INTELIGÊNCIA ARTIFICIAL

Prof. Luciano Frontino de Medeiros

TEMA 1 – APRENDIZADO DE MÁQUINA

A ideia de máquinas inteligentes, que tenham alguma capacidade de aprendizado durante a sua operação, não é antiga no campo da Inteligência Artificial. Ainda em 1950, Turing já havia considerado em realmente se ter máquinas inteligentes, as quais pudessem ser ensinadas. O aprendizado é agora o método preferencial de se criar agentes inteligentes, tendo ainda a vantagem de se tornar mais competente do que apenas tendo o seu conhecimento prévio (Russel; Norvig, 2004, p. 51).

O *aprendizado de máquina computacional* se refere à aplicação de técnicas computacionais na busca de padrões que porventura estejam ocultos em um conjunto de dados (Amaral, 2016, p. 81). *Machine learning*, como mencionado no original em inglês, é um conjunto de ferramentas estatísticas ou geométricas e de algoritmos que permitem a automatização da construção de uma função de predição a partir de um conjunto de observações chamados de *conjunto de treinamento* (Lemberger et al, 2015, p. 110).

A noção chave por trás do conceito de aprendizado de máquina é a predição. Prever o futuro é um dos sonhos mais antigos da humanidade. Desde a antiguidade, infinitos recursos de fantasia e criatividade estão neste projeto com motivações muito variadas: prever o comportamento de um adversário antes de um combate militar, antecipar a chegada da chuva para resolver a semeadura e a colheita, prever eclipses do sol, dentre outros (Lemberger et al, 2015, p. 108).

Nos dias atuais, existe a preocupação no mundo empresarial da obtenção de vantagem competitiva frente aos concorrentes, buscando antecipar o comportamento dos clientes. Desta forma, alguns dos usos mais comuns do aprendizado de máquina no âmbito empresarial são (Lemberger et al, 2015, p. 108):

- Detecção de comportamentos de fraude em transações financeiras *online*;
- Estimar uma taxa de conversão em um *site* comercial com base no número de cliques em determinadas páginas;
- Prever os riscos de insolvência de um cliente com base em seus recursos e perfil socioprofissional;
- Antecipar intenções de encerrar um serviço com base nas atividades de um assinante;

- Descobrir as preferências de um cliente que queremos manter para sugerir produtos e serviços adequados aos seus gostos e necessidades.

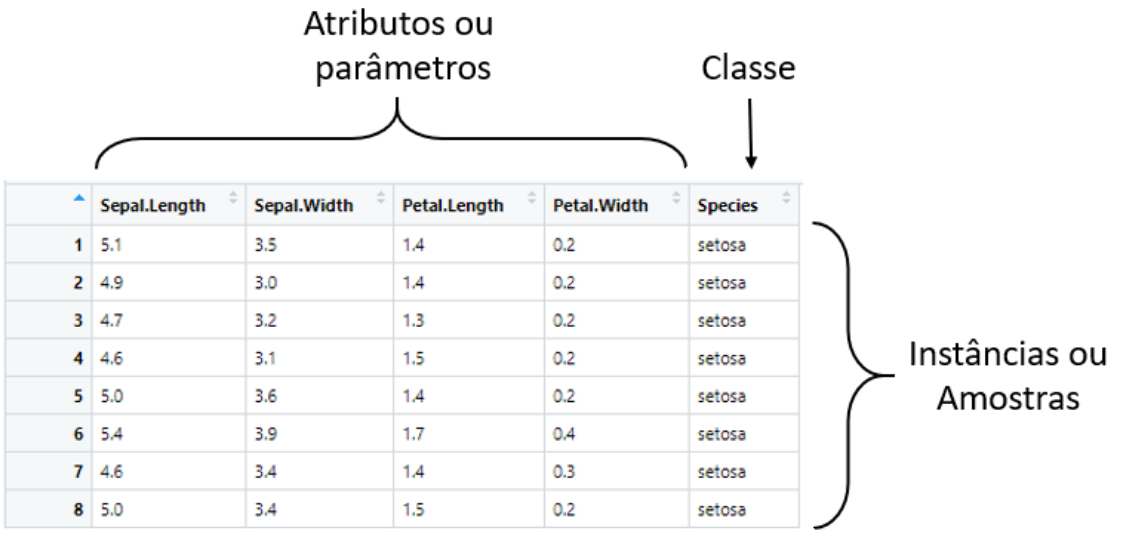
Predição é diferente de compreensão. Tendo observado pacientemente o movimento dos planetas, os astrônomos da Antiguidade deduziram dele, com grande ingenuidade, é preciso reconhecer, as leis das recorrências de eclipses. No entanto, a ciência não para por aí, pois tem a ambição não apenas de prever, mas também de entender o fenômeno observado, por meio de um modelo explicativo. Assim, as leis da gravitação descobertas por Newton séculos mais tarde forneceram um modelo explicativo para as previsões dos primeiros astrônomos. O equivalente a leis determinísticas da gravitação, em um contexto comercial, pode ser um sistema de regras de negócios destinadas ao cálculo do preço de venda de um produto de acordo com parâmetros, como a margem desejada, o tempo de fabricação e o preço dos componentes (Lemberger et al, 2015, p. 108).

Entretanto, em um contexto social como o do mundo dos negócios, tal abordagem das ciências determinísticas é impraticável se deseja-se fazer previsões. O aprendizado de máquina faz parte dessa estrutura. Cada observação passada de um fenômeno é descrita usando dois tipos de variáveis: variáveis preditivas (ou atributos ou parâmetros), das quais esperamos poder fazer previsões. As variáveis preditivas *p* associadas a uma observação serão anotadas como um vetor x=(x_1,x_2,…,x_p) com componentes *p*. Um conjunto de N instâncias ou amostras será constituído por N vetores x^((1)),x^((2)),…,x^((N)). E as variáveis alvo, cujo valor queremos prever para eventos ainda não observados, que conterão a classe (Figura 1) (Lemberger et al, 2015, p. 109).

A partir dessa definição, pode-se afirmar que a análise de dados mediante o aprendizado de máquina visa a construção de um modelo. Um modelo de aprendizado de máquina é um processo algorítmico específico que permite a construção de uma função de predição *f* a partir de um conjunto de dados de aprendizado.

A construção de *f* constitui o aprendizado ou o treinamento do modelo. Uma previsão corresponde à avaliação f(x) da função de predição *f* nas variáveis preditivas de uma observação *x*. Pode-se também denotar os valores da função *f* com a variável *y* (Lemberger et al., 2015, p. 110).

