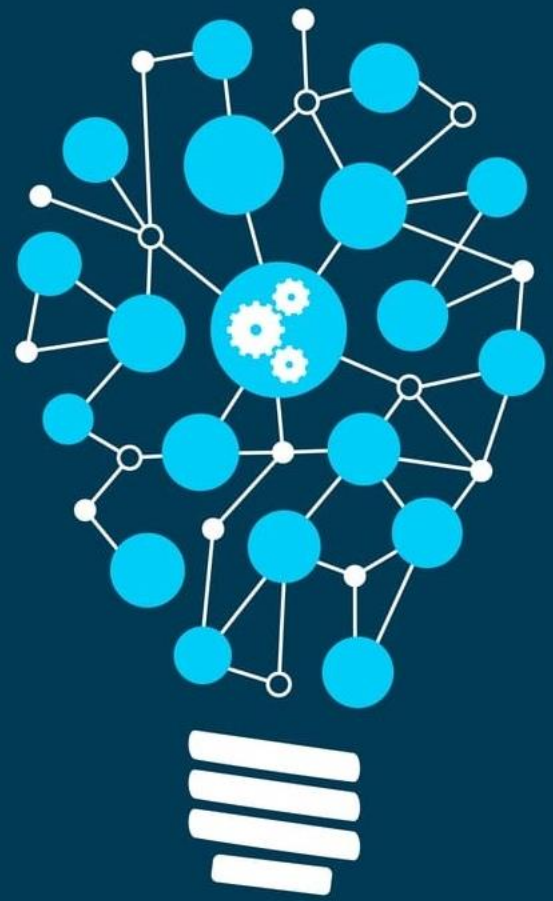
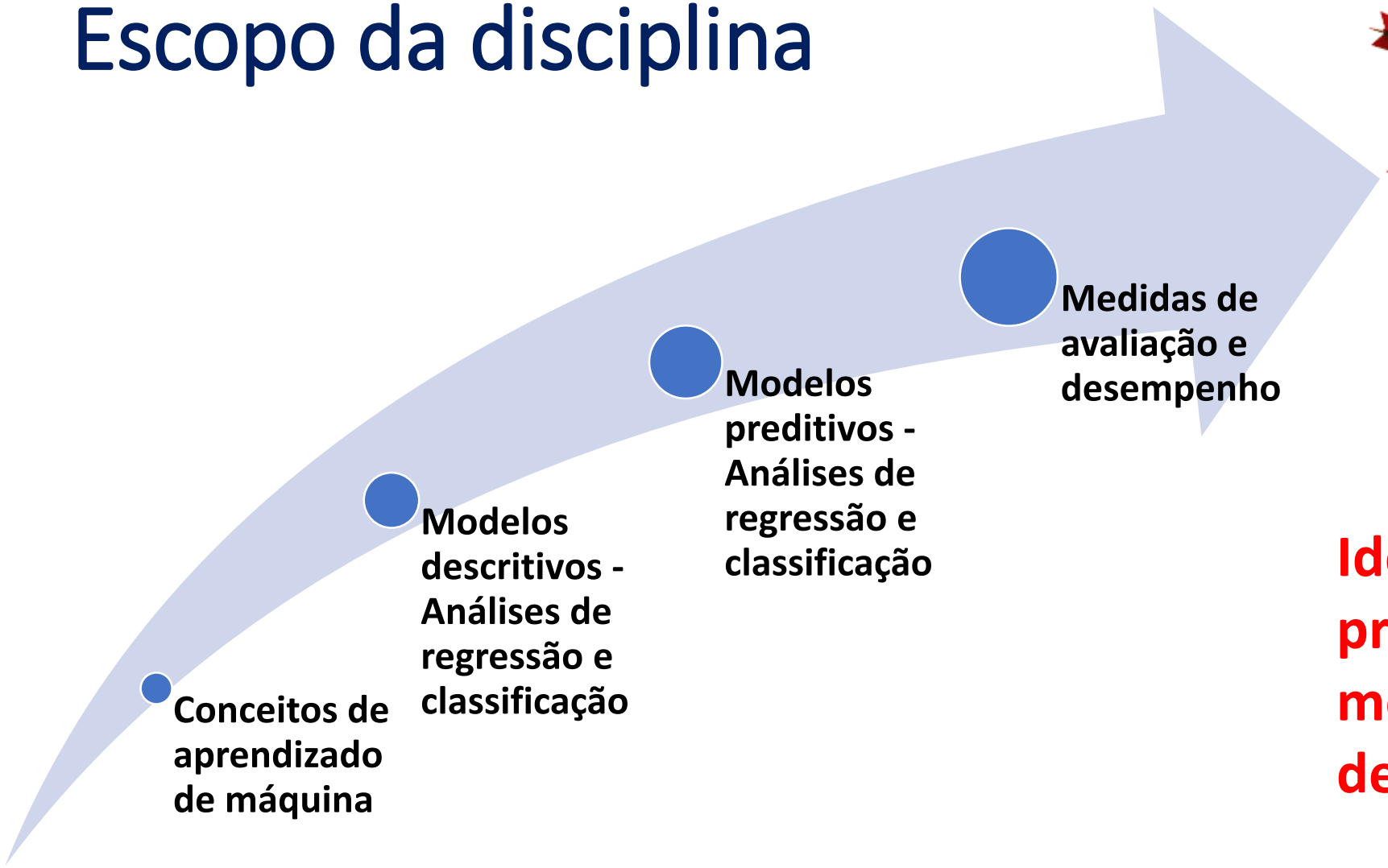
Modelos de aprendizado de máquina aplicados a dados



Escopo da disciplina



Identificar e entender problemas tratáveis por modelos de aprendizado de máquina

Escopo da disciplina

Using machine learning for predicting intensive care unit resource use during the COVID-19 pandemic in Denmark

The COVID-19 pandemic has put massive strains on hospitals, and tools to guide hospital planners in resource allocation during the ebbs and flows of the pandemic are urgently needed. We investigate whether machine learning (ML) can be used for predictions of intensive care requirements a fixed number of days into the future. Retrospective design where health Records from 42,526 SARS-CoV-2 positive patients in Denmark was extracted. Random Forest (RF) models were trained to predict risk of ICU admission and use of mechanical ventilation after *n* days (*n* = 1, 2, …, 15). An extended analysis was provided for *n* = 5 and *n* = 10. Models predicted *n*-day risk of ICU admission with an area under the receiver operator characteristic curve (ROC-AUC) between 0.981 and 0.995, and *n*-day risk of use of ventilation with an ROC-AUC between 0.982 and 0.997. The corresponding *n*-day forecasting models predicted the needed ICU capacity with a coefficient of determination (R^2) between 0.334 and 0.989 and use of ventilation with an R^2 between 0.446 and 0.973. The forecasting models performed worst, when forecasting many days into the future (for large *n*). or *n* = 5, ICU capacity was predicted with ROC-AUC 0.990 and R^2 0.928, and use of ventilator was predicted with ROC-AUC 0.994 and R^2 0.854. Random Forest-based modelling can be used for accurate *n*-day forecasting predictions of ICU resource requirements, when *n* is not too large.

Fonte: https://www.nature.com/articles/s41598-021-98617-1

O que é um modelo?

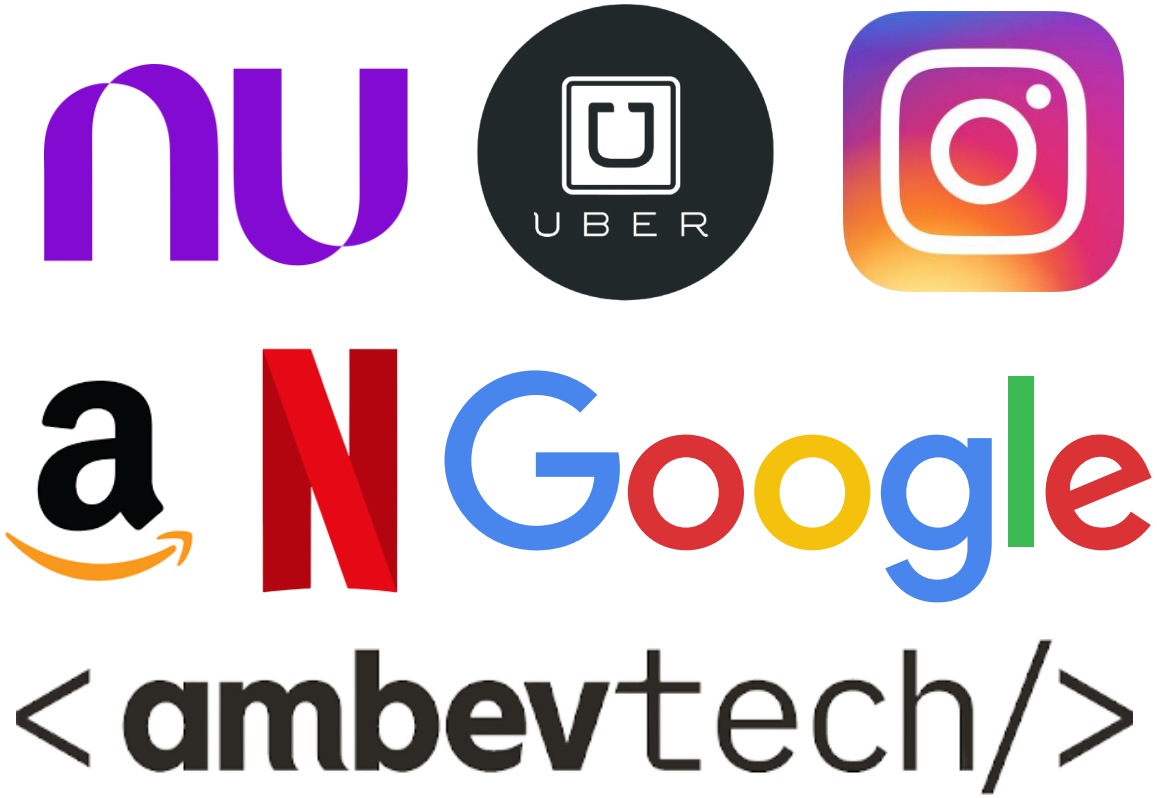
Um modelo é a **representação de um sistema** através de conceitos matemáticos

Geralmente, **especifica a relação entre variáveis**

O que é aprendizado de máquina?

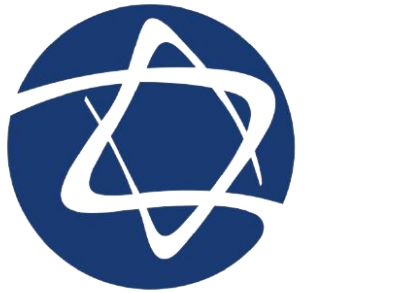
Aprendizado de máquina é uma área da **inteligência artificial** que consiste de ferramentas para **automatizar a construção de modelos** a partir de **dados**

Quais tipos de aplicações podemos criar?



- Definição de limite de crédito
- Precificação de uma viagem de carro
- Previsão de tempo de entrega de produtos
- Recomendação de conteúdo digital
- Ranking de conteúdo

Quais tipos de aplicações podemos criar?



- Previsão de admissão em UTIs para casos confirmados de Covid
- Identificação de pacientes que não comparecerão às consultas
- Detecção de queda com dados de sensores
- Detecção de batimentos irregulares
- Análise de radiografia de tórax

Quais dados precisamos?

Variáveis independentes

Variável dependente

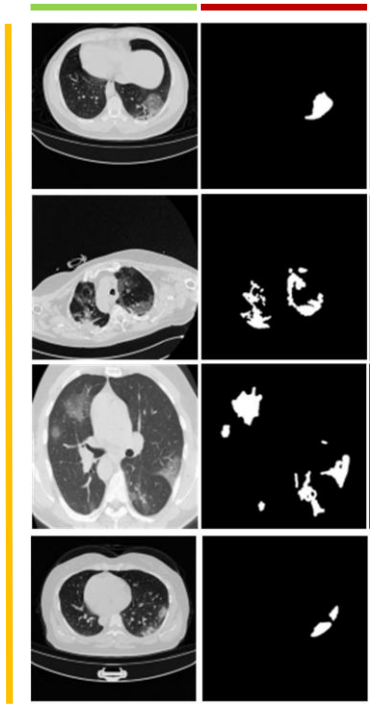
Age	Sex	ChestPain...	RestingBP	Cholesterol	RestingECG	ExerciseA...	HeartDise...
40	M	ATA	140	289	Normal	N	0
49	F	NAP	160	180	Normal	N	1
37	M	ATA	130	283	ST	N	0
48	F	ASY	138	214	Normal	Y	1
54	M	NAP	150	195	Normal	N	0
39	M	NAP	120	339	Normal	N	0
45	F	ATA	130	237	Normal	N	0
54	M	ATA	110	208	Normal	N	0
37	M	ASY	140	207	Normal	Y	1

Observações

Dados

Quais dados precisamos?

Variável independente Variável dependente

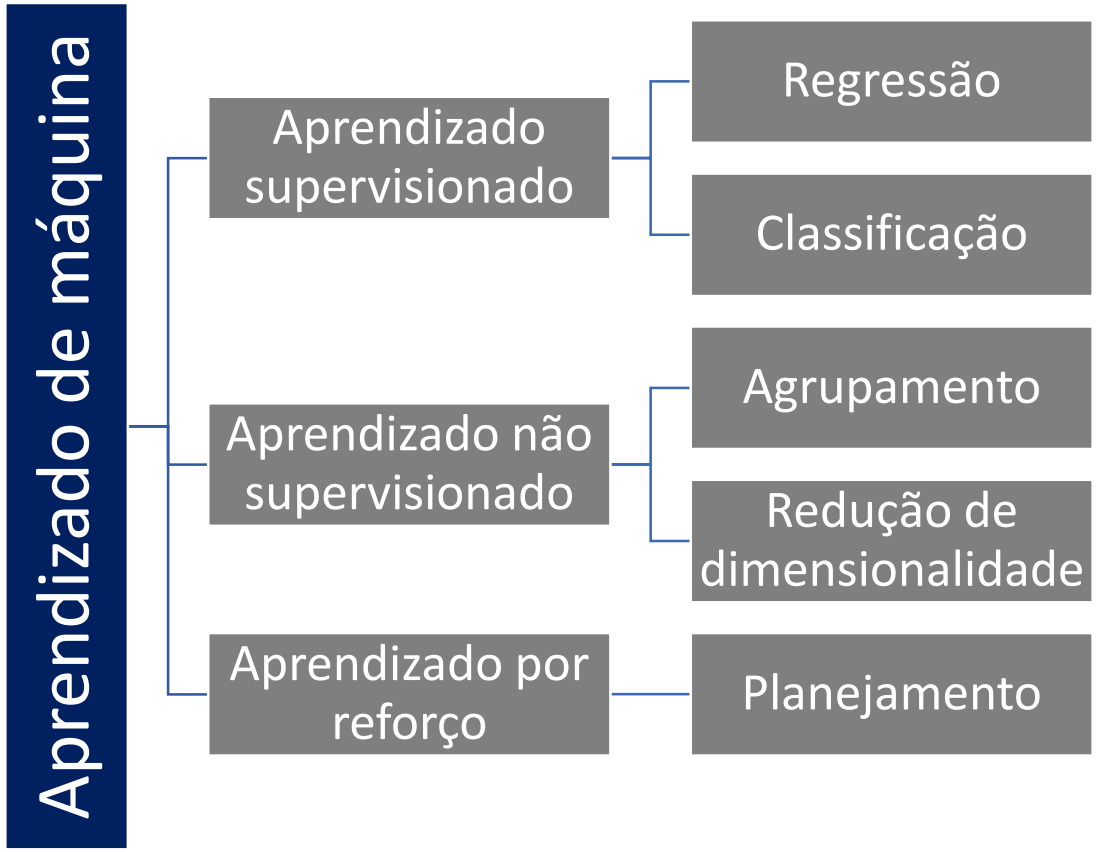


Observações

Podemos trabalhar com diferentes tipos de dados, sejam **estruturados** ou **não estruturados**

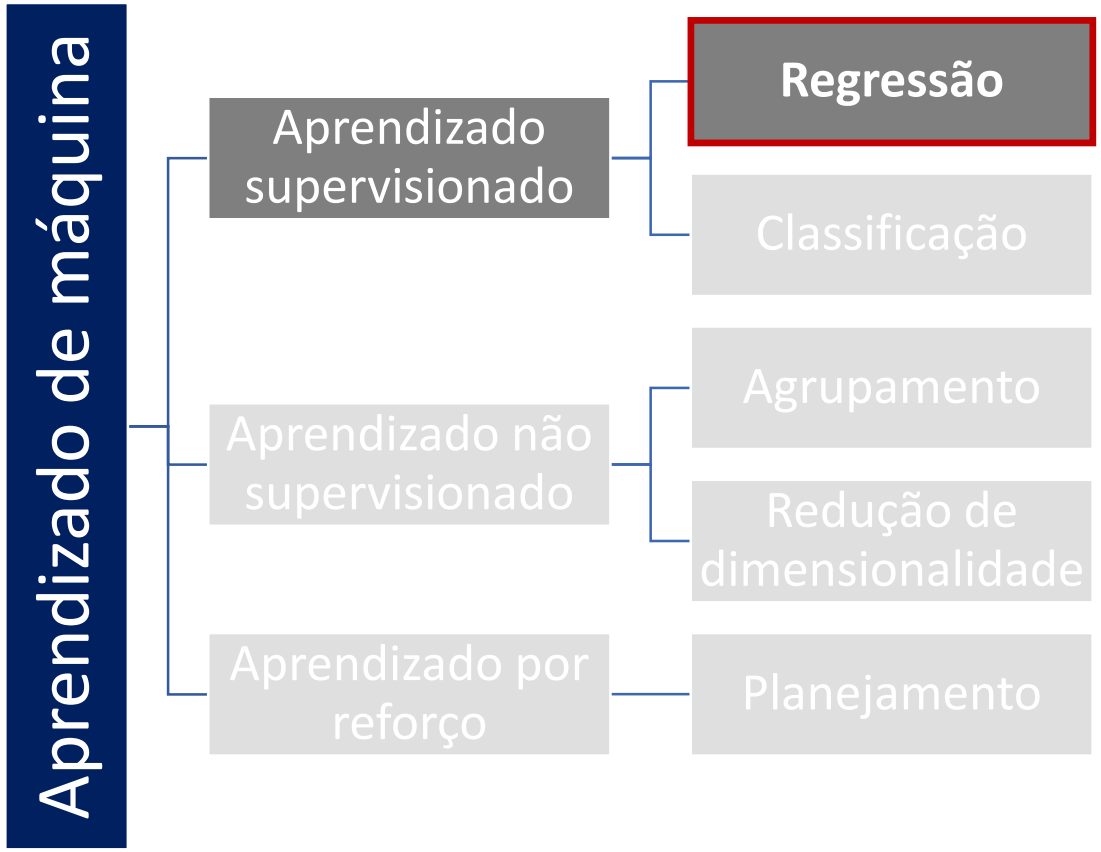
Independente da forma, para descrever ou prever relações entre variáveis **precisamos de conjuntos de dados anotados**

