

### 1.12.9 AUTOMPG

Este conjunto de dados foi extraído da *UCI Machine Learning Repository.*[11] O objetivo é prever o consumo de um carro (*mpg*) a partir de caraterísticas especificadas na tabela seguinte:

[11] Dua, D. and Graff, C. (2019). *UCI Machine Learning Repository* [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. Disponível em: https://archive.ics.uci.edu/ml/datasets/auto+mpg.

**Tabela 1.18 –** Variáveis do caso AutoMPG

| Variável | Descrição |
|---|---|
| Mpg | Consumo (milhas por galão) na cidade |
| Cylinders | Número de cilindros (variável discreta) |
| Displacement | Cilindrada (variável contínua) |
| Horsepower | Cavalos de força |
| Weight | Peso |
| Acceleration | Aceleração |
| Year | Ano |
| Origin | Origem |
| Car name | Nome do carro |

## 1.12.10 WORLD HAPPINESS REPORT

Dados extraídos do Kaggle[12] relacionam o grau (escore) de felicidade de cada país calculado a partir dos indicadores: GDP per Capita, Family, Life Expectancy, Freedom, Generosity, Trust Government Corruption. O objetivo é encontrar um modelo capaz de prever o escore de felicidade a partir das demais variáveis. A detecção da importância de cada atributo no cálculo do escore é um dos principais objetivos após a construção do modelo de previsão.

## 1.12.11 HAPPINESS ALCOHOL COMSUMPTION

Consideremos o arquivo *HAPPINESS AND ALCOHOL CONSUMPTION,* extraído do Kaggle.[13] Nosso objetivo será prever o grau de felicidade de cada país (escala de 0 a 10) a partir dos indicadores IDH, PIB per capita e consumo per capita, de cerveja, drinks e vinho em cada país. O PIB per capita e o consumo de vinho foram transformados via logaritmo natural eliminando a forte assimetria.

[12] https://www.kaggle.com/unsdsn/world-happiness (dados de domínio público).

[13] https://www.kaggle.com/marcospessotto/happiness-and-alcohol-consumption.

### 1.12.12 MINIMARKET MM

MM é uma franquia de pequenas mercearias espalhadas pelo país. Atualmente, são 74 lojas com mais de um ano de funcionamento. O diretor da empresa deseja agrupar as lojas em *clusters* a partir das informações seguintes:

**Tabela 1.19 –** Variáveis do caso MINIMARKET MM

| Variável | Descrição |
|---|---|
| Zona | Região onde se localiza a loja |
| Idade | Idade da loja: classe 1 = 1 a 5 anos, classe 2 = 6 a 10 anos, classe 3 = mais de 10 anos |
| Lucro | Lucro no ano anterior em unidades monetárias |
| Faturamento | Faturamento no ano anterior em unidades monetárias |
| Metas | Percentual da meta cumprida pela loja |
| Funcionários | Número de funcionários |

Após agrupar as lojas, o diretor deseja saber se há alguma relação entre os diferentes *clusters* encontrados e cada uma das variáveis seguintes: a idade do gerente (*idade gerente*) e o fato de ter ou não estacionamento (*estacionamento*).

### 1.12.13 MUNICÍPIOS

Este arquivo de dados apresenta as caraterísticas dos 5.565 municípios brasileiros em 2010.[14] Nosso objetivo é obter grupos homogêneos de municípios em função dos indicadores seguintes:

[14] http://www.atlasbrasil.org.br/consulta/planilha.

**Tabela 1.20 –** Variáveis drivers do caso

| Variável | Descrição |
|---|---|
| IDHM_E | IDH municipal – dimensão educação |
| IDHM_L | IDH municipal – dimensão longevidade |
| IDHM_R | IDH municipal – dimensão renda |
| GINI | Índice de Gini |
| PIND | PIND – proporção de extremamente pobres |

Após obter os *clusters*, verifique se é possível batizar cada *cluster* conquistado em função do comportamento das variáveis acima.

Posteriormente, analisar o comportamento das variáveis seguintes nos diferentes *clusters* obtidos.

**Tabela 1.21 –** Variáveis descritivas do caso MUNICÍPIOS

| Variável | Descrição |
|---|---|
| ESPVIDA | Esperança de vida ao nascer |
| RDPC | Renda per capita |
| E_ANOSESTUDO | Expectativa de anos de estudo aos 18 anos de idade |
| T_LUZ | % da população que vive em domicílios com energia elétrica |
| ÁGUA_ESGOTO | % de pessoas em domicílios com abastecimento de água e esgotamento sanitário inadequados |

## 1.12.14 HEALTHSYSTEMS (DADOS DO BANCO MUNDIAL)

Base de dados[15] para agrupar países em *clusters* com as informações seguintes, relativas à área de saúde. O arquivo original foi editado para uso em sala de aula.

Para agrupar os países, utilizar as variáveis (*drivers*) da tabela seguinte.

[15] www.kaggle.com/danevans/world-bank-wdi-212-health-systems/data.