Figura 1 – Designação de atributos ou parâmetros, instâncias ou amostras e classes, a partir de uma seção do banco de dados de íris



	Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
1	5.1	3.5	1.4	0.2	setosa
2	4.9	3.0	1.4	0.2	setosa
3	4.7	3.2	1.3	0.2	setosa
4	4.6	3.1	1.5	0.2	setosa
5	5.0	3.6	1.4	0.2	setosa
6	5.4	3.9	1.7	0.4	setosa
7	4.6	3.4	1.4	0.3	setosa
8	5.0	3.4	1.5	0.2	setosa

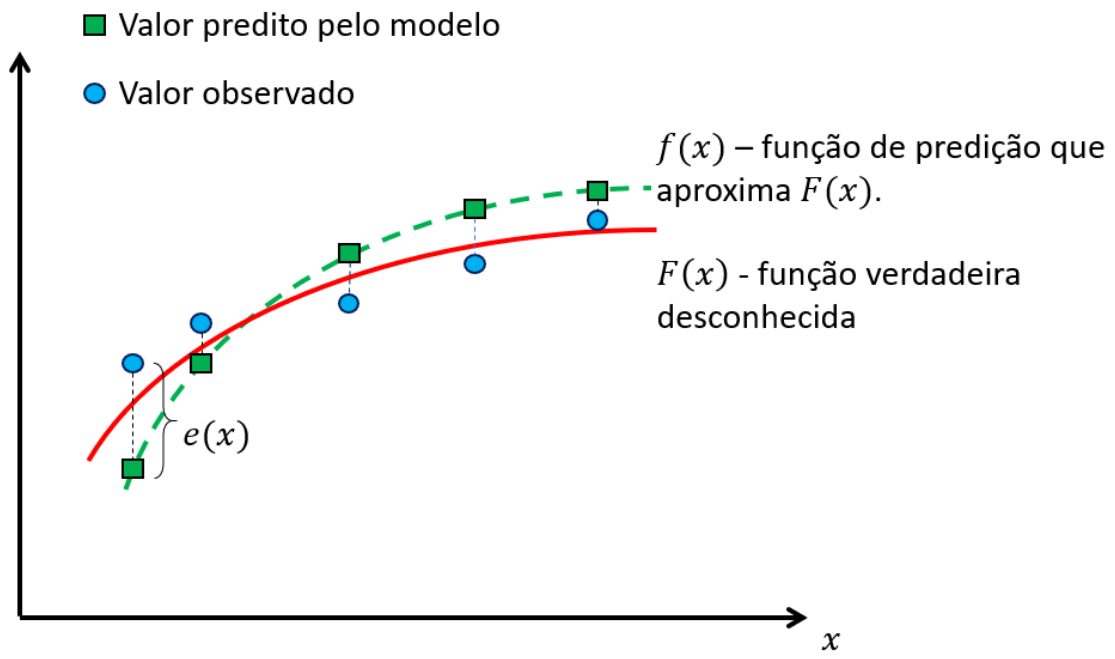
Fonte: Adaptado de Amaral, 2016, p. 84

Pode-se considerar esquematicamente que o valor observado *y* da variável alvo resulta da superposição de duas contribuições:

- Uma função F(x) das variáveis preditoras, da qual se aproxima f(x). Portanto, é uma contribuição inteiramente determinada pelas variáveis preditivas x da observação.
- Um erro aleatório e(x). É uma quantidade que engloba os efeitos combinados de um grande número de demais parâmetros, impossíveis de levar em consideração.

Dessa forma, o valor de uma variável alvo é a soma de uma função determinística (real, porém desconhecida) F(x) e de uma função erro aleatório e(x). O objetivo do aprendizado de máquina é de fornecer uma boa aproximação de F(x) a partir das observações ou amostras. A aproximação f(x) é a função de predição que permite obter as estimativas f(x) de F(x). (Lemberger et al., 2015, p. 110).

Figura 2 – O modelo f(x) com estimativas relativas às observações, como uma aproximação da função verdadeira e desconhecida F(x)



Fonte: Adaptado de Lemberger et al., 2015, p. 110

A construção de um modelo requer três fases distintas (Lemberger et al, 2015, p. 111):

1. A seleção do algoritmo de aprendizado de máquina a partir de uma biblioteca de algoritmos disponíveis. Esta biblioteca geralmente está presente em uma ferramenta disponível para a análise de dados como R ou SPSS;
2. O treinamento do algoritmo escolhido a partir dos dados, produzindo a função de predição f(x);
3. A predição propriamente dita com o modelo construído, a partir da inserção de novas observações que não fizeram parte do conjunto de dados de treinamento.

Um cientista de dados também pode ser requisitado para orientar o processo de aprendizagem, escolhendo certos dados que são mais significativos que outros, uma fase durante a qual a visualização de dados multidimensionais desempenha um papel importante. O *aprendizado de máquina* é, portanto, uma abordagem fundamentalmente orientada por dados, sendo útil para criar sistemas de suporte à decisão que se adaptem aos dados e para os quais nenhum algoritmo de processamento esteja disponível (Lemberger et al., 2015, p. 111).

Quanto aos tipos de tarefas de aprendizagem de máquina, pode-se listar três: classificação, agrupamentos e regras de associação. Quando a variável alvo é uma variável nominal (como no caso do banco de dados de íris) a tarefa do aprendizado de máquina é a descoberta das classes a que pertencem as amostras, ou seja, a classificação propriamente dita. Quando a variável é numérica, a tarefa a ser executada é a regressão (Amaral, 2016, p. 87).

A partir das tarefas, a construção de um modelo implica a escolha de técnicas e algoritmos de aprendizado de máquina. A técnica é diferente do algoritmo. Uma *técnica* é uma forma de resolver uma tarefa de aprendizado, enquanto o *algoritmo* é a implementação da técnica (Amaral, 2016, p. 87). Por exemplo, pode-se utilizar a técnica de redes neurais artificiais para o aprendizado de máquina, e um algoritmo que implementa redes neurais é o perceptron multicamada (MLP).

Outro conceito importante para a compreensão das tarefas de aprendizado de máquina diz respeito à supervisão das tarefas. Um aprendizado supervisionado considera as classes as quais as amostras são definidas previamente, e um aprendizado não supervisionado não tem uma classe prévia das amostras. As tarefas de classificação ou regressão são supervisionadas, enquanto os agrupamentos ou regras de associação são tarefas não supervisionadas (Amaral, 2016, p. 87).

TEMA 2 – PROCESSOS DE MINERAÇÃO DE DADOS

A atividade de mineração de dados em uma empresa com disposição a investir na ciência de dados como alternativa para alcançar vantagem competitiva pressupõe uma série de fases, indo desde a compreensão do negócio até a implementação de soluções. Dois padrões de processos podem ser utilizados: o CRISP-DM – Cross Industry Standard Process for Data Mining – e o KDD – Knowledge-Discovery in Databases (Amaral, 2016, p. 84).

CRISP-DM traduz-se para Processo Padrão Genérico para Mineração de Dados. É o mais conhecido e adotado, prevendo seis fases (Amaral, 2016, p. 85-86):

- **Entendimento do negócio**: inicialmente, deve-se compreender as características do negócio onde a mineração de dados será utilizada. Consiste numa etapa chave para o sucesso de todo o processo.