Tipos de aprendizado de máquina



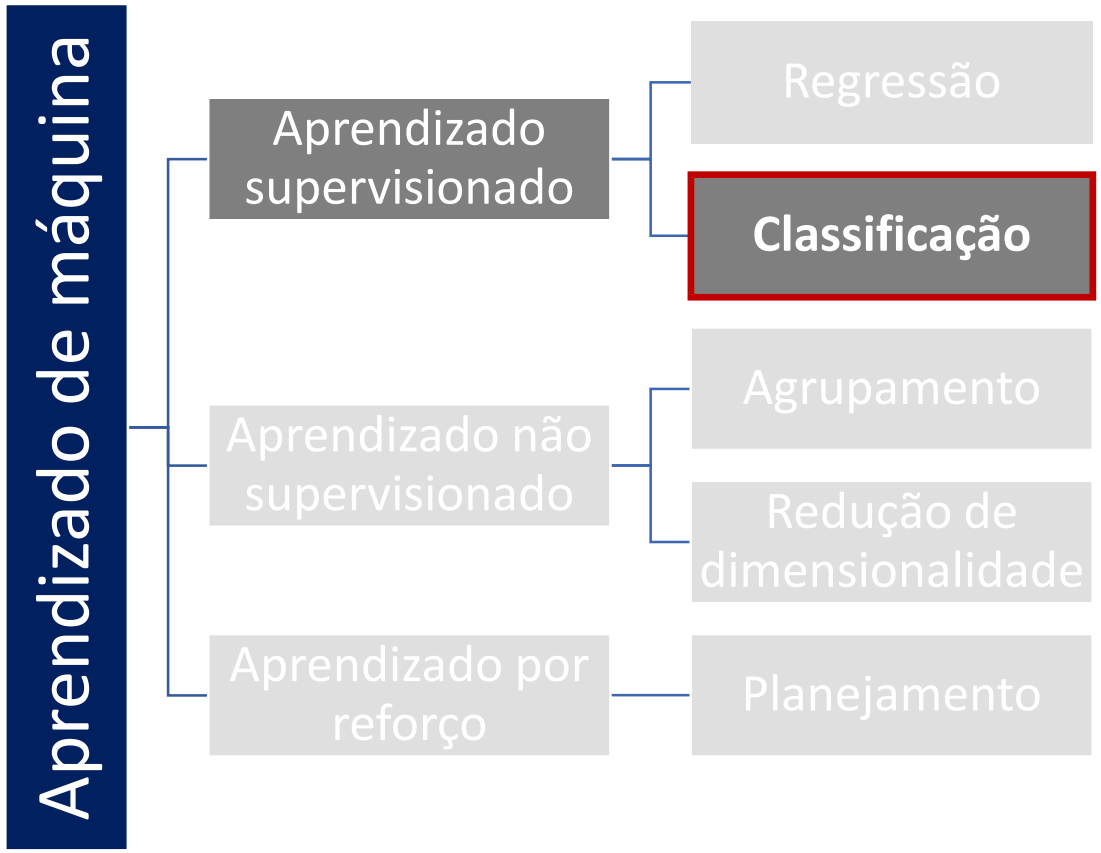
Nesta disciplina, focaremos no **aprendizado supervisionado**, discutindo suas **principais tarefas** e como construir modelos **descritivos** e **preditivos**

Tipos de aprendizado de máquina



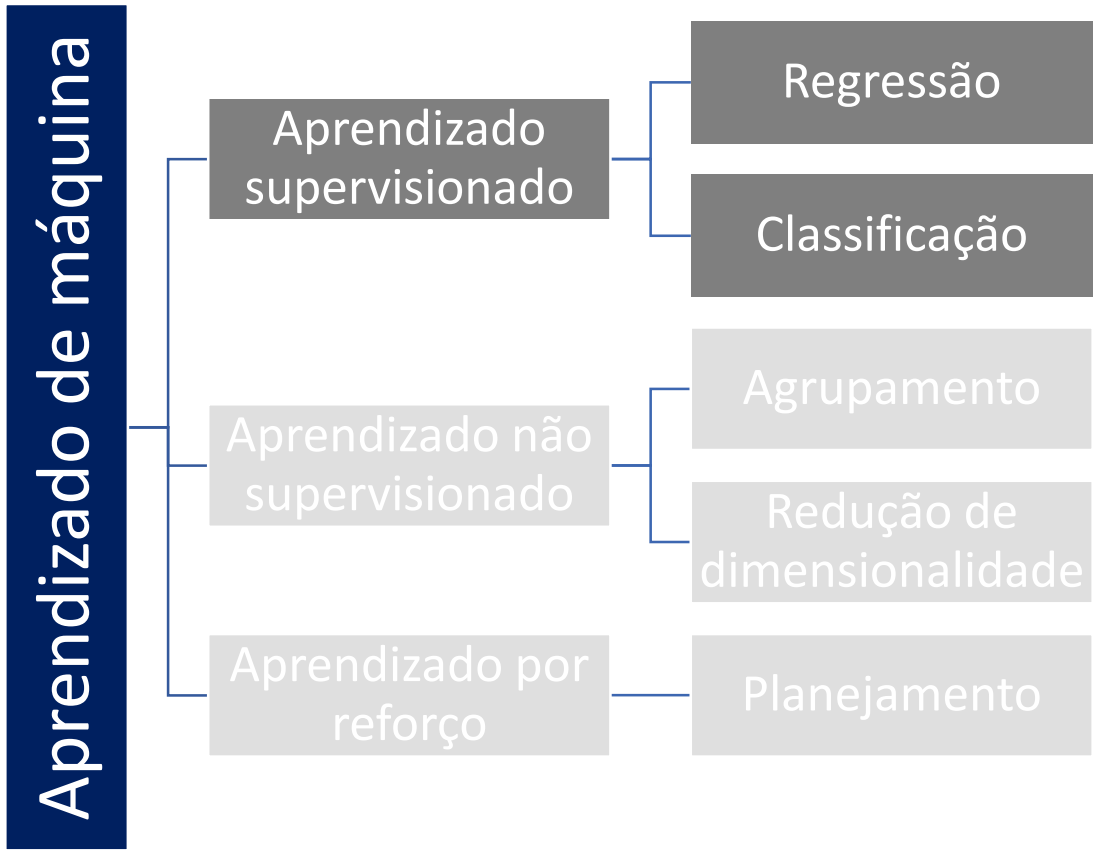
Tarefa que consiste em **aprender um modelo** para prever **variáveis contínuas**

Tipos de aprendizado de máquina

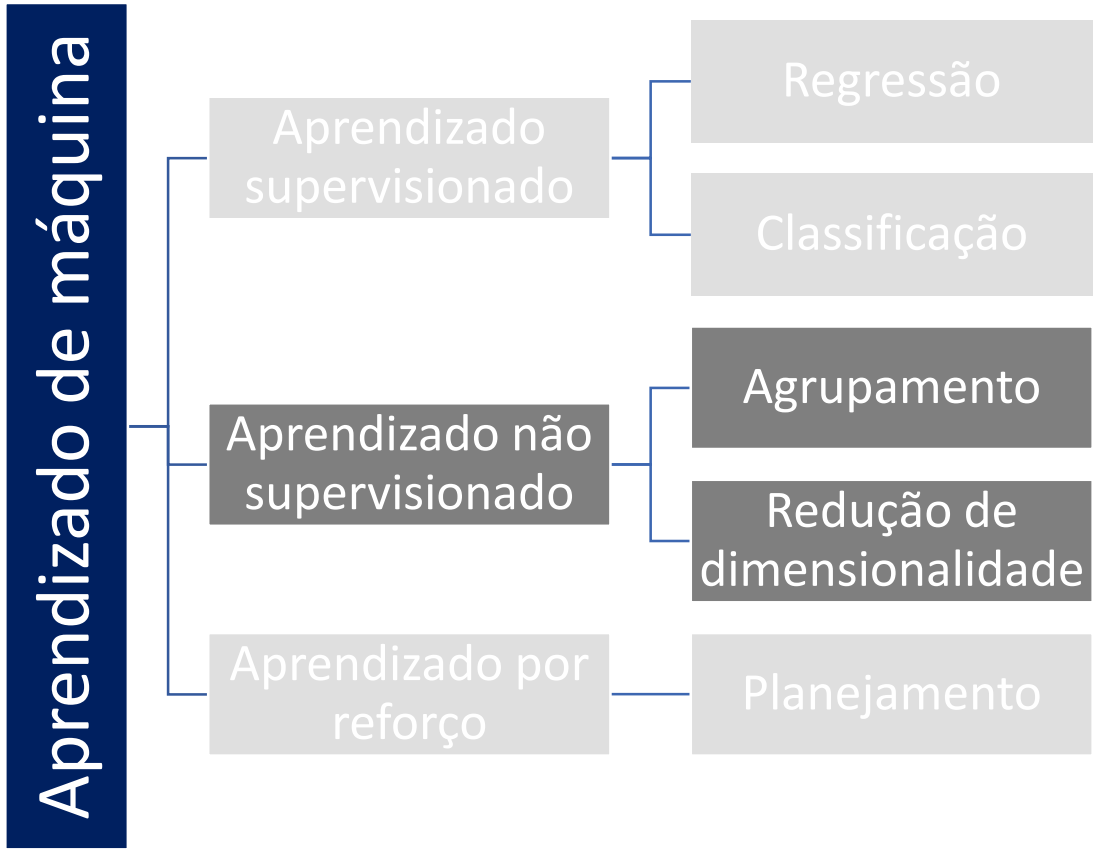


Tarefa que consiste em **aprender um modelo** para prever **variáveis discretas**

Tipos de aprendizado de máquina



Tipos de aprendizado de máquina



Fonte: https://www.youtube.com/watch?v=4b5d3muPQmA

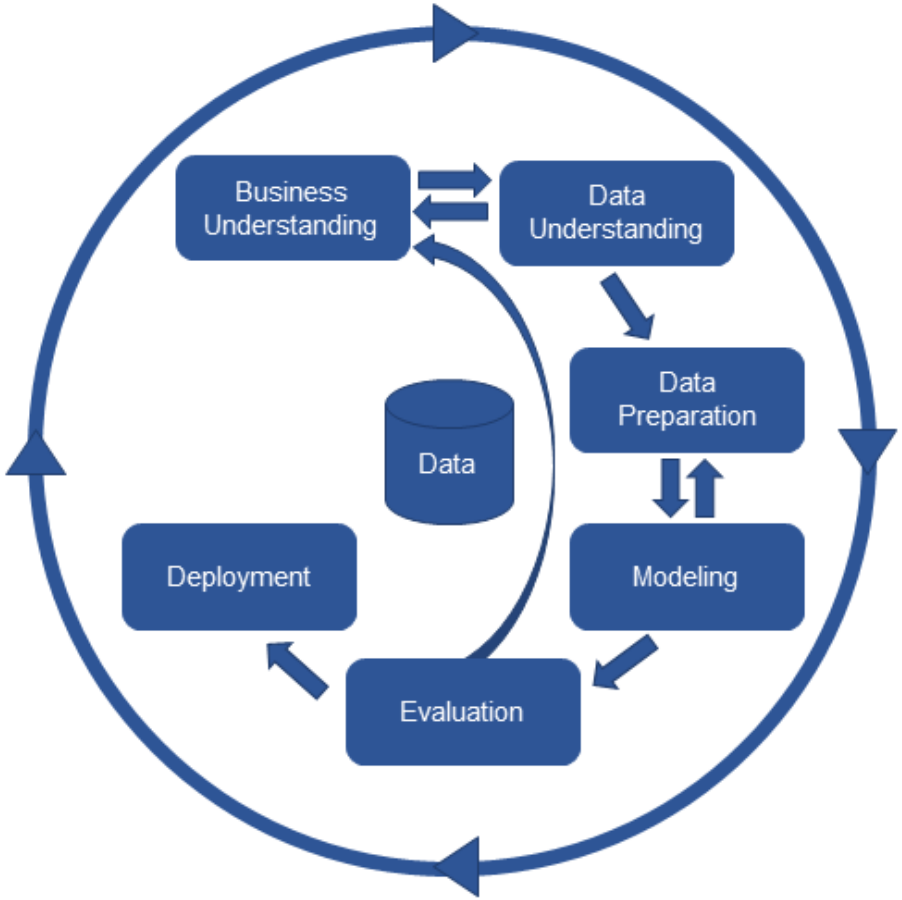
Tipos de aprendizado de máquina

Fonte: https://www.youtube.com/watch?v=WXuK6gekU1Y

Classificação ou Regressão?

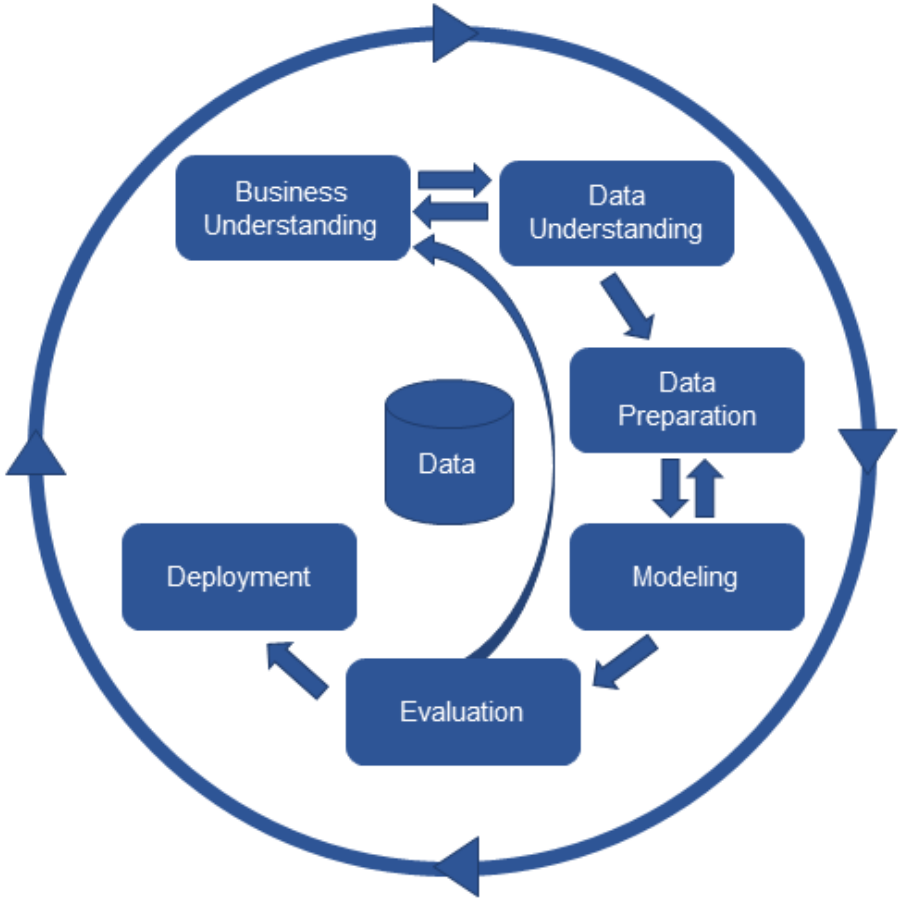
Predição de no show em consultas	Classificação
Precificação de planos de saúde	Regressão
Triagem de pacientes	Classificação
Identificação de melanoma	Classificação e Regressão
Predição de duração de cirurgias	Regressão
Gestão da sala de operações	Otimização

Etapas do desenvolvimento de uma aplicação de aprendizado de máquina



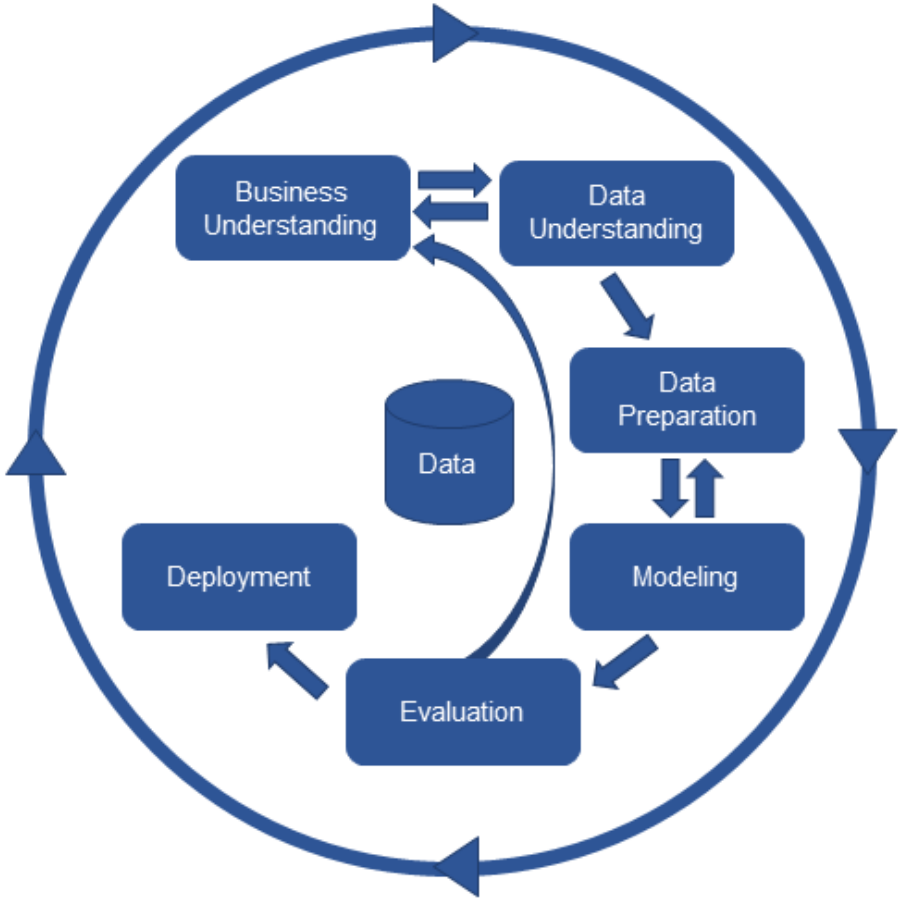
- **Entendimento do negócio**
  - Definir o objetivo do ponto de vista de negócio
  - Definir o objetivo do ponto de vista de dados
  - Definir o critério de sucesso do projeto
  - **Entregável: contrato de planejamento do projeto**