**Tabela 1.22 –** Variáveis drivers do caso HEALTHSYSTESMS

| Variável | Descrição |
|---|---|
| WB name | *World_Bank_Name*: refere-se ao nome dos países |
| H1 | Gasto do PIB em saúde em 2016 |
| H2 | Porcentagem de gasto público em saúde em 2016 |
| H3 | Gastos em saúde, em USD, feitos diretamente pelas famílias em 2016 |
| H4 | Gasto per capita em saúde, em USD, em 2016 |
| H5 | Porcentagem de recursos externos em saúde, compostos de investimentos estrangeiros diretos e transferências externas de fontes públicas e privadas em 2016 |

Analisar o comportamento das variáveis descritivas seguintes nos *clusters* obtidos.

**Tabela 1.23 –** Variáveis descritivas do caso HEALTHSYSTEMS

| Variável | Descrição |
|---|---|
| Phys | Médicos (clínicos gerais e especialistas) para cada 1.000 habitantes em 2016 |
| Nurse | Enfermeiras para cada 1.000 habitantes em 2016 |
| Surgic | Especialistas cirurgiões para cada 1.000 habitantes em 2016 |

## 1.12.15 UN NATIONAL STATS

Estes dados foram extraídos do package carData do software R.[16] Contém informações de diferentes países, obtidos a partir das bases das Nações Unidas. Os países devem ser agrupados utilizando as técnicas de *cluster analysis*. As variáveis estão listadas a seguir. Todas devem ser utilizadas para determinar os *clusters*, analisar e caracterizar os *clusters* obtidos:

[16] Os dados foram coletados via http://unstats.un.org/unsd/demographic/ products/socind on April 23, 2012. OECD membership is from https://www.oecd.org/, accessed May 25, 2012. Dados utilizados no livro: Weisberg, S. (2014). Applied Linear Regression, 4th edition. Hoboken NJ: Wiley.

**Tabela 1.24 –** Variáveis drivers do caso UN NATIONAL STATS

| Variável | Descrição |
|---|---|
| Country | País |
| Region | Região geográfica |
| Group | Especifica o grupo a que pertence o país (OECD, África, outros) – 2012 |
| Fertility | Taxa de fertilidade (número de crianças/mulher) |
| ppgdp | Produto interno bruto em US$ |
| lifeExpF | Esperança em anos ao nascer, para o sexo feminino |
| pctUrban | Porcentagem de habitantes em áreas urbanas |
| infantMortality | Crianças que falecem até um ano de vida por 1.000 nascimentos |

## 1.12.16 SUPERMERCADOS GUDFUD

A empresa GUDFUD entrega produtos alimentícios por meio de pedidos pelo telefone ou internet. Para melhor entendimento, seu mercado decidiu segmentar os clientes de acordo com os tipos de produtos solicitados, classificados em três categorias: hortifruti, carnes (incluindo aves e peixes) e laticínios. Uma amostra dos clientes que realizaram seis ou mais pedidos nos últimos três meses foi selecionada aleatoriamente. Os valores médios dos pedidos, em reais, para cada uma dessas categorias encontram-se no arquivo gudfud.xlsx.

**Tabela 1.25 –** Variáveis do caso SUPERMERCADOS GUDFUD

| Variáveis | Descrição |
|---|---|
| Cliente | Identificação do cliente |
| Hortifruti | Valor das compras nos últimos três meses em u.m. |
| Carnes | Valor das compras nos últimos três meses em u.m. |
| Laticínios | Valor das compras nos últimos três meses em u.m. |
| Canal | Canal pelo qual cliente realiza as compras |
| Pedidos | Número de pedidos nos últimos três meses |
| Sexo | Sexo do cliente |

Inicialmente, devemos segmentar os clientes considerando apenas as variáveis hortifruti, carnes, laticínios. Após a segmentação, devemos descrever as diferenças entre os *clusters,* considerando todas as variáveis da base de dados.

A segmentação deve ser repetida considerando as variáveis hortifruti, carnes, laticínios, canal e número de pedidos.

**O desenvolvimento de modelos de análise aplicados à realidade empresarial se tornou competência crítica em todos os setores de atividade econômica e os profissionais enfrentam duas questões. Como geraremos valor a partir dos dados? Como usaremos corretamente os métodos de análise? Existe um crescente arsenal de técnicas de *Machine Learning*, provocando efeitos colaterais indesejáveis, como a adoção de técnicas apenas por parecerem modernas.**

Neste livro, Abraham Laredo Sicsú, André Samartini e Nelson Lerner Barth expõem as técnicas obrigatórias, prática e objetivamente, suplantando as barreiras para o aprendizado de *Data Science*. Apresentam de forma equilibrada e rigorosa tanto os conceitos fundamentais quanto os exemplos de aplicação prática.

ISBN 978-65-5506-396-7

www.blucher.com.br

Blucher


Clique aqui e:

VEJA NA LOJA

# Técnicas de machine learning

Abraham Laredo Sicsú, André Samartini, Nelson Lerner Barth

ISBN: 9786555063967
Páginas: 394
Formato: 17 x 24 cm
Ano de Publicação: 2023