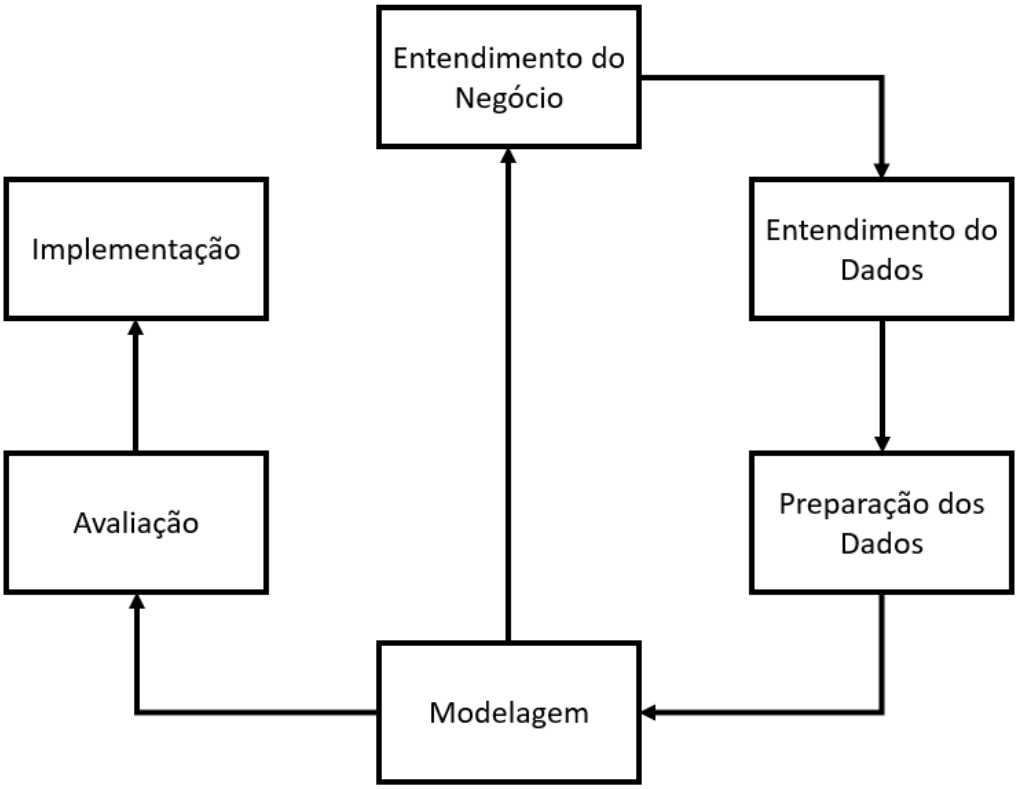
- **Entendimento dos dados**: após o entendimento do negócio, a fase a seguir se ocupa dos dados necessários à mineração em termos de alguns elementos: estrutura, relacionamentos, qualidade, quantidade e acesso aos dados.
- **Preparação dos dados**: a aplicação de qualquer algoritmo de aprendizado de máquina requer que os dados estejam devidamente organizados, selecionados e limpos. A preparação ainda envolve tarefas como discretização (os dados são transformados em nominais) e a normalização (dados bem comportados, dentro de uma faixa determinada).
- **Modelagem**: o produto da tarefa de aprendizado de máquina é um modelo. Este modelo será utilizado para classificar novas amostras, dados que não foram alimentados ao modelo durante a fase de treinamento.
- **Avaliação**: O modelo é testado quanto ao seu desempenho, a partir de critérios a serem satisfeitos.
- **Implementação**: o processo de mineração é implantado, fazendo parte dos processos da organização.

O outro modelo, KDD, traduzido como “descoberta de conhecimento em banco de dados”, provém da área da gestão do conhecimento e está dividido em cinco fases:

- **Entendimento do Negócio**: semelhante ao entendimento do negócio visto no padrão anterior.
- **Pré-processamento**: equivalente ao entendimento dos dados no processo CRISP-DM.
- **Transformação**: semelhante à preparação dos dados no modelo anterior.
- **Mineração de Dados**: equivalente à fase de modelagem do CRISP-DM.
- **Interpretação e avaliação**: equivalente às fases de avaliação e implementação do CRISP-DM.

Pode-se notar a equivalência existente entre os dois padrões para mineração de dados.

Figura 3 – O processo de mineração de dados CRISP-DM



Fonte: Amaral, 2016, p. 84

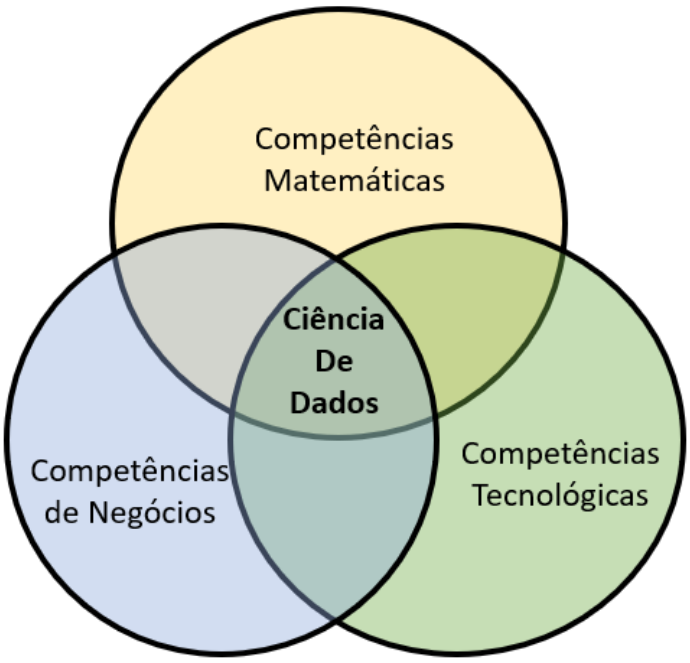
O processo de mineração de dados tem como protagonista o cientista de dados. O *cientista de dados* é uma profissão recente, que surgiu há alguns anos de "Big Data" e cujos contornos ainda são bastante vagos. Muitos debates acontecem sobre o leque de habilidades que o caracterizam, sobre a diferença com os perfis dos estatísticos "clássicos", seu lugar na organização, sua vida cotidiana. Essa profissão apareceu em um contexto de profunda evolução tecnológica, maior disponibilidade e baixo custo de poder computacional e explosão das fontes de dados disponíveis (Lemberger et al., 2015, p. 73).

Contribuem de forma significativa para o contexto das competências do cientista de dados três dimensões (Lemberger et al., 2015, p. 74):

- **Dimensão matemática / estatística**: obviamente, é nessa dimensão que pensamos em primeiro lugar. O conhecimento de estatística e algoritmos de aprendizado de máquina é essencial. O cientista de dados deve ser capaz de entender um conceito como o nível de significância de um teste, corrigir vieses, calcular probabilidades etc.

- **Dimensão tecnológica / informática**: É aqui que o termo cientista de dados se destaca. Em seu trabalho, o cientista de dados raramente está satisfeito com o *software* tradicional: planilhas, suítes tradicionais de BI (geração de relatórios / exploração) ou mesmo *software* de análise e estatística de dados. A explosão de volumes de dados e a disponibilidade de inúmeras estruturas de código aberto destinadas a operar transformações e enriquecimentos complexos em larga escala em dados, leva à necessidade de usar uma gama de tecnologias e linguagens de programação muito mais amplas do que no passado.
- **Dimensão empresarial**: Provavelmente seria um exagero erguer essa dimensão "profissão" como uma nova característica da profissão de cientista de dados. Os estatísticos sempre objetivaram otimizar certas facetas dos negócios de seus negócios. No entanto, essa dimensão está crescendo apenas em um contexto de Big Data e esperamos muito de um cientista de dados nesse ponto. Compreender seu trabalho, analisar os desafios comerciais de seu setor e de sua empresa fazem parte de suas responsabilidades. Confrontado com numerosas figuras e estatísticas, ele também deve ser capaz de entender as escalas usadas, compreender as sutilezas para extrair as informações relevantes para seus negócios.