Etapas do desenvolvimento de uma aplicação de aprendizado de máquina



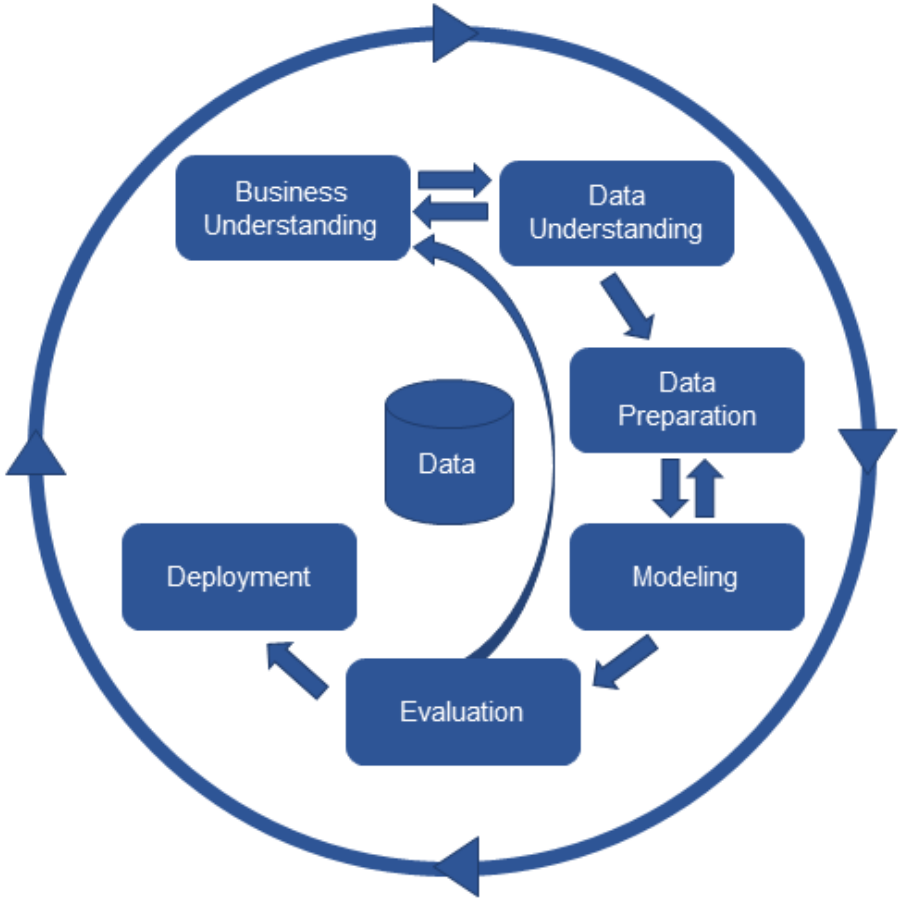
- **Entendimento dos dados**
  - Coleta e armazenamento de dados
  - Análise da qualidade de dados
  - Análise exploratória de dados
  - **Entregável: hipóteses de modelagem / insights das análises**

Etapas do desenvolvimento de uma aplicação de aprendizado de máquina



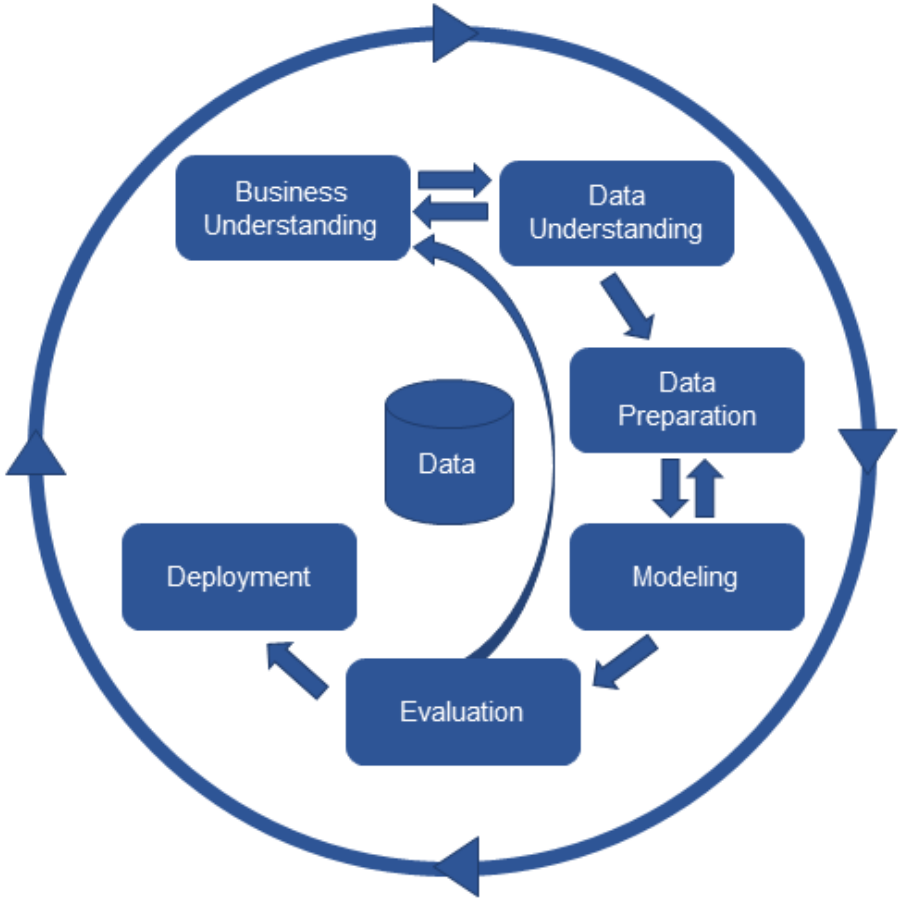
- **Preparação dos dados**
  - Limpeza de dados
  - Integração de dados
  - Normalização de dados
  - Transformação de dados
  - **Entregável: conjunto de dados para o uso das ferramentas de aprendizado de máquina**

Etapas do desenvolvimento de uma aplicação de aprendizado de máquina



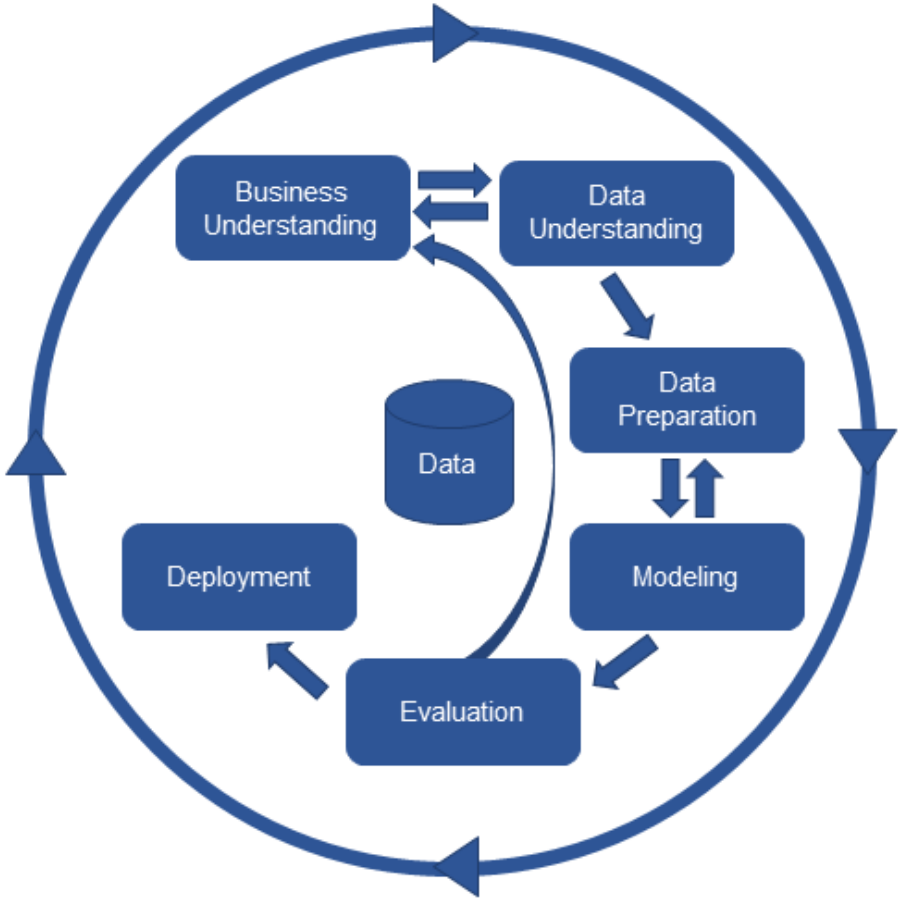
- **Modelagem**
  - Escolha das técnicas de modelagem
  - Desenho e execução do experimento
  - **Entregável: modelo resultante da experimentação**

Etapas do desenvolvimento de uma aplicação de aprendizado de máquina



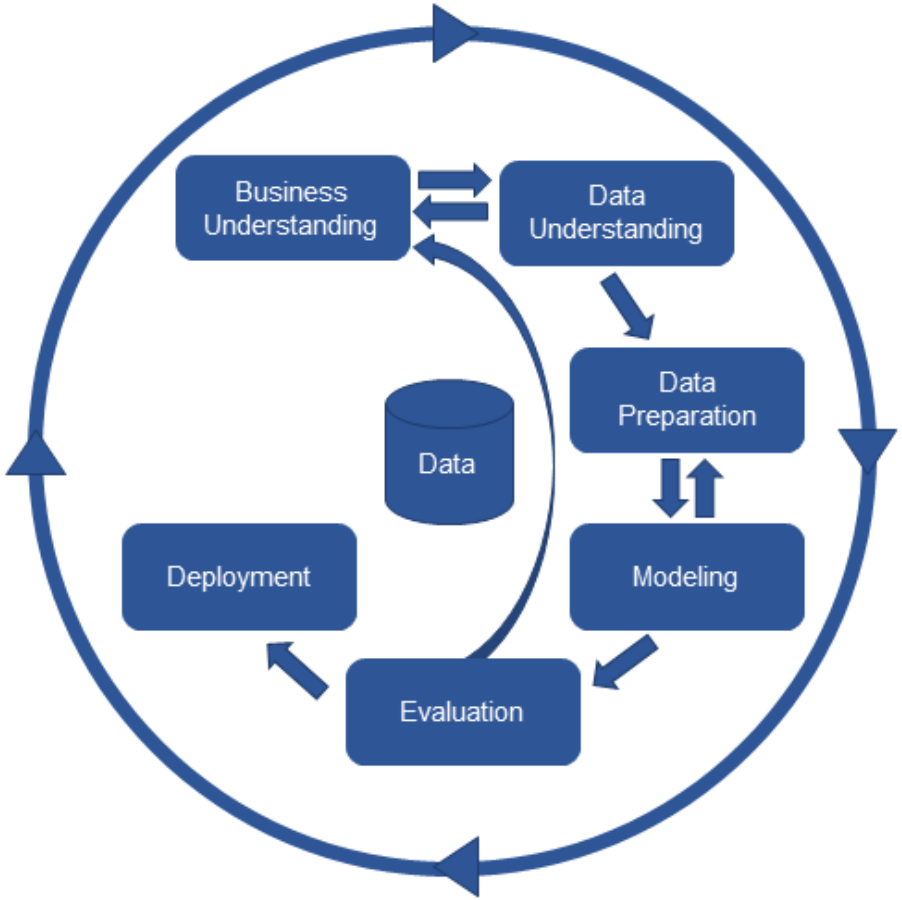
- **Avaliação**
  - Interpretação dos resultados sob o ponto de vista do negócio
  - Revisão do processo
  - **Entregável: análise dos resultados**

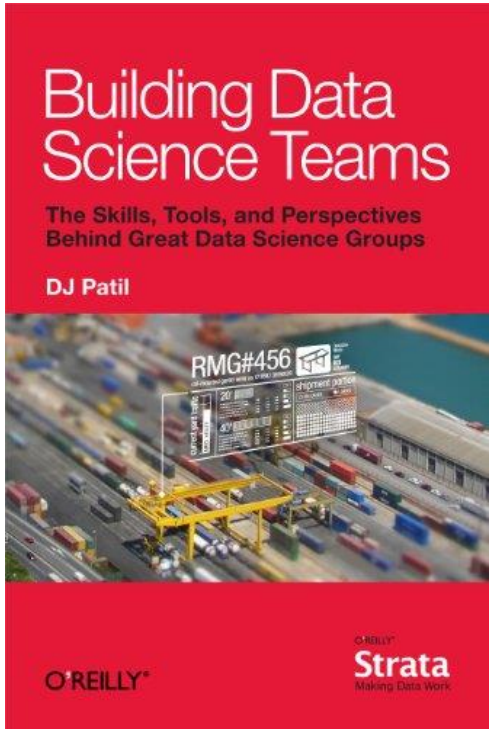
Etapas do desenvolvimento de uma aplicação de aprendizado de máquina



- **Implantação**
  - Definição da estratégia de integração do modelo com o processo da empresa
  - Definição de um plano de monitoramento e manutenção do modelo
  - **Entregável: modelo integrado ao processo da empresa**

Etapas do desenvolvimento de uma aplicação de aprendizado de máquina





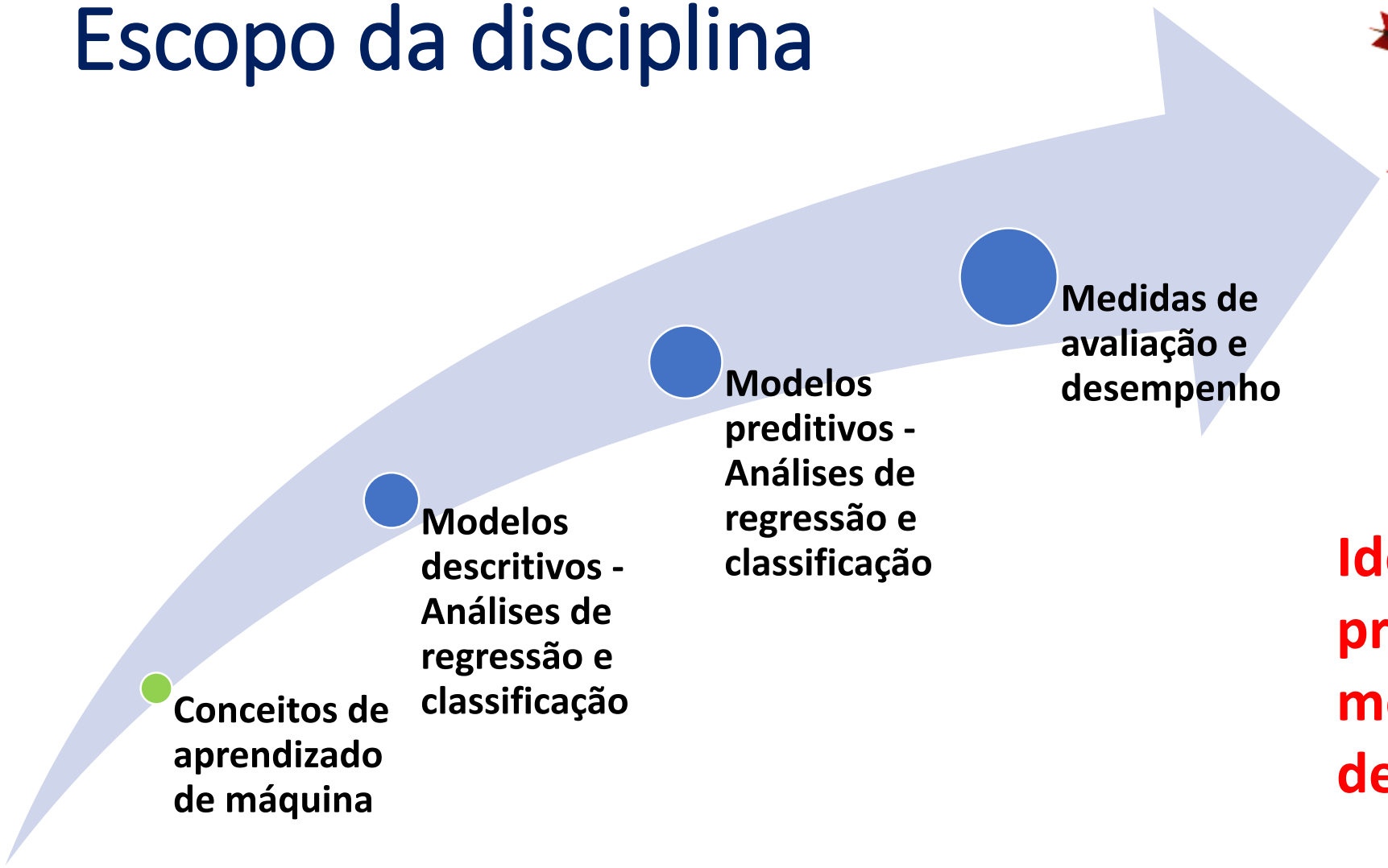
Fonte: http://www.datascienceassn.org/sites/default/files/Building%20Data%20Science%20Teams.pdf

Etapas do desenvolvimento de uma aplicação de aprendizado de máquina

Se você fosse responsável por implantar uma solução de aprendizado de máquina na sua empresa, você optaria por uma **consultoria**, a **contratação de profissionais** ou **ferramentas de mercado**?

Quais perfis de profissionais da área de dados você precisaria?

Escopo da disciplina



Identificar e entender problemas tratáveis por modelos de aprendizado de máquina

O que é um modelo descritivo?

Exemplo: A empresa FRITZ MÜLLER está enfrentando um problema recorrente de pacientes que **marcam consultas e não comparecem ao consultório** na data agendada. O time de executivos da empresa precisa definir uma estratégia para reduzir a quantidade de ***no show*** **de pacientes** e decidiu contratar uma consultoria para entender quais fatores influenciam os pacientes a não comparecerem às consultas.

Modelos descritivos simplificam um sistema para interpretarmos a importância das variáveis independentes na predição da variável dependente

Como construir um modelo descritivo?

Com o conjunto de dados anotado e o problema definido, qual o **primeiro passo para desenvolver nosso modelo**?

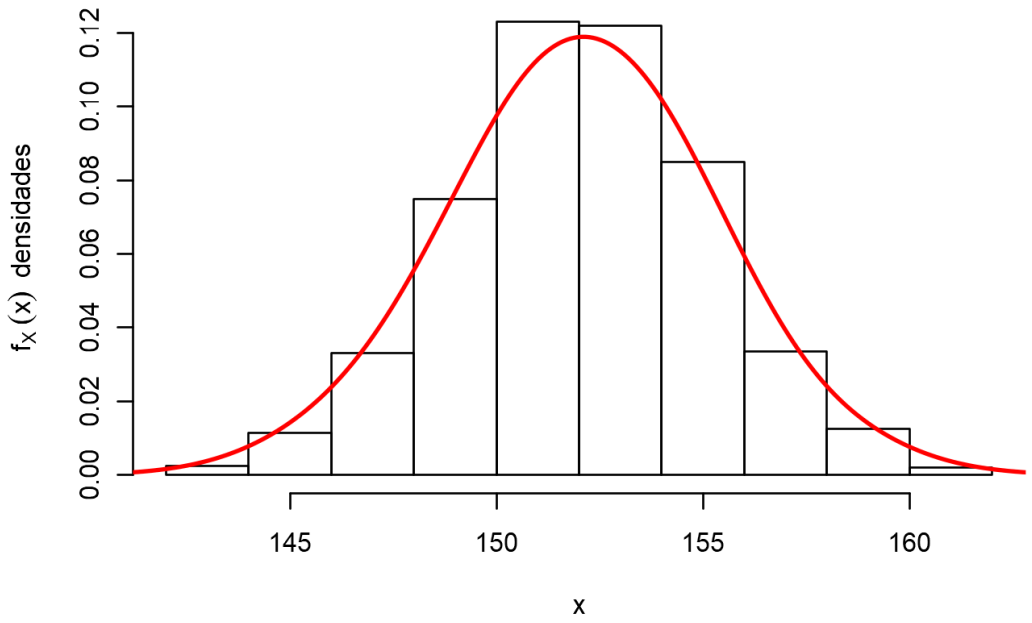
Entender os dados. Para isso, utilizamos ferramentas da **estatística descritiva.** Vamos relembrar essas ferramentas e entender como elas estão relacionadas com os nossos modelos

Como construir um modelo descritivo?

Nosso conjunto de dados é composto de múltiplas variáveis. Conduzimos nossa exploração analisando os dados de forma individual, em pares e em conjunto.

Exploramos a **distribuição das variáveis**, calculamos **medidas de posição** e **dispersão**

Também aplicamos técnicas de **limpeza de dados**, como o tratamento de dados faltantes, padronização de unidades de medidas e nomenclaturas

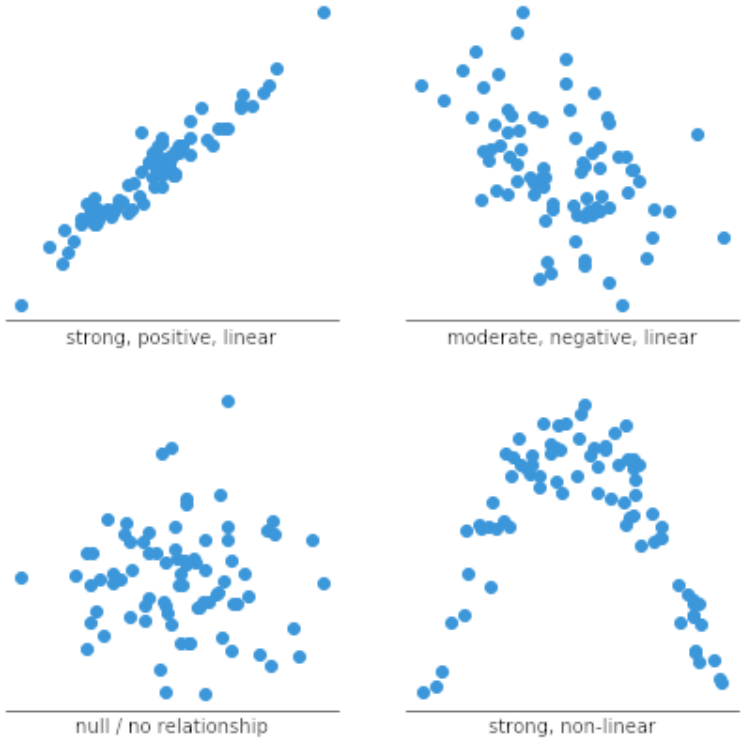


Como construir um modelo descritivo?

Nosso conjunto de dados é composto de múltiplas variáveis. Conduzimos nossa exploração analisando os dados de forma individual, em pares e em conjunto.

Exploramos a **associação entre as variáveis** através de **medidas de correlação** e **gráficos de dispersão**

Essas informações são relevantes para definirmos as **variáveis do nosso modelo** e até **qual modelo utilizar**

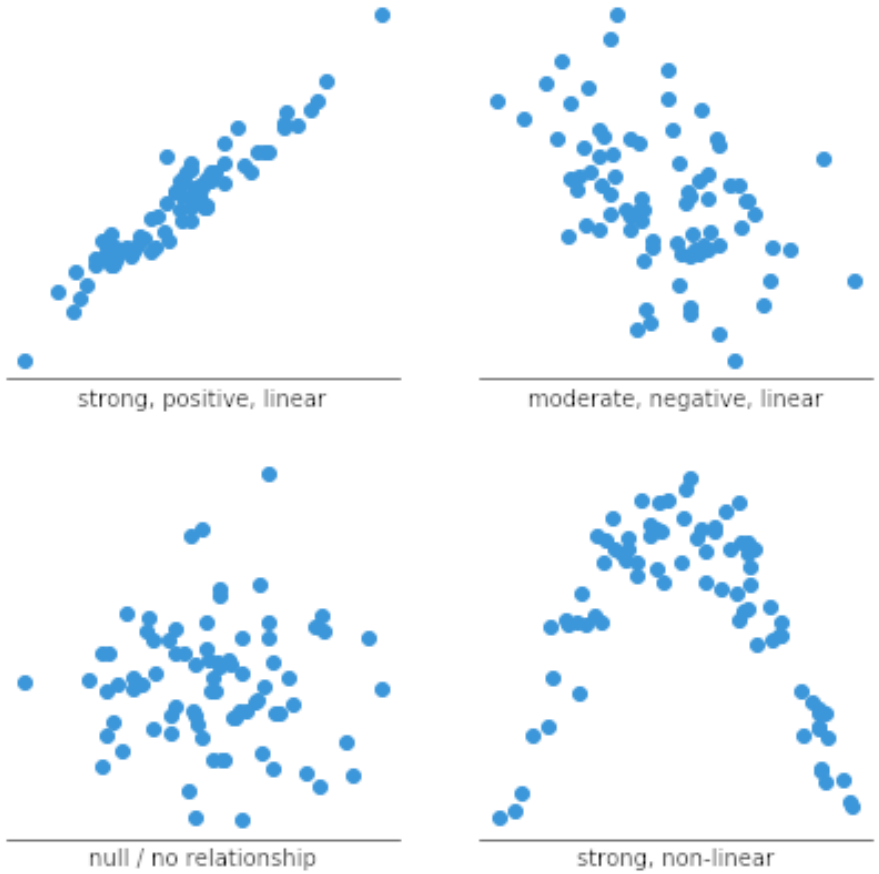


Como construir um modelo descritivo?

Podemos utilizar a correlação entre as variáveis para construir o nosso modelo. Lembrando: modelo é a representação de um sistema com conceitos matemáticos.

Qual conceito podemos utilizar para representar as correlações ao lado?

Podemos utilizar **funções** para representar a **relação entre variáveis**



Como construir um modelo descritivo?

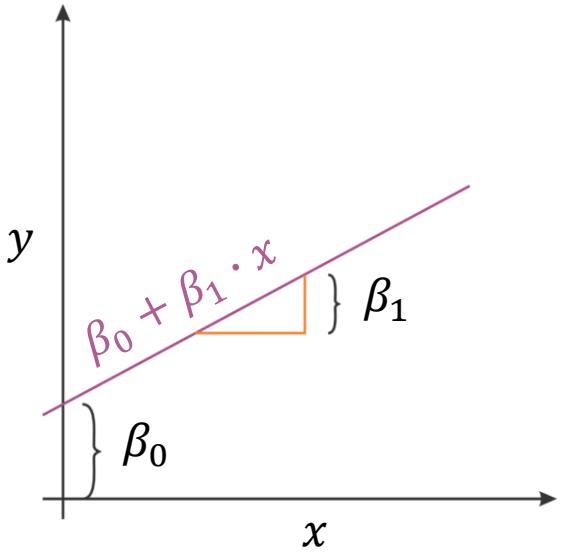
Um dos modelos mais utilizados para representar a relação entre variáveis é a **regressão linear**

- A equação de uma reta é definida como

$$Y_i = \beta_0 + \beta_1 \cdot x_i + \varepsilon_i \text{ para } i = 1,2,3,\ldots,n$$

- Onde:
  - Y_i é a variável dependente
  - x_i é a variável independente
  - β_1 é o coeficiente angular
  - β_0 é o intercepto
  - ε_i é o erro aleatório

} Parâmetros a serem estimados



Como construir um modelo descritivo?

Equação da reta, com 1 variável

Equação da reta, com 2 variáveis

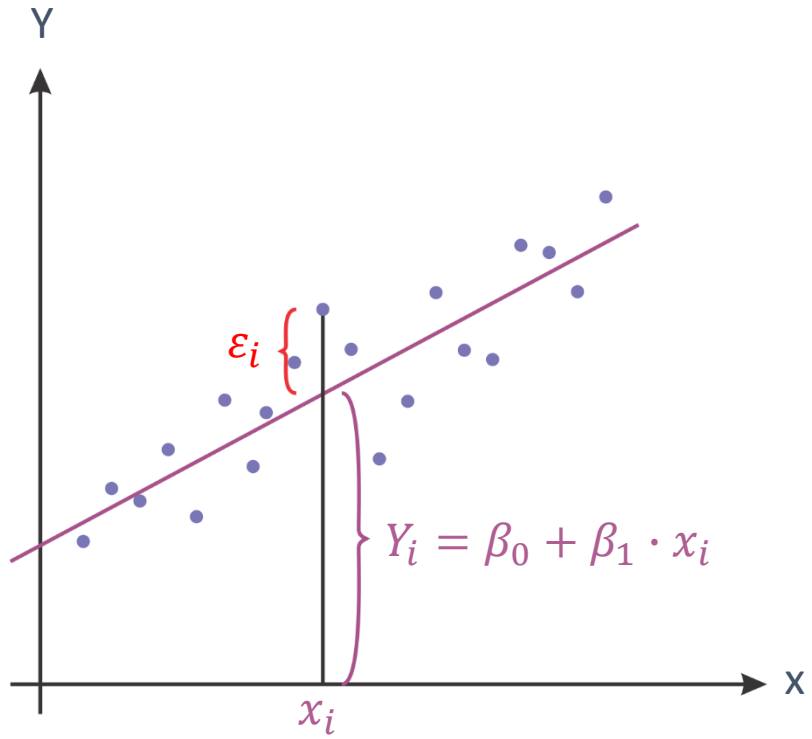
Como construir um modelo descritivo?

- **Como estimar β_0 e β_1**
  - O objetivo é minimizar os desvios (ε_i) entre os valores observados e os estimados.

$$\varepsilon_i = Y_i - \beta_0 - \beta_1 \cdot x_i$$

- Método dos Mínimos Quadrados

$$L = \sum_{i=1}^{n} {\varepsilon_i}^2 = \sum_{i=1}^{n} [Y_i - \beta_0 - \beta_1 \cdot x_i]^2$$



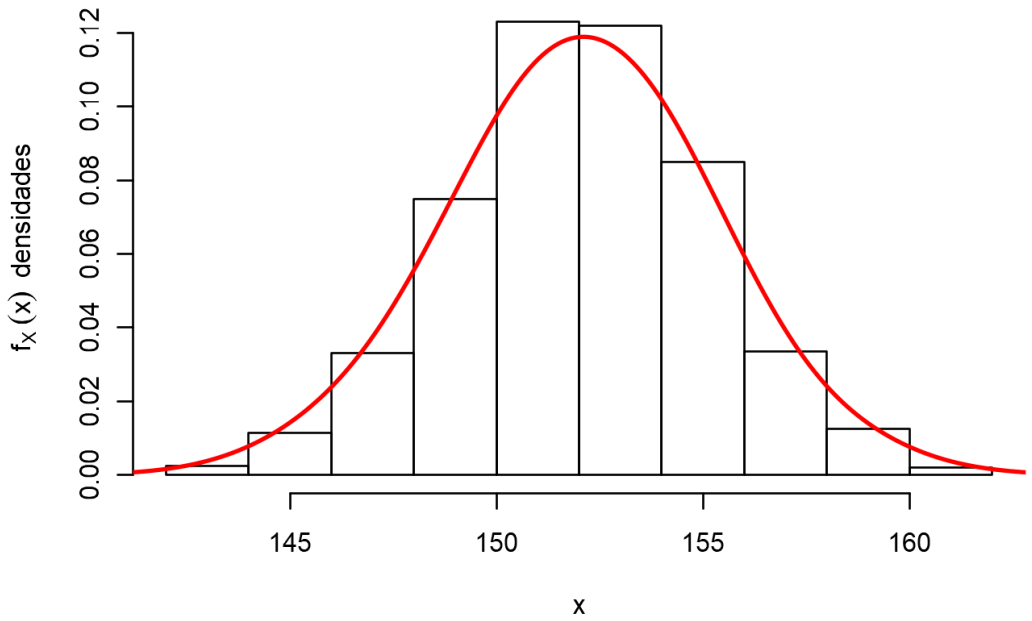
Condições para um bom ajuste do modelo

Para utilizarmos corretamente o modelo de **regressão linear**, precisamos que os dados estejam de acordo com algumas **condições assumidas pelo modelo**

Normalidade dos resíduos

Os resíduos gerados pelo ajuste da reta devem seguir uma **distribuição normal**

Resíduos: valores que representam o erro entre o valor estimado pelo modelo e o valor real



Condições para um bom ajuste do modelo

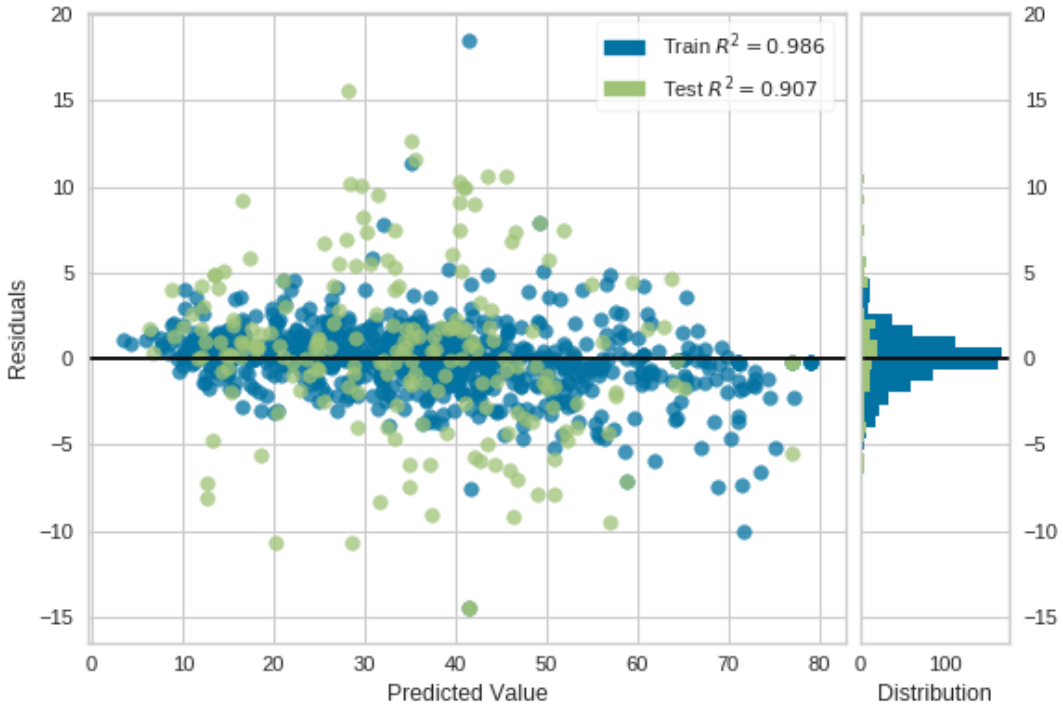
Para utilizarmos corretamente o modelo de **regressão linear**, precisamos que os dados estejam de acordo com algumas **condições assumidas pelo modelo**

Homocedasticidade

É preciso que a variância da variável Y seja constante para todos os valores de X

Independência

É necessário que as variáveis sejam independentes, para que os resíduos sejam independentes e identicamente distribuídos



Como interpretar o modelo?

Os parâmetros no modelo de regressão linear são **estimativas pontuais**, para interpretá-los, precisamos antes **testar a sua significância**, ou seja, com qual nível de confiança conseguimos afirmar que a **estimativa é diferente de zero**.