Figura 4 – Perfil de competências de um cientista de dados



Fonte: Adaptado de Lemberger et al., 2015, p. 73

O trabalho de um cientista de dados consiste em projetar e executar, do início ao fim, o protótipo de um novo produto que implementa técnicas de análise preditiva. Desde o *design* de um novo serviço até o desenvolvimento do protótipo, incluindo a coleta e limpeza de dados, ele terá que ser tecnicamente autônomo, mostrar sua imaginação e estar atento às necessidades básicas que ele deverá formalizar para implementá-los.

O *workflow* referente às tarefas de um cientista de dados pode ser descrito em seis fases (Lemberger et al., 2015, p. 81-87):

1. **Imaginar um produto ou serviço**: Durante essa primeira fase, envolve passar de uma descrição informal de uma necessidade ou oportunidade de negócios para uma formulação mais rigorosa capaz de ser implementada em um modelo preditivo. O cientista de dados, nesse caso, "pensará no produto", primeiro fazendo a pergunta: o que queremos impactar em termos de emprego? A resposta a esta pergunta dependerá, em particular, da definição de uma (ou mais) variável(s) alvo que se deseja prever por meio de diferentes variáveis preditivas com as quais está correlacionada. A definição de uma variável de destino às vezes é menos óbvia do que parece.
2. **Coleta dos dados**: a coleta de dados é uma fase que varia consideravelmente de um projeto para outro. Portanto, é arriscado propor uma abordagem padrão que se aplique a todas as situações. No máximo, podemos listar as questões a serem tratadas aqui.

a. **Disponibilidade dos dados**: depois de ter imaginado as variáveis preditivas e a variável de destino na qual o modelo preditivo será baseado, ainda é necessário garantir a disponibilidade efetiva desses dados. Eles são acessíveis em volume suficiente? Se sim, a que custo? Podemos comprá-los de organizações de terceiros, operadores, fornecedores de arquivos de *marketing*, sensos ou listas de reclamações? Existem fontes de "dados abertos"?

b. **Qualidade dos dados**: intimamente ligada à primeira pergunta, a qualidade dos dados tem várias facetas. A precisão dos dados é suficiente? As fontes de erros foram identificadas para que possam ser corrigidas, se necessário? Os dados são afetados por um viés que poderia influenciar as previsões?

c. **Questões técnicas**: a recuperação de dados pode levantar questões relacionadas aos formatos de dados e às tecnologias que permitem extraí-los das habilidades das pessoas responsáveis por sua coleta.
d. **Questões legais**: os dados que planejamos usar estão livres de direitos? Existe o risco de violar a legislação, em particular no que diz respeito à privacidade, pela liberação que possibilita a verificação cruzada entre vários conjuntos de dados? Existem problemas relacionados à localização geográfica em que os dados desejados serão armazenados?
e. **Questões políticas**: a posse de dados está frequentemente associada a problemas de poder nas empresas. O uso descoordenado, portanto, corre o risco de gerar conflitos. Portanto, um departamento de *marketing* não deseja divulgar seus dados, considerando-os como um "tesouro de guerra". As equipes de TI, preocupadas em não desestabilizar os aplicativos pelos quais são responsáveis, ficarão relutantes em abrir o acesso sem saber com antecedência como e depois de quem serão usados.

3. **Preparação**: depois que os dados forem recuperados das várias fontes, eles ainda precisarão ser utilizados pelos algoritmos de aprendizado. Deve-se mencionar aqui que se trata de homogeneizar os formatos das diferentes fontes, de limpar os dados para excluir os registros que contêm dados ausentes ou então preenchê-los por meio de cálculos apropriados ou fontes de terceiros. Operações de dimensionamento destinadas a harmonizar as unidades usadas para os diferentes parâmetros podem fazer parte deste trabalho de preparação. As operações de cruzamento dos dados iniciais com outras fontes intervêm nesse estágio. Os dados comportamentais extraídos do log de *sites* podem, portanto, ser cruzados com os dados de compra extraídos de uma base transacional.
4. **Modelagem**: A modelagem geralmente ocorre iterativamente com várias tentativas e erros. No entanto, duas tarefas principais podem ser distinguidas. Em primeiro lugar, trata-se de escolher um número restrito de variáveis que serão usadas para alimentar os algoritmos de aprendizado, denominadas *variáveis preditivas*. A atividade de usar o conhecimento de negócios para imaginar novas variáveis que melhorem a qualidade preditiva de um modelo é chamada de *engenharia de requisitos*. É principalmente nesse nível que entram em cena a criatividade e o conhecimento profissional dos cientistas de dados, que, por esse motivo,

consideram essa atividade a "parte nobre e recompensadora" de sua profissão, em contraste com as preocupações muito práticas da preparação de dados. Quando o número dessas variáveis é muito alto (mais de dez), pode ser necessária uma fase de redução de dimensionalidade. Diferentes possibilidades estão disponíveis para o cientista de dados para reduzir o número de parâmetros usados em seus modelos preditivos. É então uma questão de escolher um algoritmo de aprendizado adaptado à situação. A intuição do cientista de dados diante de uma situação específica e a experiência acumulada da comunidade de estatísticos desempenham um papel fundamental aqui na escolha. Mesmo em situações em que é difícil justificar racionalmente a escolha de um algoritmo em detrimento de outro, existem práticas que se provaram úteis em situações semelhantes à que estamos examinando.

5. **Visualização**: No leque de habilidades esperadas de um cientista de dados ou, mais realista, de um laboratório de dados, a comunicação não é a menos importante. A linguagem das estatísticas e dos algoritmos, no centro de sua própria profissão, não é a dos especialistas em negócios e muito menos a dos clientes para os quais os serviços que ele projeta são destinados. Portanto, ele terá que fazer um trabalho educacional e ser capaz de adaptar sua comunicação de acordo com os interlocutores a quem se dirige. A visualização dos dados será seu melhor trunfo para tornar palpáveis as intuições e conclusões que ele extrai de suas análises estatísticas: "uma imagem vale mais que mil palavras", diz o ditado.
6. **Otimização**: A otimização de um sistema preditivo ocorre naturalmente de maneira iterativa. As abordagens ágeis são todas indicadas aqui. O objetivo que deve orientar o cientista de dados ou o laboratório de dados é a rápida realização, de ponta a ponta, de uma solução simples e funcional do serviço previsto. A complexificação ocorre gradualmente após a validação por todas as partes interessadas no projeto. Como em qualquer tarefa criativa, o cientista de dados nunca pode considerar uma análise como final, mas deve, pelo contrário, desafiá-la permanentemente a reexaminar o mesmo fenômeno sob diferentes ângulos. Parte da atividade de otimização abrange a implantação de um aplicativo preditivo na produção. Os primeiros testes em uma população pequena às vezes revelam vieses inesperados que terão que ser remediados. Uma das variáveis que

descobrimos mais tarde não está disponível quando queremos fazer as previsões.