Exemplo da equação da reta ajustada

$$E(Y) = 150 + 1{,}5 \cdot idade + 75 \cdot fumante$$

Associado a cada parâmetro teremos um **p-valor**, resultante do teste de hipótese

Assumindo $\alpha = 0.05$, **rejeitamos a hipótese nula** de que a estimativa é zero

Podemos interpretar o modelo da seguinte maneira

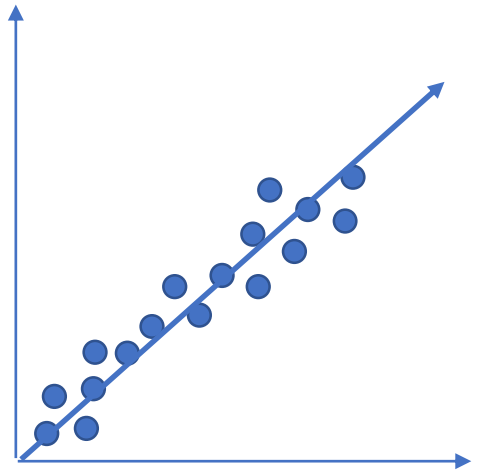
O plano tem um custo médio de 150 reais, independente da idade e risco dos clientes

Para cada ano, o custo do plano sobe 1,5 reais. Pessoas com 10 anos de idade pagariam uma mensalidade de R$165,00, enquanto pessoas de 20 anos pagariam uma mensalidade de R$180,00

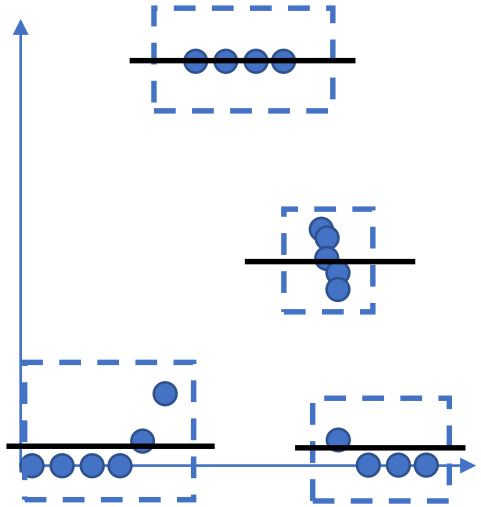
Existe uma variável binária **fumante**, que caso esteja ativa, é considerada um risco para o plano, portanto a faixa de preço do plano muda, custando 75 reais a mais

E se o modelo linear não for adequado?

- Aplicar transformações de escala em variáveis
- Adicionar termos polinomiais ao modelo linear
- Escolher outro tipo de modelo, **não linear**



Modelo linear

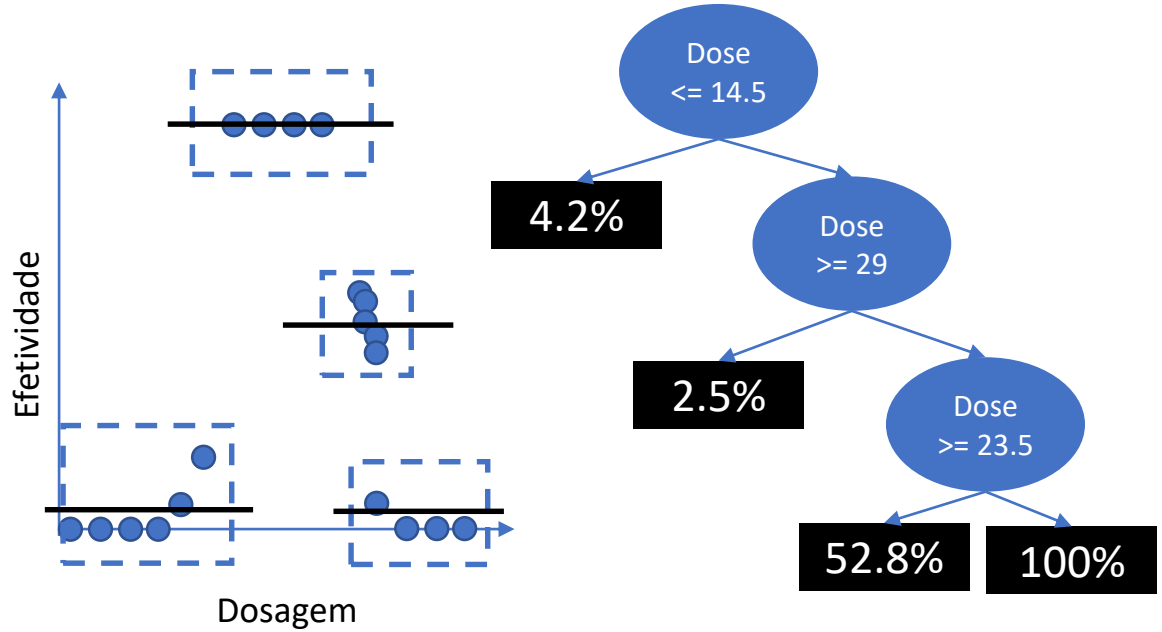


Modelo não linear

Modelos não lineares

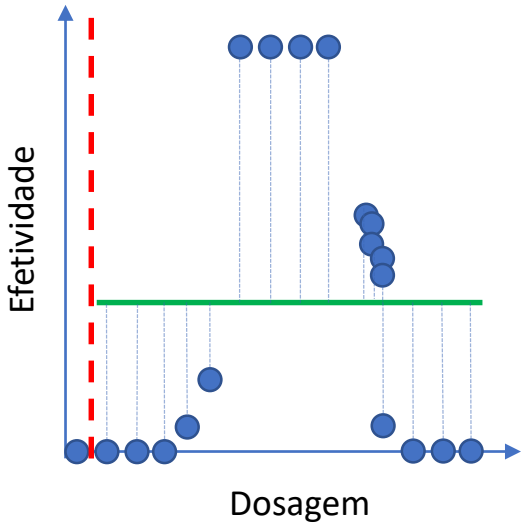
Modelos baseados em árvores de decisão são uma alternativa comum para representar problemas **não lineares**

- A estrutura da árvore permite **organizar o processo de decisão**
- O conhecimento é explícito e interpretável
- A tomada de decisão é, além de entendível, explicável
- É representada na forma de regras **se/então**

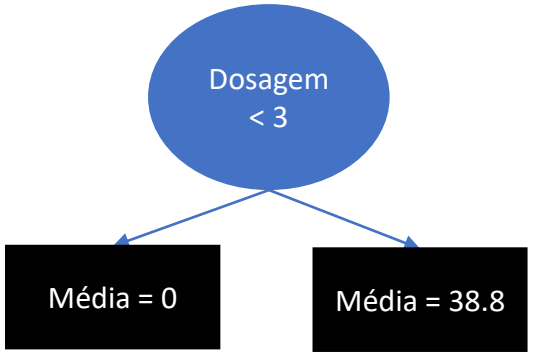


Como construir modelos não lineares

Modelos de árvore são construídos iterativamente, **aplicando cortes** nos valores das variáveis e medindo a **redução no erro** ao criar um novo nó na árvore



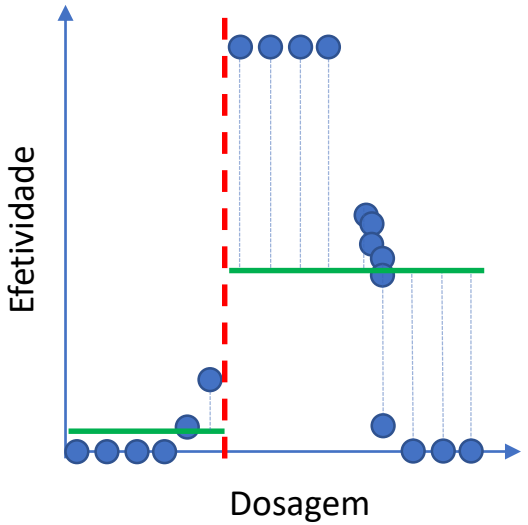
$$SSR = \sum_{i=1}^{n} (x_i - \bar{x})^2$$

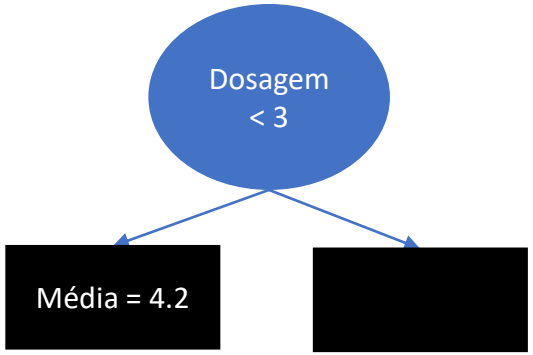


Fonte: https://www.youtube.com/watch?v=g9c66TUylZ4

Como construir modelos não lineares

Modelos de árvore são construídos iterativamente, **aplicando cortes** nos valores das variáveis e medindo a **redução no erro** ao criar um novo nó na árvore

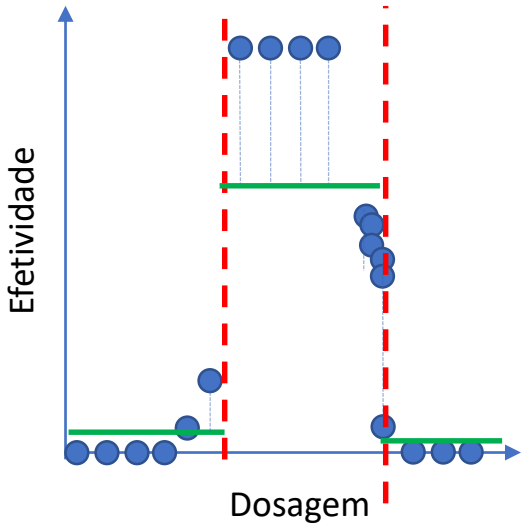


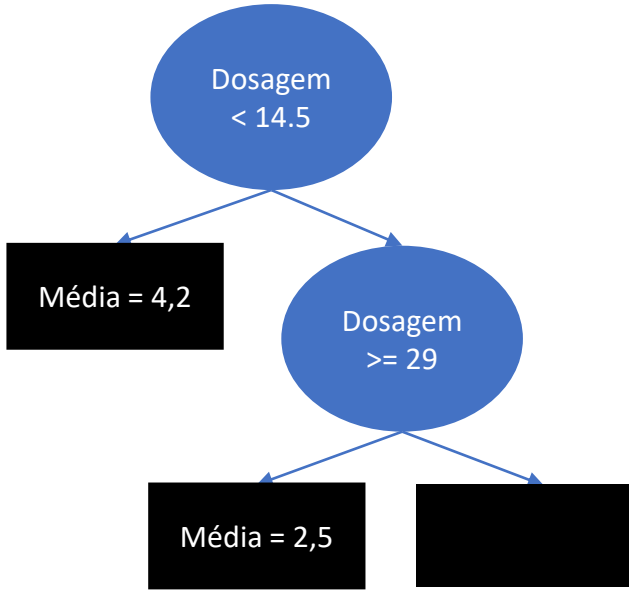


Fonte: https://www.youtube.com/watch?v=g9c66TUylZ4

Como construir modelos não lineares

Modelos de árvore são construídos iterativamente, **aplicando cortes** nos valores das variáveis e medindo a **redução no erro** ao criar um novo nó na árvore

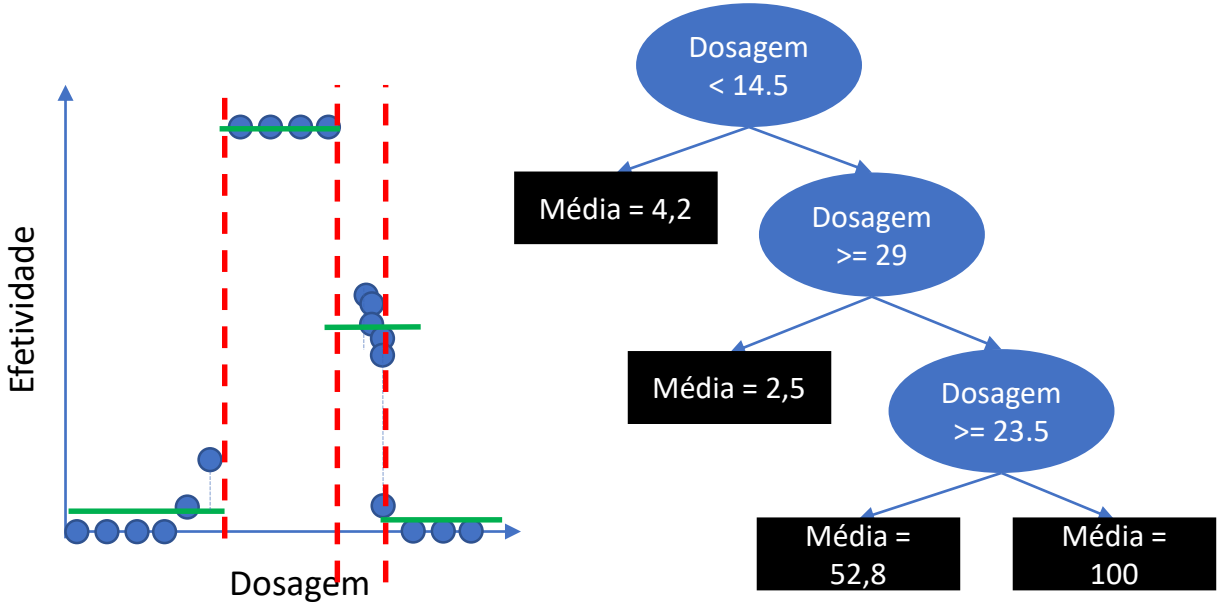




Fonte: https://www.youtube.com/watch?v=g9c66TUylZ4

Como construir modelos não lineares

Modelos de árvore são construídos iterativamente, **aplicando cortes** nos valores das variáveis e medindo a **redução no erro** ao criar um novo nó na árvore



Fonte: https://www.youtube.com/watch?v=g9c66TUylZ4

Como construir modelos descritivos de classificação

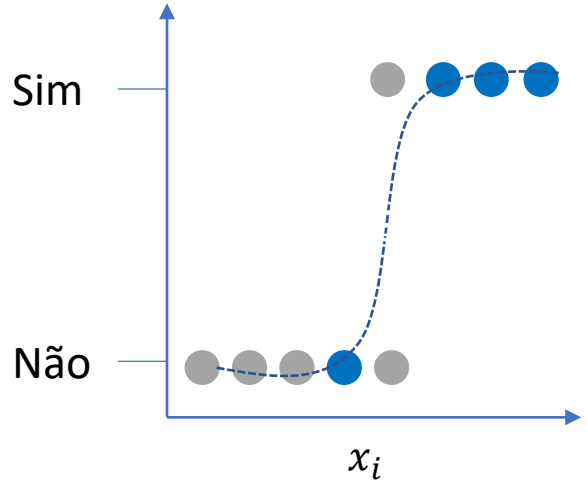
Até o momento, vimos como podemos utilizar a equação da reta e árvores para criar modelos descritivos para resolver **problemas de regressão**. As mesmas ferramentas podem ser utilizadas para resolver **problemas de classificação**, com algumas alterações

- A equação da reta para problemas de classificação é definida como

$$Y_i = \sigma(\beta_0 + \beta_1 \cdot x_i) \text{ para } i = 1,2,3, \dots, n$$

- Onde:
  - Y_i é a variável dependente
  - x_i é a variável independente
  - β_1 é o coeficiente angular
  - β_0 é o intercepto
  - σ é a função sigmoid

Parâmetros a serem estimados



Como construir modelos descritivos de classificação

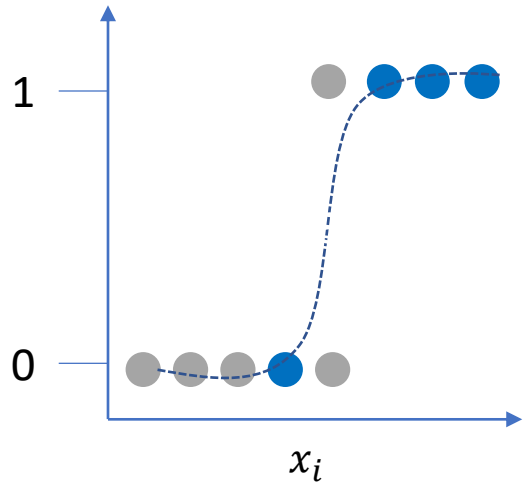
Até o momento, vimos como podemos utilizar a equação da reta e árvores para criar modelos descritivos para resolver **problemas de regressão**. As mesmas ferramentas podem ser utilizadas para resolver **problemas de classificação**, com algumas alterações

- A equação da reta para problemas de classificação é definida como

$$Y_i = \sigma(\beta_0 + \beta_1 \cdot x_i) \text{ para } i = 1,2,3,\dots,n$$

- Onde:
  - Y_i é a variável dependente
  - x_i é a variável independente
  - β_1 é o coeficiente angular
  - β_0 é o intercepto
  - σ é a função sigmoid

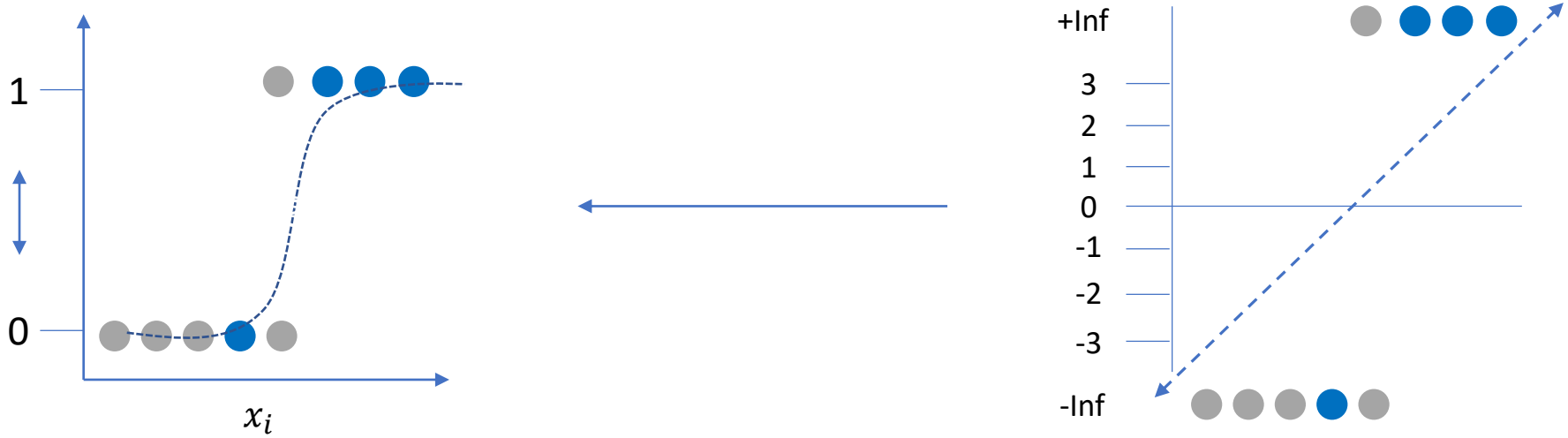
Parâmetros a serem estimados



Como construir modelos descritivos de classificação

Até o momento, vimos como podemos utilizar a equação da reta e árvores para criar modelos descritivos para resolver **problemas de regressão**. As mesmas ferramentas podem ser utilizadas para resolver **problemas de classificação**, com algumas alterações

- No modelo da **regressão logística**, queremos aprender a curva que melhor separa as classes
- Para encontrar essa curva, trabalhamos com uma escala de valores diferenciada, portanto, a interpretação dos fatores não pode ser literal



Como construir modelos descritivos de classificação

Como converter a probabilidade da observação pertencer a uma classe para uma decisão?

Quando trabalhamos com **modelos probabilísticos** para fazer classificação, determinamos um **limiar de decisão**, ou seja, um valor de probabilidade que separa as classes

Por padrão, esse valor é definido como

t = 0,5

Como interpretar o modelo?

Assim como na regressão linear, na regressão logística os parâmetros no modelo de classificação são **estimativas pontuais** e precisamos antes **testar a sua significância.** Além disso, ao interpretar os coeficientes, a análise deve ser **relativa**

Exemplo da equação da reta ajustada

$$E(Y) = \sigma(0{,}125 \cdot \text{dias} - 1{,}5 \cdot ligacao - 0{,}67)$$

Associado a cada parâmetro teremos um **p-valor**, resultante do teste de hipótese

Assumindo $\alpha = 0.05$, **rejeitamos a hipótese nula** de que a estimativa é zero

Podemos interpretar o modelo da seguinte maneira

A variável **ligação** tem coeficiente negativo, o que contribui para evitar que os pacientes faltem às consultas**.** Ela pode ser até 16 vezes mais importante que a quantidade de dias para determinar o **no show**

A variável **dias de antecedência** tem coeficiente positivo, o que indica que conforme a quantidade de dias de antecedência aumenta, os pacientes esquecem e faltam às consultas

Quem marcou a consulta há muito tempo, tem uma alta chance de no show

E se nossa classe tiver mais de 2 valores?

Exemplo: A empresa FRITZ MÜLLER está com problemas de lotação por conta de pacientes com COVID-19. Para planejar melhor a quantidade de UTIs e unidades de tratamento semi-intensivas. Para otimizar o planejamento de abertura e manutenção das unidades, o time de executivos da empresa precisa entender os fatores que levam os pacientes infectados a não ocuparem leitos, precisarem dar entrada em unidades semi-intensivas ou em UTIS.

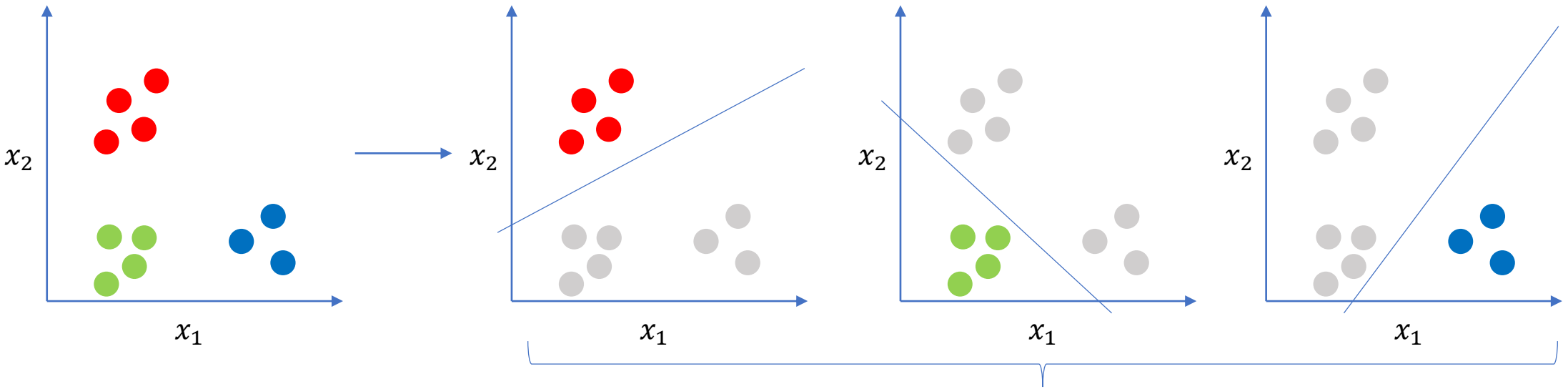
Como estruturar um problema que possui mais de 2 classes?

Para problemas multi-classe, podemos adotar uma estratégia chamada "Um contra todos", onde criamos um modelo por classe.

Para determinar a probabilidade e a classe correta, precisamos avaliar todos os modelos e selecionar o modelo com o melhor resultado

E se nossa classe tiver mais de 2 valores?

Representação visual da estratégia “um contra todos”

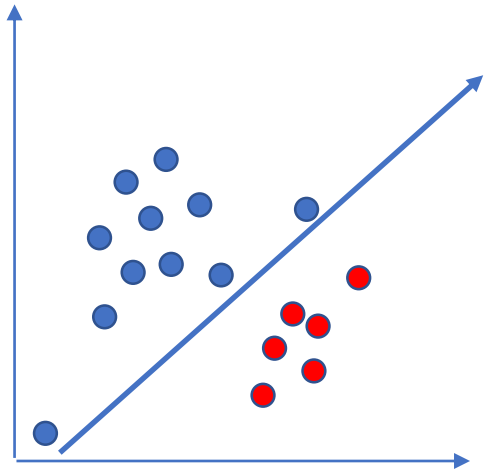


Exemplo de uso:

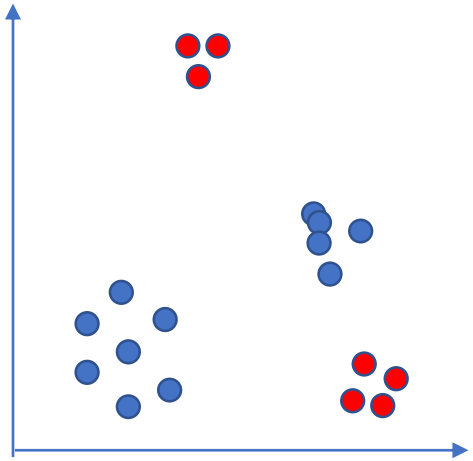
- Modelo 1 = 0,5
- Modelo 2 = 0,6
- Modelo 3 = 0,9

E se o modelo linear não for adequado?

- Aplicar transformações de escala em variáveis
- Adicionar termos polinomiais ao modelo linear
- Escolher outro tipo de modelo, **não linear**



Modelo linear

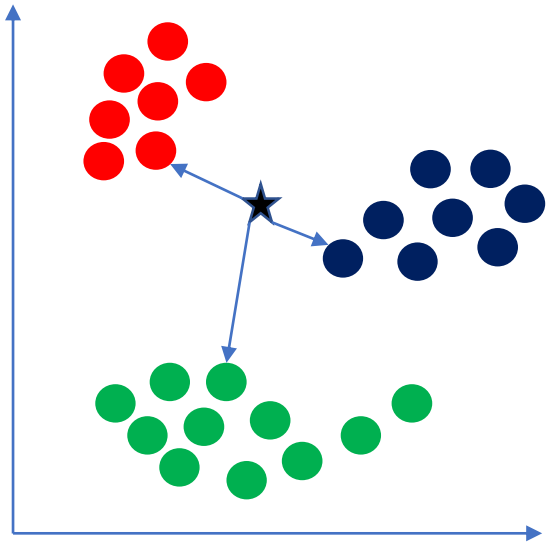


Modelo não linear

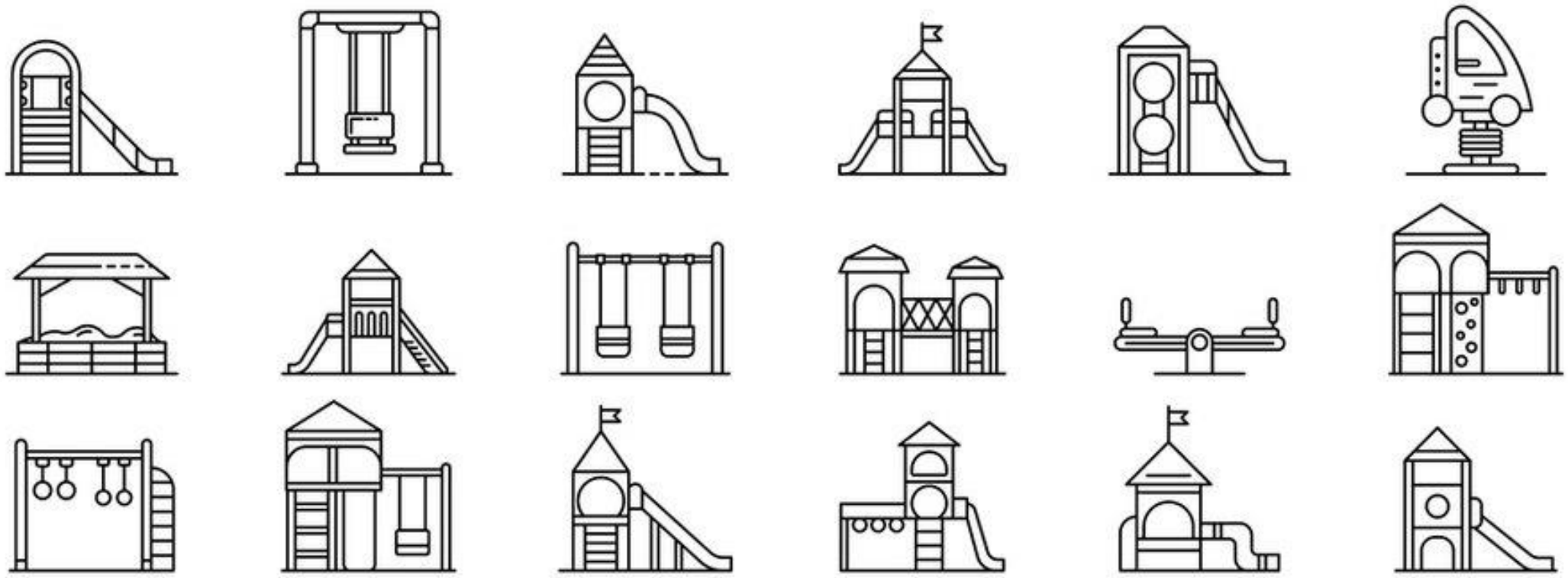
Modelos não lineares

Além de modelos baseados em árvores, outro modelo não linear bastante comum é o **K-Vizinhos mais Próximos**

- É um modelo bastante simples e intuitivo
- O conhecimento é a própria base de dados
- A explicabilidade das decisões é feita com base na interpretação dos exemplos mais similares



Ferramentas interativas



[Tensorflow Playground](#) [ML Playground](#)

Aplicação de um modelo descritivo

- Defina um problema a ser analisado através de um modelo descritivo
- Contextualize o problema, descreva o impacto financeiro e o retorno esperado da solução
- Descreva quais são as variáveis independentes disponíveis, a variável dependente da análise e se há variáveis que ainda precisam ser coletadas
- Levante pontos de risco para a execução da análise na sua instituição – resistência a novas tecnologias, falta de pessoal qualificado, qualidade dos dados, segurança da informação, custos, etc
- Detalhe o time que vai cuidar da análise: consultoria, parceria com universidade? Um ou mais especialistas? Quais especialidades?
- Apresente o seu problema para a turma

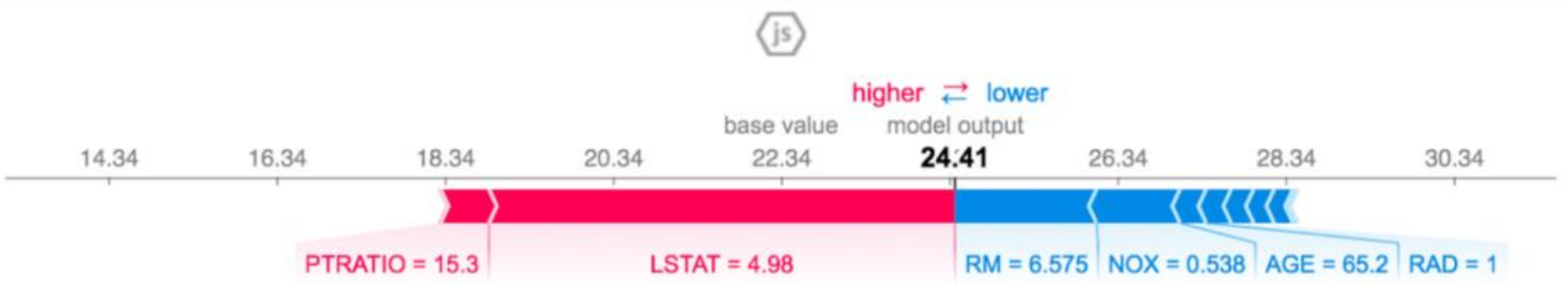
Ressalvas sobre modelos descritivos

- Normalmente, a capacidade de explicações de um modelo é **inversa a sua complexidade**
- Na análise descritiva, utilizamos modelos mais simples para capturar relações, obter insights e entender melhor o impacto dos fatores em determinado sistema
- Através da análise descritiva, estudamos e explicamos um evento que ocorreu, por isso é importante respeitar as premissas dos modelos, para garantir que a interpretação faça sentido

Ressalvas sobre modelos descritivos

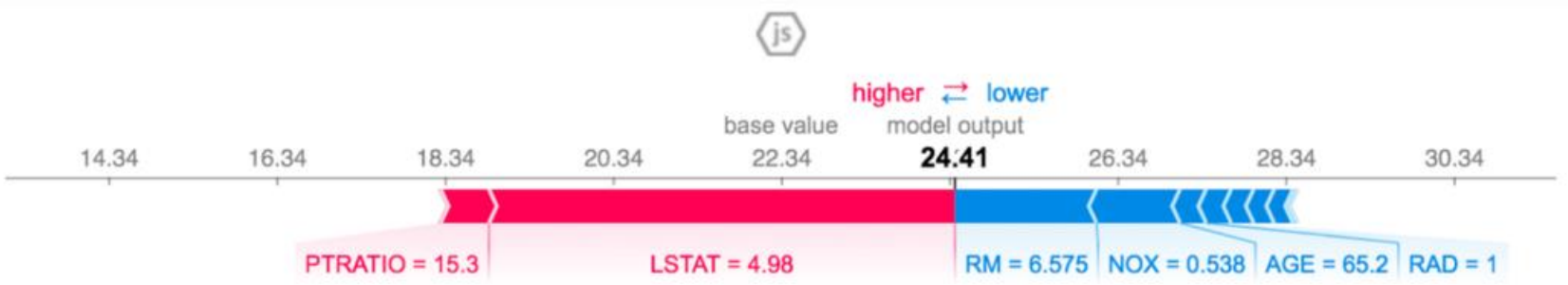
- Além da utilização de modelos descritivos, existem outras formas de interpretar modelos, mesmo modelos mais complexos como **redes neurais**
- Uma ferramenta bastante usada para fazer isso é o **SHAP**
- O SHAP pode ser aplicado a qualquer modelo e nos dá a importância de cada fator na predição da variável dependente

Interpretando valores SHAP



- O gráfico acima é chamado de **gráfico de forças**, aplicado a um modelo treinado para um dataset de previsão de valores imobiliários
- O gráfico mostra, para uma observação, a predição do modelo é **24,41**
- Cada variável usada na predição tem uma barra, onde o tamanho da barra indica o seu efeito e a cor o sinal
- O valor 4,98 para a variável LSTAT fez com que o modelo contribuísse com aproximadamente 6 (tamanho da barra) à resposta do modelo

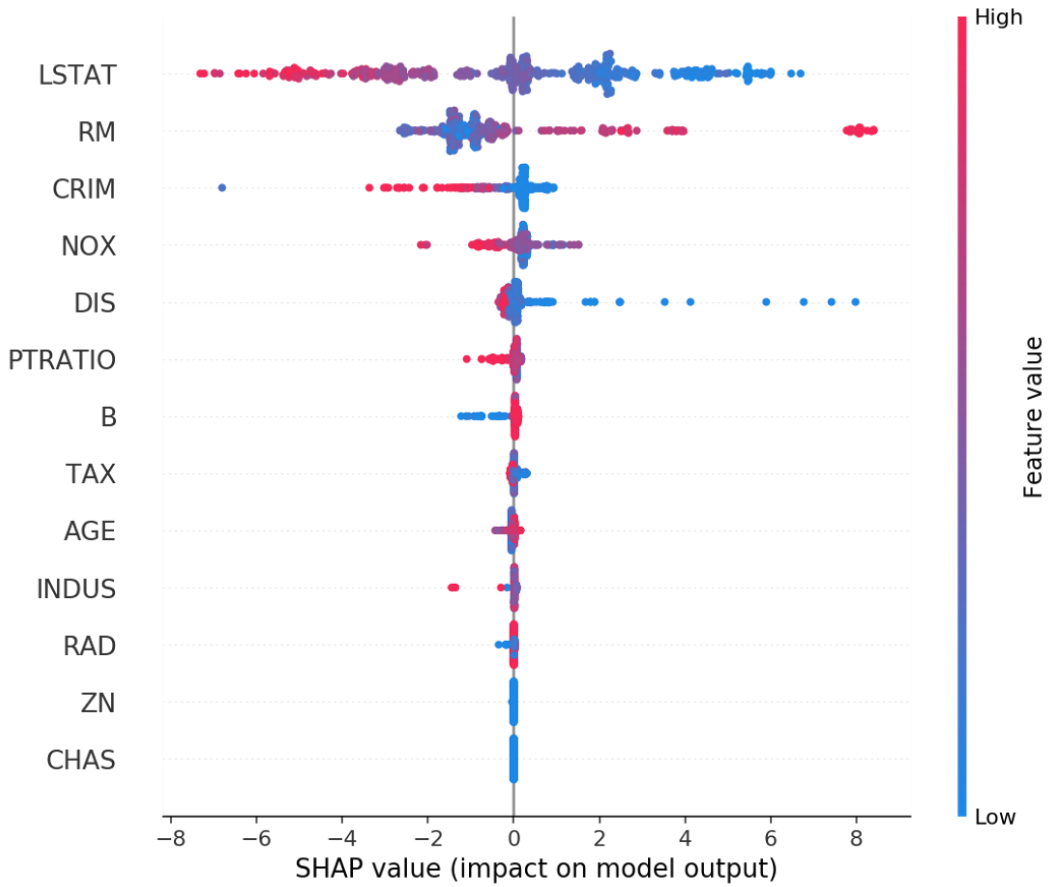
Interpretando valores SHAP



- Os valores SHAP tem uma unidade, que é a mesma da variável de resposta
- É importante ressaltar que o fato da variável ter um efeito positivo **nesta observação** não implica que ele será positivo para todas outras observações
- Também não significa que, ao aumentar o valor da variável, o efeito aumentaria
- O valor SHAP só mostra qual é a importância para a observação específica, ao assumir esses valores das variáveis