7. **Implementação**: Os dois principais problemas a serem resolvidos durante uma implantação de produção de um aplicativo preditivo são a expansão e a industrialização. Em contextos de Big Data, a expansão às vezes envolve o processamento de volumes de dados de várias ordens de magnitude maiores do que aquelas usadas durante o design e os primeiros testes. Às vezes, parte do código precisa ser reescrita por razões de confiabilidade ou desempenho. Isso diz respeito tanto ao processamento usado durante a preparação dos dados quanto aos próprios mecanismos de aprendizagem e previsão. Esse trabalho de reescrita geralmente será realizado pelas equipes de TI, e não pelos cientistas de dados.

TEMA 3 – CLASSIFICAÇÃO

A *classificação* é uma das tarefas de aprendizado de máquina aplicada quando a classe é um dado nominal, e que acontece de forma supervisionada, ou seja, já existe um conjunto de dados que estão previamente classificados (Amaral, 2016, p. 88). Dessa forma, existem dados históricos sobre o problema o qual se deseja aplicar um processo de classificação. Por exemplo, em análise de crédito, pode-se dispor de um cadastro que informe os clientes que foram bons ou maus pagadores. Com base nestas informações, o cientista de dados busca construir um modelo que possa auxiliar na previsão do comportamento de novos clientes.

No aprendizado da ciência de dados, é comum acessar bancos de dados disponíveis na *Web*, *open source*, de diferentes contextos de aplicação, os quais já foram utilizados para a validação de algoritmos de classificação. Um exemplo de bancos de dados é o UCI Machine Learning Repository (Disponível em: <https://archive.ics.uci.edu/ml/datasets.php>. Acesso em: 4 mar. 2020). Até o momento da escrita deste material, este repositório de bancos de dados contava com 488 diferentes conjuntos de dados. Ainda que sirvam para as outras tarefas como regressão e agrupamentos, a maior parte é dedicada a processos de classificação. A Tabela 1 apresenta alguns bancos de dados (inclusive o IRIS), no qual pode-se identificar o número de instâncias disponíveis, bem como o número de atributos.

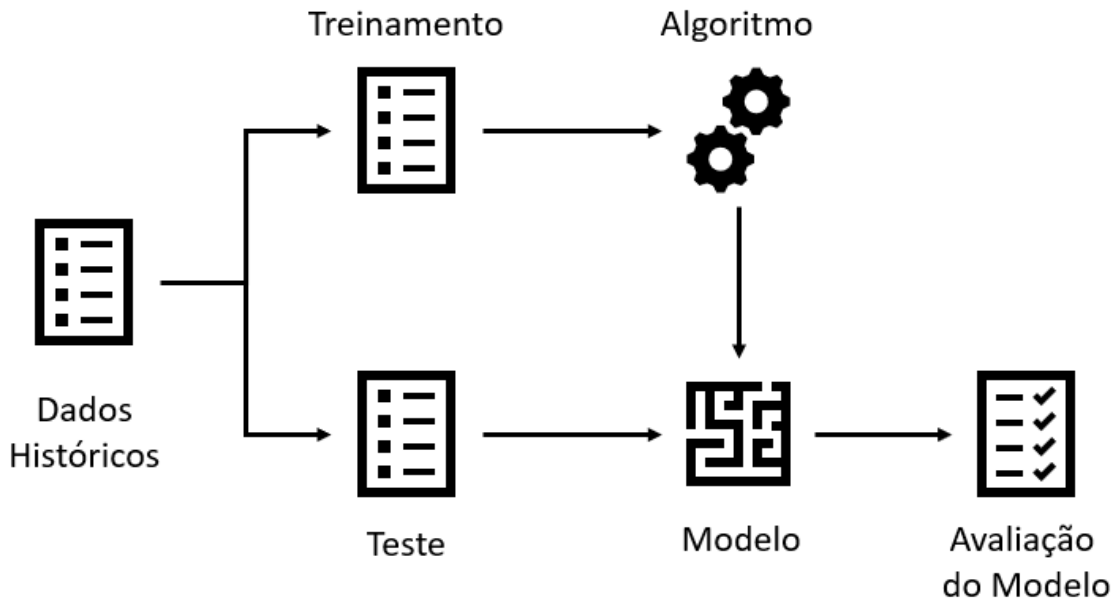
Tabela 1 – Exemplos de Conjuntos de Dados do Repositório UCI ML

Nome	Descrição	Tipos dos Atributos	N. de Instâncias	N. de Atributos
Soybean (Small)	BD de doenças da soja	Categórico	47	35
Autistic Spectrum Disorder	Dados descritivos de crianças no espectro autista	Numérico (Inteiros)	292	21
Comportamento de tráfego em São Paulo	BD com registros sobre o comportamento de tráfego na cidade de São Paulo	Numérico (Inteiros e reais)	135	18
LIBRAS Movement	BD sobre o movimento de mãos na produção de sinais na linguagem LIBRAS	Numérico (Reais)	360	91
IRIS	BD sobre informações da flor iris	Numérico (Reais)	150	4
Breast Cancer Wisconsin (Original)	BD sobre câncer de mama (Dados do Hospital da Universidade do Wisconsin)	Numérico (Inteiro)	699	10

Fonte: Disponível em: <https://archive.ics.uci.edu/ml/datasets.php>. Acesso em: 4 mar. 2020.

O processo de aprendizado assume a construção de um modelo no qual os dados são divididos em dois grupos: o grupo de treinamento e o grupo de teste. O *grupo de treinamento* permitirá a construção do modelo, no qual o grupo de teste será aplicado para verificar o quanto o modelo consegue prever o comportamento dos dados que não fizeram parte do treinamento (Figura 5).

Figura 5 – Diagrama esquemático da construção do modelo de aprendizado de máquina



Fonte: Amaral, 2016, p. 88.

Os grupos de treinamento e teste podem ser criados utilizando-se dois procedimentos: o ***hold out***, que é a técnica mais comum, onde os dados são divididos de forma aleatória, sem substituição (ou seja, quando uma amostra é selecionada para o grupo de treinamento, ela não poderá ser utilizada para o grupo de teste), de maneira que o conjunto de treinamento fique com 70% dos dados e 30% ficando para o grupo de teste. A outra técnica é a **validação cruzada** (*cross validation).* Aqui, as amostras são utilizadas o mesmo número de vezes no grupo de treinamento e no de teste, sendo, em seguida, trocados de forma a repetir um determinado número de vezes (geralmente 10 vezes). O número de interações é conhecido como **partições** ou ***folds*** (Amaral, 2016, p. 91).

Após a submissão dos dados do conjunto de treinamento ao algoritmo, o modelo é então construído, devendo ser testado. Em primeiro lugar, aplica-se o conjunto de treinamento para verificar o quão bom o modelo consegue verificar as classes para o conjunto que serviu de treinamento do modelo. Quando se compara os resultados obtidos com a classificação prévia que os dados de treinamento possuem, cria-se a **matriz de confusão**. Por exemplo, uma matriz de confusão para um modelo para classe com dois valores possíveis, geram-se quatro índices: verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos.