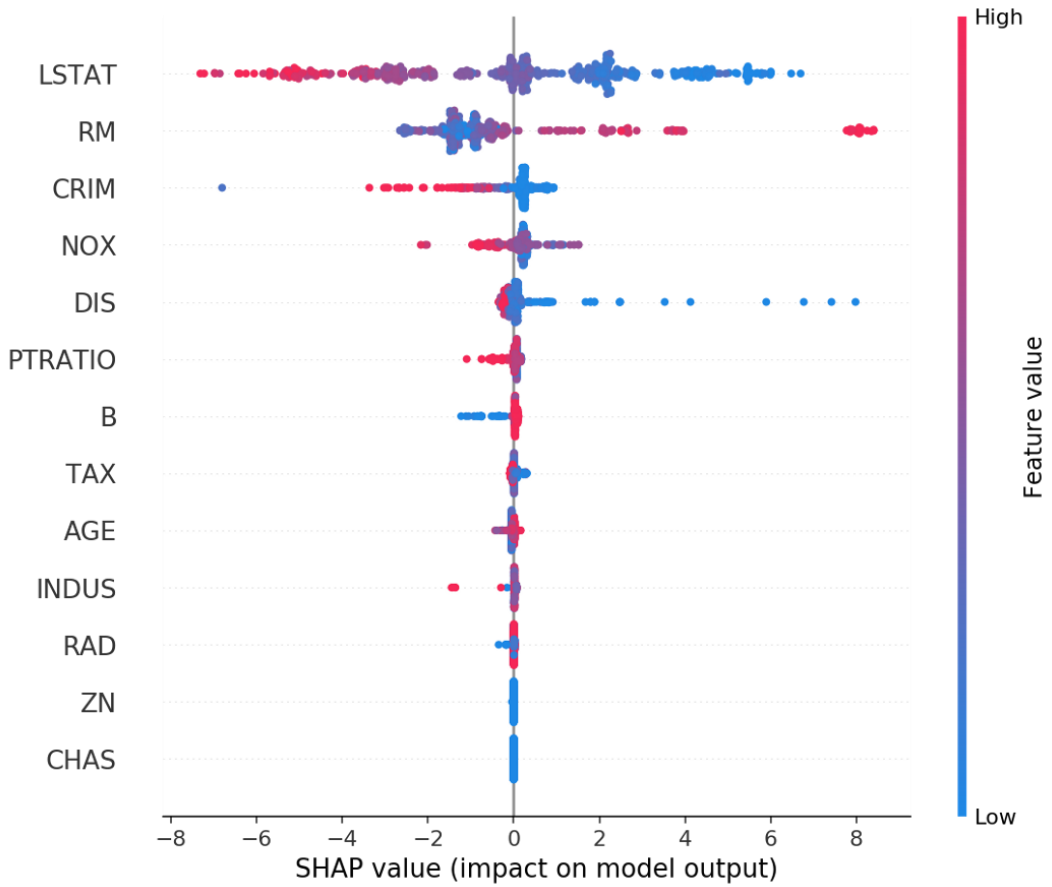
Interpretando valores SHAP

- Para extrair informações globais a partir das importâncias das observações, podemos usar o **gráfico resumo**
- Cada ponto de cada linha representa uma observação
- Pontos azuis significam que aquela observação possui um valor baixo para a variável, enquanto os vermelhos significam um valor mais alto



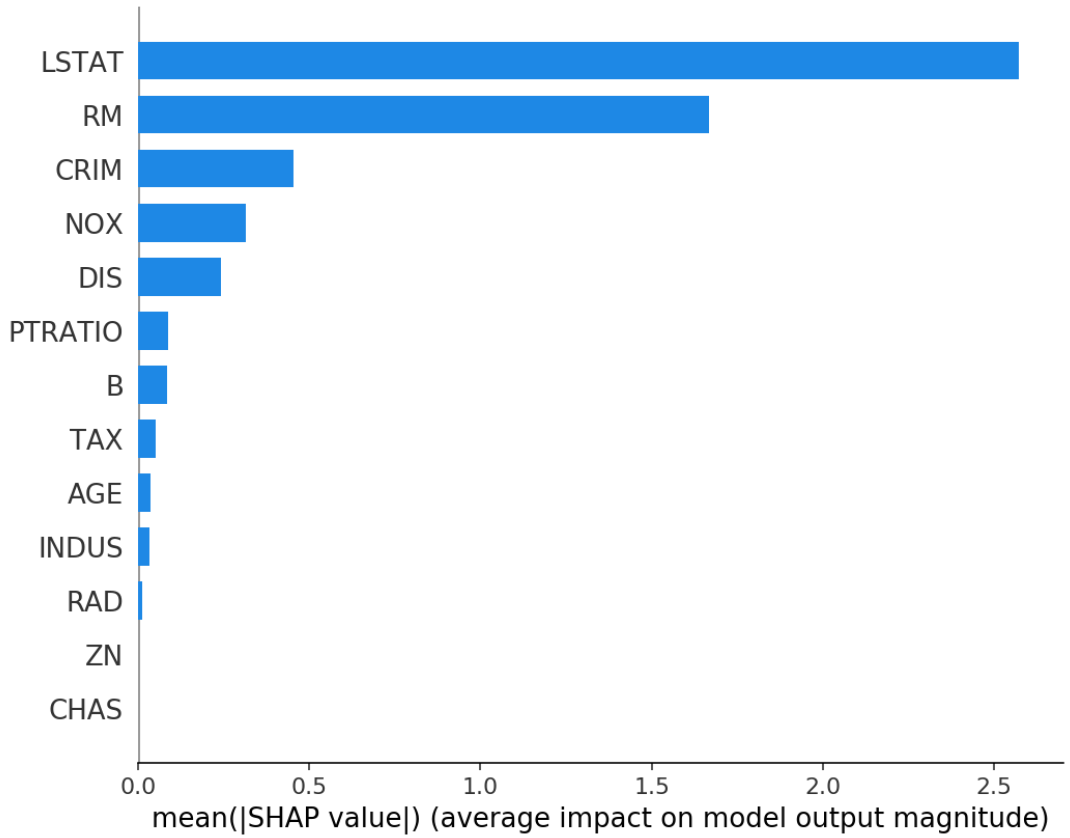
Interpretando valores SHAP

- A posição no eixo x indica o efeito daquela observação
- Quanto mais a direita, mais positiva é a contribuição da variável para a observação
- As variáveis são ordenadas pela sua importância
- Podemos observar que para a variável LSTAT, quanto maior os seus valores, mais negativa é a sua contribuição para o resultado do modelo



Interpretando valores SHAP

- Para visualizar a importância global das variáveis, podemos usar o **gráfico de resumo em barras**
- No gráfico, a importância global das variáveis é calculada a partir das importâncias locais como a média dos valores absolutos das contribuições



Interpretando valores SHAP

- SHAP é uma ferramenta útil para obter a importância das variáveis independente do modelo utilizado
- Uma vantagem do SHAP é que seus valores estão na mesma unidade da variável dependente, facilitando a sua interpretação
- Os gráficos disponibilizados pela ferramenta nos permitem entender melhor os nossos modelos e extrair informações sobre o seu comportamento

Escopo da disciplina

Identificar e entender problemas tratáveis por modelos de aprendizado de máquina

O que é um modelo preditivo?

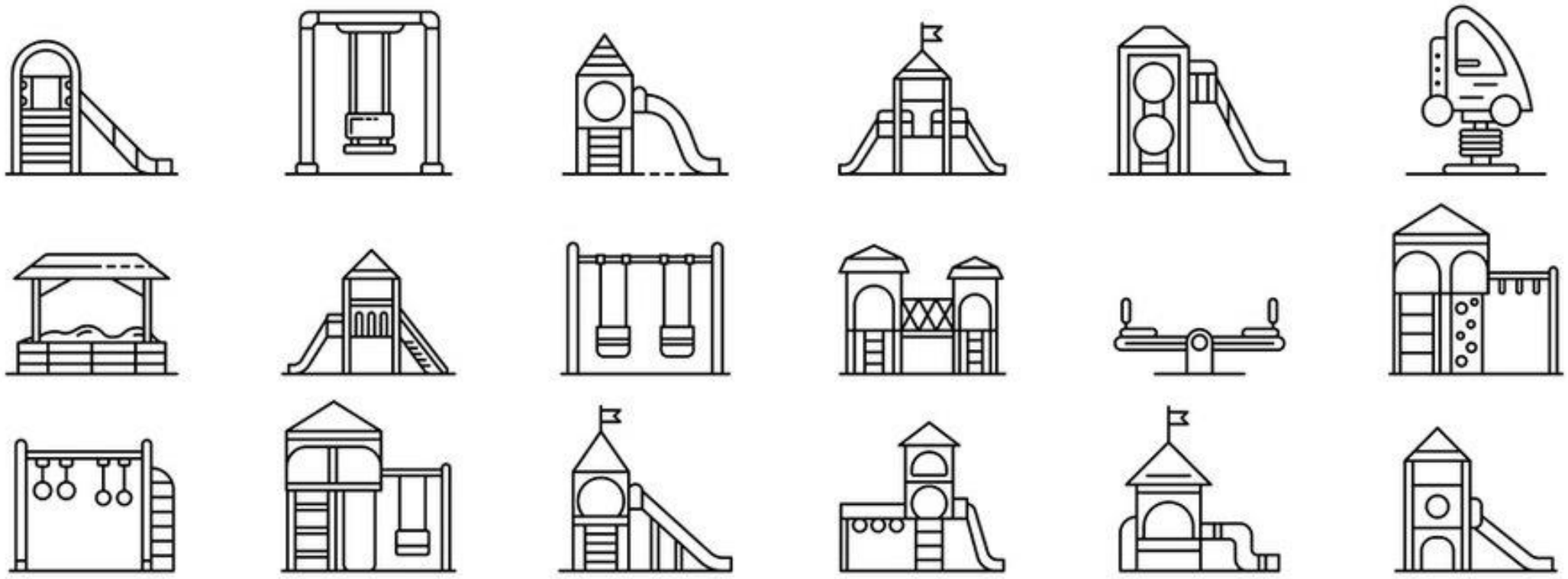
Exemplo: A empresa FRITZ MÜLLER está enfrentando um problema de superlotação da sua unidade de tratamento intensiva de forma recorrente. Os gerentes da empresa desejam saber com 1 semana de antecedência a demanda de uso das UTIs para evitar prejuízos.

Modelos preditivos utilizam os padrões entre as variáveis independentes e a variável dependente para prever resultados futuros

Comparação da abordagem descritiva e preditiva

Explicação de eventos passados	Previsão de eventos futuros
Pode conter dados coletados após a obtenção da variável dependente	Não pode conter vazamento de dados para não enviesar o modelo
Respeito às premissas dos algoritmos; avaliação dos pesos	Não é dada importância às premissas desde que os resultados do modelo sejam precisos
Modelagem simples, mas interpretável	Modelagem complexa, na maioria dos casos não interpretável
Não há necessidade de tentar generalizar os resultados	Objetivo de generalizar para conjuntos de dados não vistos anteriormente

Alguns exemplos de modelos preditivos



Reconhecimento de dígitos

Reconhecimento de sketchs

Classificação de imagens

Quais algoritmos podemos utilizar?

Podemos utilizar os **mesmos algoritmos** que vimos aplicados à modelagem descritiva

Apesar disso, **o processo de construção dos modelos é diferente** para as duas abordagens

Na modelagem preditiva, normalmente **dividimos nosso conjunto de dados**

Por que dividir os conjuntos de dados?

- O objetivo da modelagem preditiva é encontrar um algoritmo capaz de gerar modelos que **generalizem** além do conjunto de dados para o qual foram treinados
- Para fazer isso é comum criar um experimento para avaliar diferentes algoritmos, com suas respectivas parametrizações, aplicados a um mesmo conjunto de dados
- Existem diferentes abordagens, que podem variar de acordo com a complexidade do problema e o volume de dados disponível

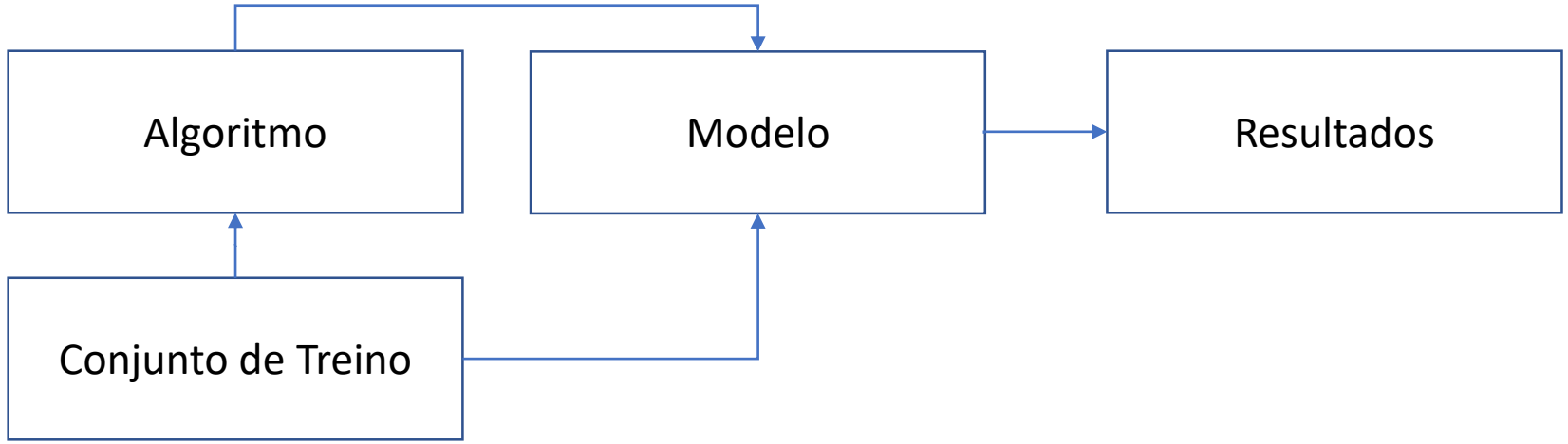
Estratégias para a divisão dos conjuntos de dados

Existem diferentes estratégias para dividirmos o nosso conjunto de dados e validarmos os resultados dos nossos modelos

As três estratégias mais comuns são: **resubstituição**, **holdout** e **cross validation**

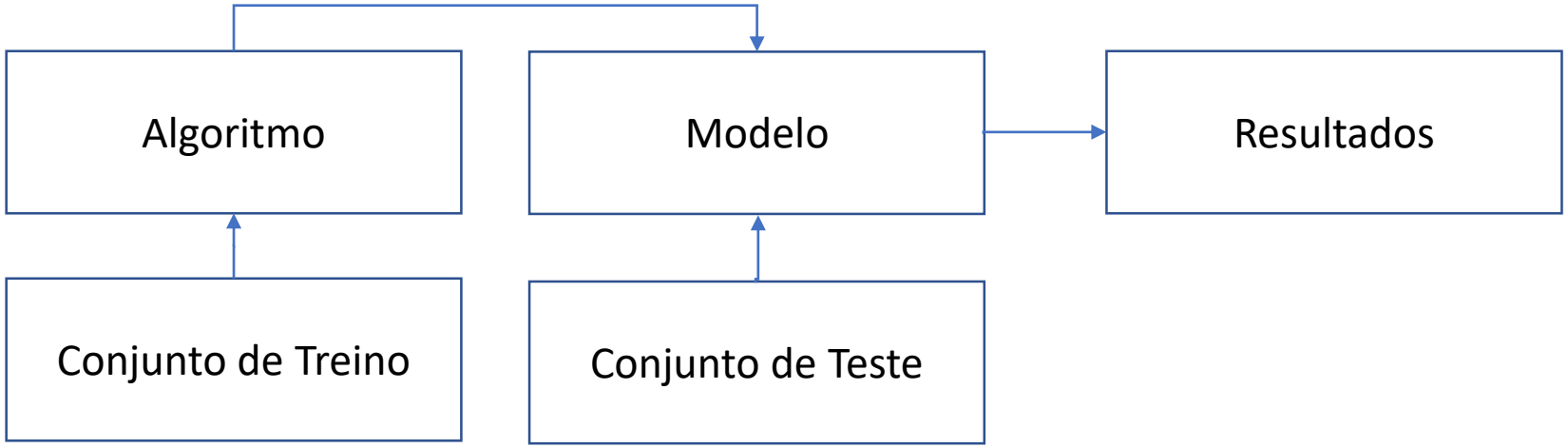
Estratégias para a construção de modelos preditivos – Resubstituição

O método de **resubstituição** consiste em construir o modelo e testar seu desempenho no mesmo conjunto de dados, ou seja, o conjunto de teste é idêntico ao conjunto de treino. Este estimador é enviesado e possui uma estimativa altamente otimista



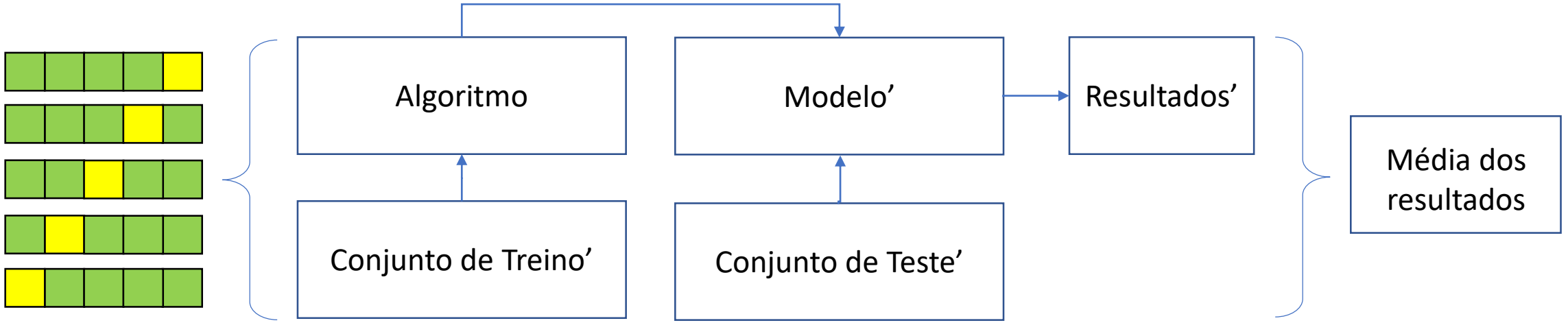
Estratégias para a construção de modelos preditivos – Holdout