A Tabela 2 mostra um exemplo de matriz de confusão para a classificação dos dados do BD íris. O banco de dados foi dividido conforme o critério ***hold out***, com 70% das amostras para o conjunto de treinamento. O algoritmo utilizado foi o Naive Bayes. Pode-se verificar que a classificação dos dados apresentados para o treinamento mostrou que o modelo classifica corretamente quando a amostra é da classe “setosa”. Entretanto, quando é “versicolor”, o modelo classifica duas amostras erroneamente como “virginica”. E duas amostras que deveriam ser classificadas como “virginica”, são classificadas como “versicolor”.

Tabela 2 – Um exemplo de matriz de confusão para o BD íris

		Classe Prevista		
	IRIS	Setosa	Versicolor	Virginica
Classe Real	Setosa	40	0	0
	Versicolor	0	36	2
	Virginica	0	2	32

Um conjunto de indicadores pode ser utilizado para indicar a eficiência do processo de classificação obtido. Na Tabela 3 estão listados alguns destes indicadores. A Tabela 4 é a adaptação da Tabela 3, incluindo mais uma linha e uma coluna ao final, resumindo as informações de falso negativo para a coluna e de falso positivo para a linha, de maneira a permitir o cálculo dos indicadores. As Tabelas 5, 6 e 7 mostram o cálculo dos indicadores para cada classe.

Tabela 3 – Indicadores utilizados para verificação da eficiência de um modelo

Indicador	Fórmula
Precisão	Precisão = TP / (TP + FP)
***Recall* (Lembrança)**	Recall = TP/(TP+FN)
Acurácia	Acurácia = (TP+TN)/(TP+FP+TN+FN)

Fonte: Adaptado de Amaral, 2016, p. 9

Tabela 4 – Adaptação da tabela 3

IRIS	Setosa	Versicolor	Virginica	FN
Setosa	40	0	0	0
Versicolor	0	36	2	2
Virginica	0	2	32	2
FP	0	2	2	4

Tabela 5 – Indicadores para a classe “setosa”

#setosa	
40 (TP)	0 (FN)
0 (FP)	68 (TN)
Precisão = 40/(40+0)=100% **Recall = 40/(40+0)=100%** **Acurácia = (40+68)/(40+68+0+0)=100%**	

Tabela 6 – Indicadores para a classe “versicolor”

#versicolor	
36 (TP)	2 (FN)
2 (FP)	72 (TN)
Precisão = 36/(36+2)=94.7% **Recall = 36/(36+2)=94.7%** **Acurácia = (36+72)/(36+72+2+2)=96.4%**	

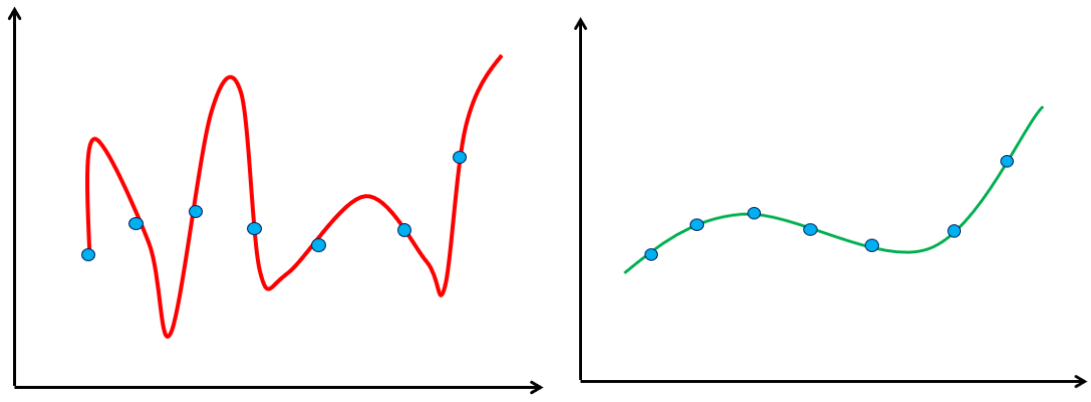
Tabela 7 – Indicadores para a classe “virginica”

#virginica	
32 (TP)	2 (FN)
2 (FP)	76 (TN)
Precisão = 32/(32+2)=94.1% **Recall = 32/(32+2)=94.1%** **Acurácia = (32+76)/(32+76+2+2)=96.4%**	

A tarefa do cientista de dados, na construção do modelo, está orientada para os indicadores de maneira a maximizar a precisão e minimizar a taxa de erros. Dessa forma, está à mão do cientista de dados um arsenal de algoritmos para efetuar esta tarefa da melhor forma possível. No entanto, alguns cuidados devem ser tomados na fase da construção do modelo.

Podem acontecer problemas relacionados à obtenção de uma ótima precisão, *recall* ou acurácia durante o treinamento, porém um péssimo desempenho na fase de testes. Este problema é conhecido como ***overfitting*** (superajuste). No *overfitting*, o modelo fica superadaptado ao conjunto de treinamento; no entanto, as amostras do conjunto de teste geram uma alta taxa de erros. A ideia é de que, durante o treinamento, o modelo atinja um nível adequado de **generalização**, de forma que as amostras de teste possam indicar o mais próximo do valor esperado, conforme o resultado obtido para o modelo (Amaral, 2016, p. 95).

Figura 6 – Modelo com *overfitting* à esquerda e modelo com boa generalização à direita



Outro problema associado à construção de modelos se refere à **maldição da dimensionalidade**. Este problema está relacionado com o alto número de atributos presentes em um modelo, induzindo-o a não funcionar corretamente.

Nem sempre uma grande quantidade de características presentes em um modelo leva a maior eficiência do mesmo. A maldição da dimensionalidade pode ser evitada utilizando-se técnicas de seleção de atributos (Amaral, 2016, p. 95). Pode-se também fazer uma transformação dos dados de forma a gerar um novo conjunto de atributos como, por exemplo, utilizando a técnica de análise de componentes principais.

Uma série de algoritmos pode ser utilizada para a construção do modelo de dados. A Tabela 8 mostra alguns destes algoritmos.

Tabela 8 – Exemplos de Algoritmos para Classificação de Dados

Algoritmo	Descrição	Exemplo
Classificadores Bayesianos	Baseados na teoria de probabilidade condicional de Thomas Bayes. Considera que não existe dependência entre os atributos usados na construção do modelo.	Naive Bayes
Redes Neurais Artificiais	Os atributos são colocados como neurônios na camada de entrada, enquanto as classes são colocadas na camada de saída. As conexões ou pesos, entre neurônios, irão guardar os valores calculados pelo algoritmo de treinamento. O treinamento acontece de maneira que a RNA vai minimizando o erro na classificação das amostras apresentadas à rede.	Multilayer Perceptron
Árvores de Decisão	Uma árvore com vários nós, contendo os valores das instâncias, é construída, de forma particionar continuamente os dados e mostrar qual a classe à qual a amostra pertence.	J48 C-Tree
Máquina de Vetores de Suporte	Modelo que maximiza a margem ou divisão entre as classes, criando um vetor otimizado para fazer a separação entre as instâncias. Podem suportar muitos atributos e serem bastante robustos quanto à maldição da dimensionalidade.	SVM

Fonte: Adaptado de Amaral, 2016, p. 101-104

Os problemas de regressão diferem dos de classificação, no sentido de que a classe é substituída por um atributo numérico. A regressão busca então um relacionamento matemático entre os atributos e o valor que se espera obter. Este relacionamento é denominado de *correlação*. A correlação pode ser medida de forma a indicar a variação em uma faixa de -1 a 1. Quanto mais próximo de -1 ou

1, mais forte será a correlação entre os dados; quanto mais próxima de zero, mais fraca ou indicativa de nenhuma relação entre eles, a correlação será (Amaral, 2016, p. 105).

TEMA 4 – AGRUPAMENTOS

A análise por agrupamentos, também denominada de *análise de conglomerados*, busca reunir objetos com base nas características dos mesmos. Ela classifica objetos de acordo com o que cada elemento do conjunto tem de similar em relação a outros. O grupo resultante desta classificação deve exibir um alto grau de homogeneidade interna (em relação aos outros membros da classe) e alta heterogeneidade externa (em relação aos membros das outras classes) (Polmann, 2017, p. 325).

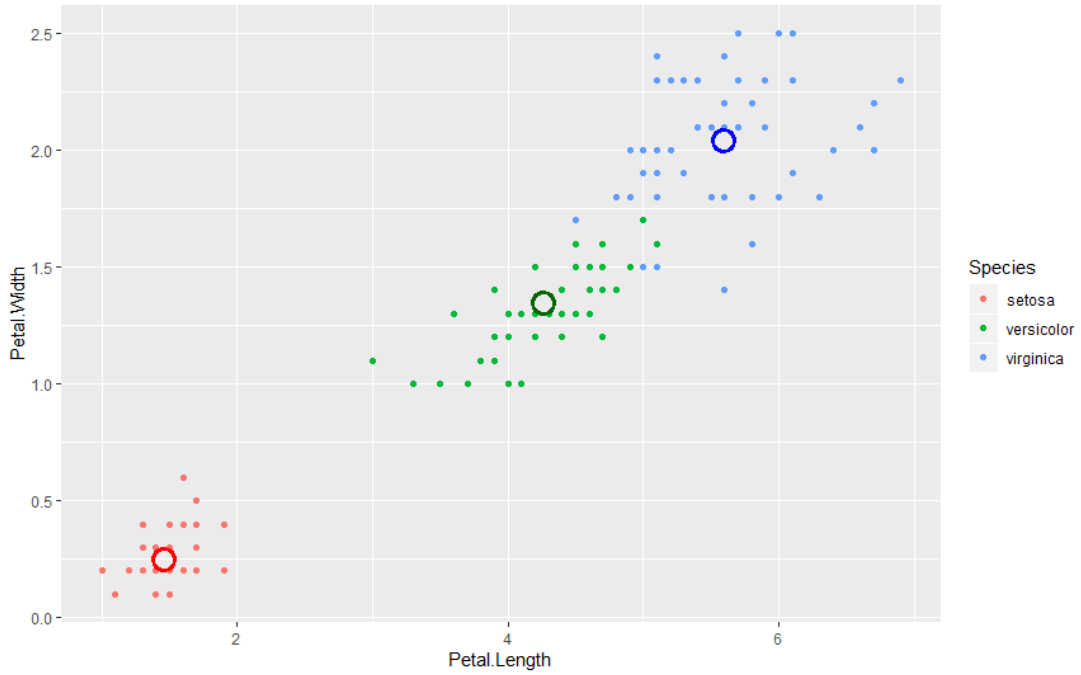
Diferente da classificação vista anteriormente, agrupamentos constituem tarefas de mineração de dados que não são supervisionadas, pois não há uma classe para atribuir a priori. Exemplos de aplicações de tarefas de agrupamento são a identificação de grupos de clientes para direcionamento de campanhas, agrupamento de clientes de seguradoras que são indenizados com mais frequência, identificação de fraudes ou ainda classificação de instâncias quando não se dispõe de classes conhecidas (Amaral, 2016, p. 108).

Os agrupamentos do tipo **particional** dividem as instâncias, cada uma, em grupos únicos. Os agrupamentos particionais podem ser diferenciados pelo uso de modelos baseados em **protótipos** ou baseados em **densidade**. Naqueles baseados em protótipo, as instâncias são agrupadas conforme sua proximidade a um protótipo, podendo ser um **centroide** (centro de uma forma geométrica) ou um ***medoid*** (semelhante ao centroide, porém assumido como uma das instâncias).

O algoritmo **K-Means** é um tipo de agrupamento baseado em protótipos com centroides. Neste tipo de algoritmo, o número de centroides é definido pelo usuário e sempre todas as instâncias serão agrupadas. O algoritmo K-Means é um algoritmo não determinístico, ou seja, seus pontos de início na definição dos agrupamentos são executados de forma aleatória (Amaral, 2016, p. 109).

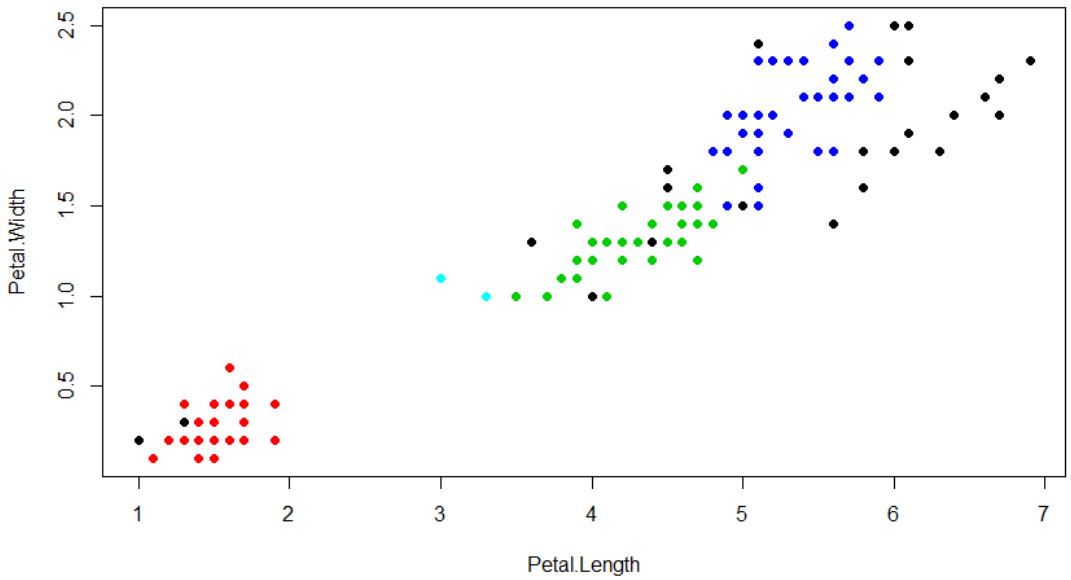
A figura 7 mostra um exemplo de agrupamento com dois atributos do banco de dados de íris, utilizando-se k-means. Note que os centroides são posicionados de acordo com a distribuição de cada um dos conjuntos de amostras (setosa, virginica e versicolor).

Figura 7 – Exemplo de agrupamento com k-means



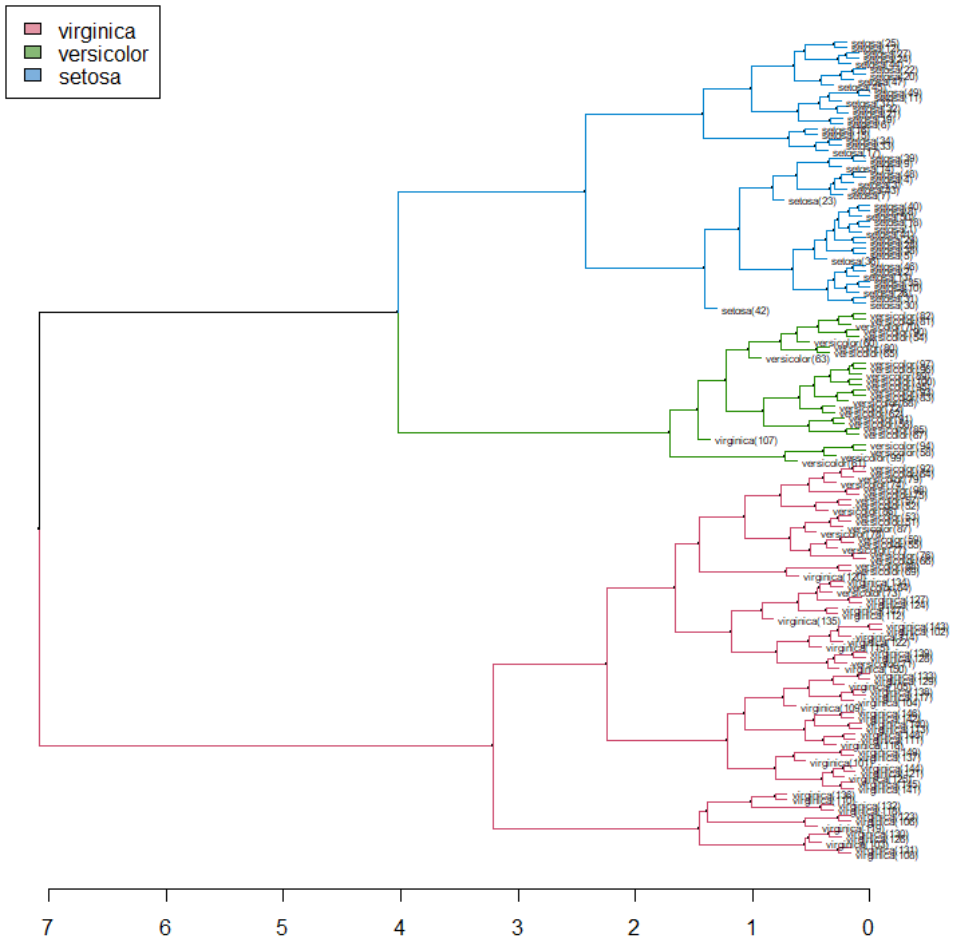
Um algoritmo baseado em densidade de dados é o DBSCAN. Neste algoritmo, o número de grupos é definido automaticamente. Este algoritmo pode ou não classificar elementos em nenhum dos grupos, sendo considerados como ruídos (Amaral, 2016, p. 110). Na Figura 8, está representado o mesmo gráfico da Figura 7, porém após a execução do algoritmo DBSCAN.

Figura 8 – Exemplo com o algoritmo DBSCAN



Além do agrupamento particional, existe o **agrupamento hierárquico**. Este tipo de agrupamento permite que um elemento tenha grupos pais e subgrupos filhos, formando uma estrutura hierárquica. Agrupamentos hierárquicos são normalmente representados por diagramas conhecidos como **dendogramas**. NA Figura 9, encontra-se representado na forma de um dendograma o agrupamento hierárquico a partir do banco de dados de íris.

Figura 9 – Exemplo de um dendograma



TEMA 5 – REGRAS DE ASSOCIAÇÃO

O terceiro método de análise implícita se refere ao uso de regras de associação. É utilizado quando se deseja encontrar uma associação entre diferentes objetos em um conjunto de dados, encontrar padrões frequentes em um banco de dados de transações, bancos de dados relacionais ou qualquer outro repositório de informações.

Um uso frequente de regras de associação pode ser encontrado com a análise de cesta de compras. Um algoritmo de aprendizado de máquina irá minerar as transações em busca de associações entre os itens comprados. Imaginando um conjunto de transações feitas onde cada linha a seguir se relaciona a uma cesta de compras:

Pão, Leite

Pão, Fraldas, Cerveja, Ovos

Leite, Fraldas, Cerveja, Refrigerante

Pão, Leite, Fraldas, Cerveja

Pão, Leite, Fraldas, Refrigerante

Uma regra de associação é um tipo de regra de produção **A -> B**, onde é representada a relação de quem compra A, compra também B. Assim, em aplicações de varejo, a quantidade de regras geradas pode ser muito grande. Para avaliar a relevância das associações, duas métricas são utilizadas: **suporte** e **confiança**.

Supondo-se a regra **Cerveja -> Fraldas**, há um suporte de 60% (3 das 5 transações). Entretanto, a confiança é de 100%, pois quem compra cerveja, sempre compra fraldas. Já a regra **Fraldas -> Cerveja** tem também um suporte de 60%, mas a confiança é de 75% (3 das 4 transações). É comum um algoritmo de regras de associação receber como parâmetros o grau de suporte e confiança mínimos que são esperados. Dessa forma, aquelas regras que não apresentarem relevância, não serão mostradas. Um dos algoritmos mais populares utilizados para regras de associação é o **Apriori** (Amaral, 2016, p. 114).

REFERÊNCIAS

AMARAL, F. **Introdução à Ciência de Dados**: Mineração de dados e Big Data. Rio de Janeiro: Alta Books, 2016.

DUA, D.; GRAFF, C. **Iris Data Set**. UCI Machine Learning Repository. Irvine, CA: University of California, School of Information and Computer Science, 2019. Disponível em: <https://archive.ics.uci.edu/ml/datasets/iris>. Acesso em: 04 mar. 2020.

LEMBERGER, P.; MOREL, M. B. M.; RAFFAELLI, J. L. **Big Data et Machine Learning: Manuel du data scientist.** Paris-France: Dunod, 2015.

MEDEIROS, L. F. de. **Inteligência Artificial Aplicada**: Uma Abordagem Introdutória. Curitiba: InterSaberes, 2018.

POHLMANN, M. C. Análise de Conglomerados. In: CORRAR, L. J.; PAULO, E.; DIAS FILHO, J. M. **Análise Multivariada para os Cursos de Administração, Ciências Contábeis e Economia.** São Paulo: Atlas, 2017.

RUSSEL, S.; NORVIG, P. **Inteligência Artificial**. trad. da 2. ed. Rio de Janeiro: Campus, 2004.