O método **holdout** divide os exemplos em uma porcentagem fixa de exemplos p para treinamento e $(1-p)$ para teste, considerando normalmente $p > 0.5$. Valores típicos são $p = 0.70$ e $(1-p) = 0.30$, embora não exista nenhum fundamento sobre esses valores



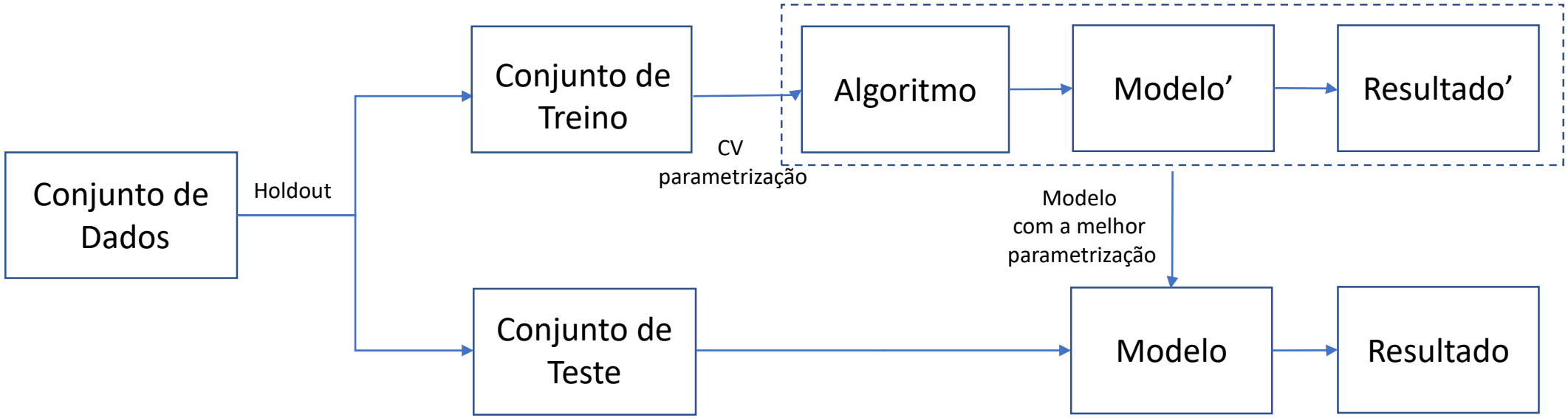
Estratégias para a construção de modelos preditivos – Cross validation

No método **cross-validation**, os exemplos são aleatoriamente divididos em k partições mutuamente exclusivas de tamanho aproximadamente igual. Os exemplos nas $(k-1)$ partições são usados para treinamento e o modelo gerado é testado na partição remanescente. Este processo é repetido k vezes, cada vez considerando uma partição diferente para teste. A medida final é calculada como a média do resultado de cada partição



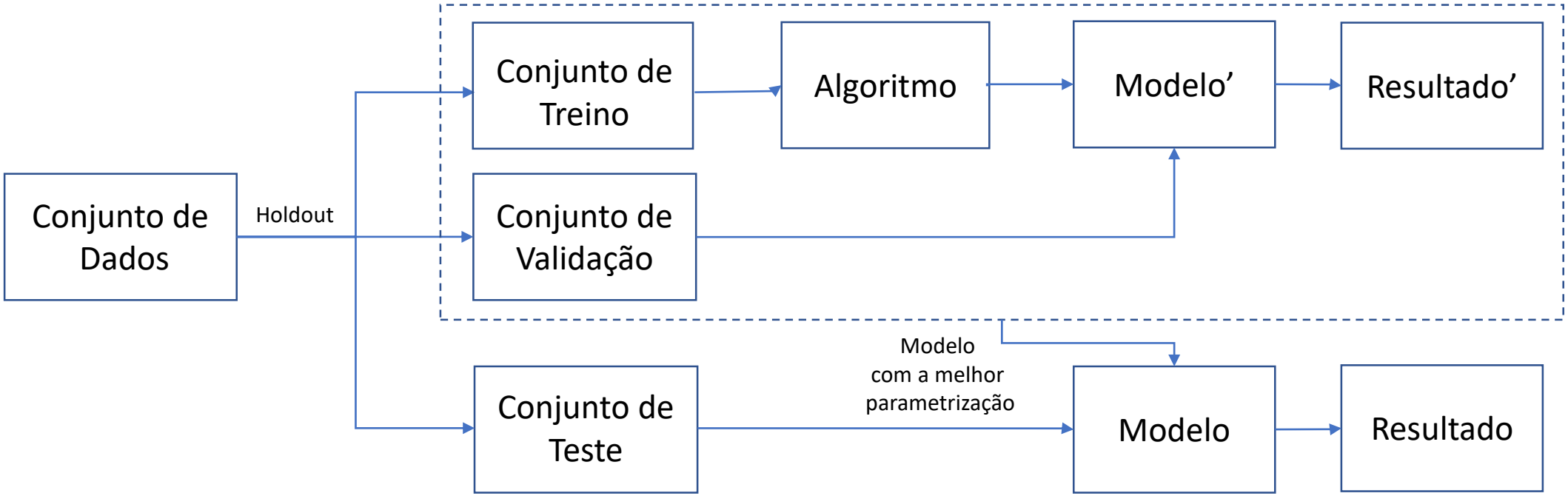
Design de experimento para construir modelos preditivos – Volume moderado de dados

No design do experimento, é necessário avaliar múltiplos algoritmos, cada um com sua própria parametrização. Por conta disso, para não enviesar nossa decisão, é comum criar um terceiro conjunto de dados, para **validação dos parâmetros** do modelo. Podemos fazer isso combinando os **métodos vistos anteriormente**



Design de experimento para construir modelos preditivos – Volume grande de dados

No design do experimento, é necessário avaliar múltiplos algoritmos, cada um com sua própria parametrização. Por conta disso, para não enviesar nossa decisão, é comum criar um terceiro conjunto de dados, para **validação dos parâmetros** do modelo. Podemos fazer isso combinando os **métodos vistos anteriormente**

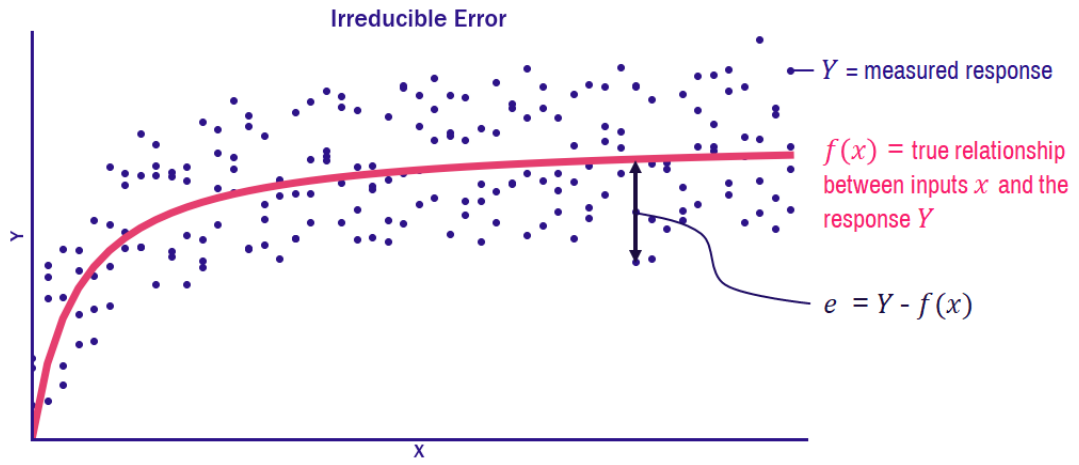


Depois de estruturar o experimento, como escolher um modelo?

Para entendermos como **escolher um bom modelo**, precisamos antes entender os conceitos de **viés** e **variância** aplicados à modelagem preditiva

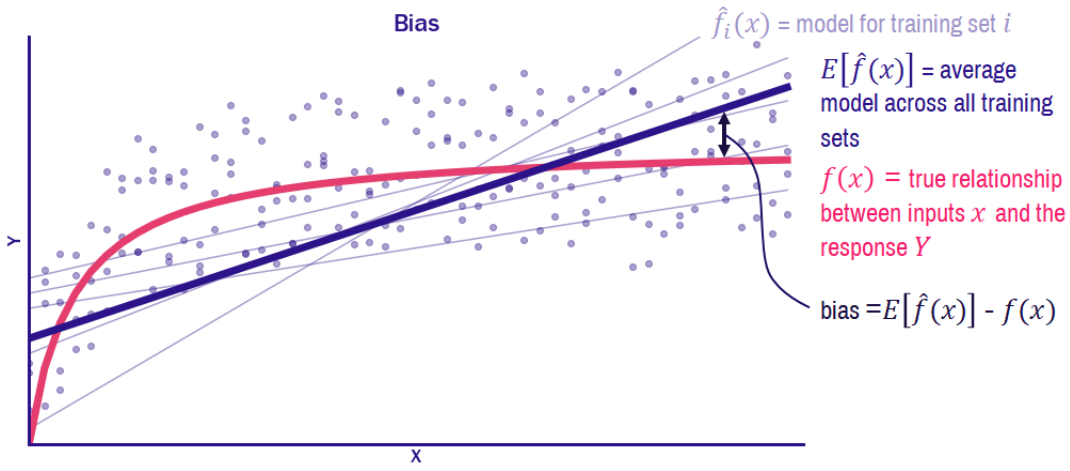
Viés e variância

- Para entender os conceitos de **viés** e **variância**, é necessário relacioná-los com o **erro irredutível**, ou erro aleatório
- O erro irredutível surge de dados faltantes, erros na medição da variável ou são relacionados ao próprio fenômeno sendo estudado
- Matematicamente, o erro irredutível é a diferença entre o resultado observado e o resultado do relacionamento verdadeiro entre as variáveis independentes e a variável dependente



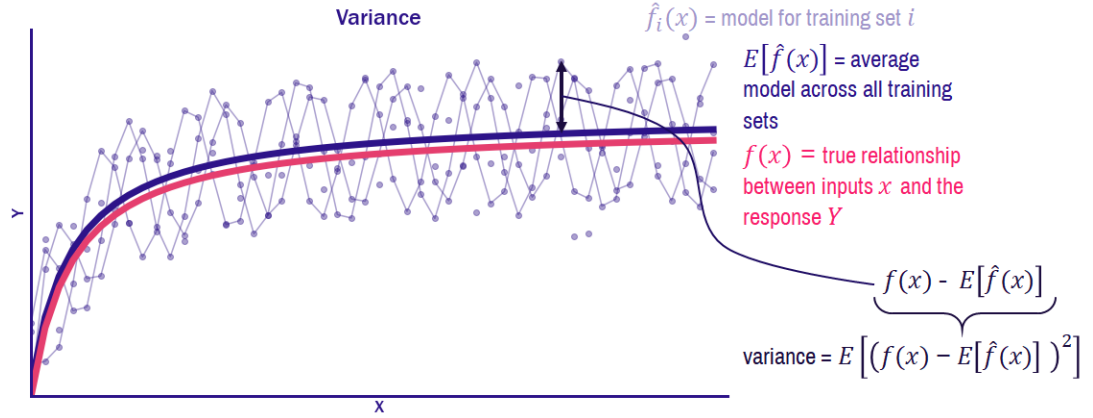
Viés e variância

- Além do **erro irredutível**, os erros do nosso modelo consistem em erros de **viés** ou de **variância**
- **Viés** é a inabilidade do modelo de aprender o relacionamento das variáveis independentes X e a variável dependente Y
- Um modelo com **viés,** para diferentes conjuntos de dados, na média, apresentará um alto valor de erro
- Também chamamos o problema de **viés** de ***underfitting*.** O modelo apresenta uma baixa performance no **conjunto de treino**



Viés e variância

- A **variância** quantifica a tendência do modelo a aprender muito sobre o relacionamento das variáveis independentes X e a variável dependente Y
- Modelos com **alta variância** aprendem ao ponto de capturar a aleatoriedade dos dados. Em outras palavras, o modelo decora o conjunto de dados. Esse fenômeno é chamado de ***overfitting***
- Modelos que sofrem de ***overfitting*** possuem uma tendência de generalizar muito mal para outros conjuntos de dados, resultando em baixa performance no **conjunto de teste**

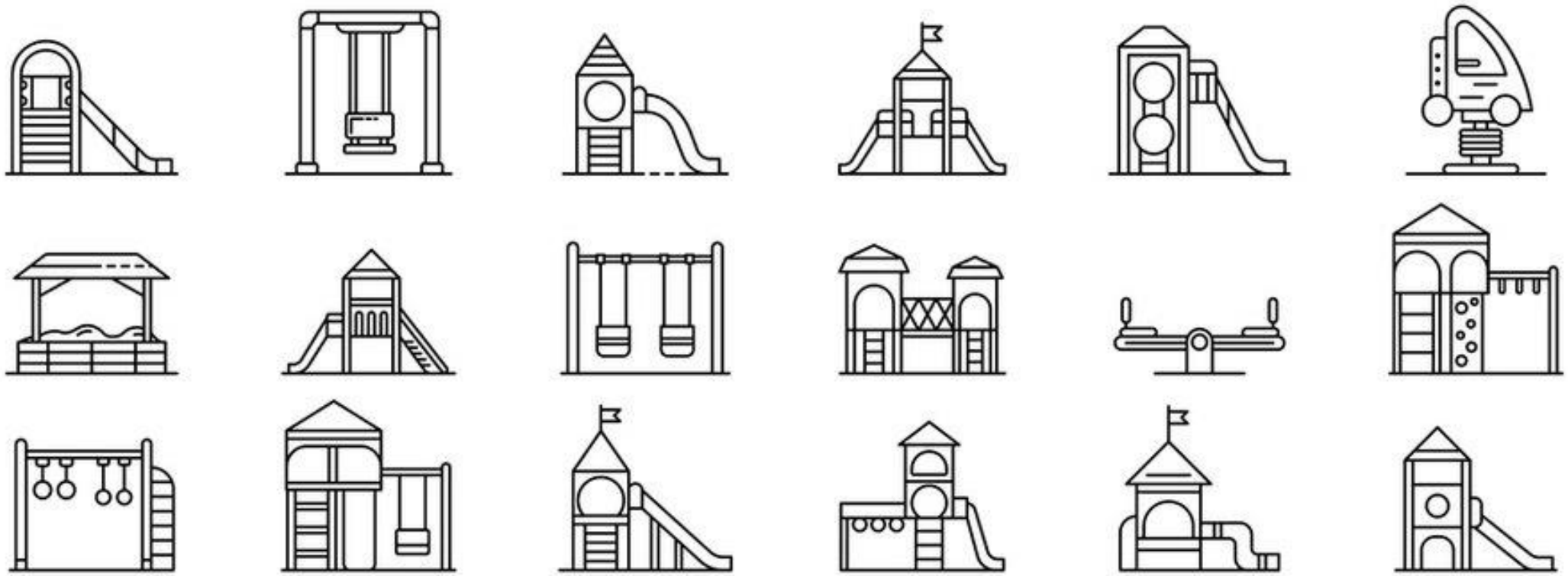


Tradeoff entre viés e variância

- Na prática nós temos que escolher **simplificar o relacionamento** entre as variáveis (reduzindo a variância, mas introduzindo viés) ou tentar **capturar mais o relacionamento** entre as variáveis (reduzindo o viés, mas aumentando a variância)
- Nós identificamos problemas com viés e variância através dos conjuntos de treino e teste
- Existem 4 possíveis combinações de viés e variância para caracterizar um modelo

Variância	**Viés**	
	Alto	**Baixo**
Alto	Péssimo	*Overfitting*
Baixo	*Underfitting*	Ótimo

Visualizando viés e variância na modelagem



Viés

Variância

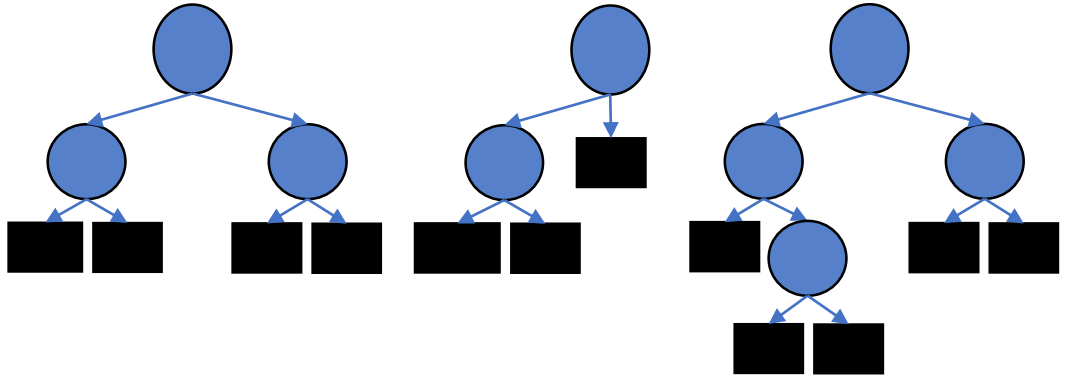
Como obter resultados melhores?

Uma estratégia bastante comum é utilizar **ensembles** (comitês) para a tomada de decisão

Ao invés de utilizar o resultado de apenas um modelo, **agregamos o resultado de múltiplos modelos**

Florestas aleatórias

- **Florestas aleatórias** consistem em criar múltiplas **árvores de decisão** para fazer uma votação, melhorando os resultados do modelo
- O **número de árvores** é um **parâmetro**
- Cada árvore é gerada a partir de uma **amostra de dados** e uma **amostra de variáveis** do conjunto de dados original
- Os resultados podem ser calculados como a **soma dos votos absolutos** (hard voting) ou a **média dos resultados** (soft voting)



$A1 = 1$

$A2 = 0 \rightarrow Resultado = 0$

$A3 = 0$

Hard voting

$A1 = 1$ com 99%

$A2 = 1$ com 49% $\rightarrow Resultado = 65{,}67\%$

$A3 = 1$ com 49%

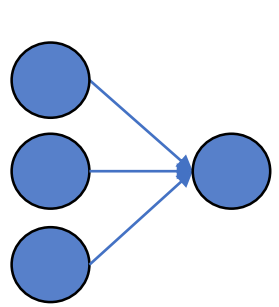
$Resultado = \frac{(99 + 49 + 49)}{3}$

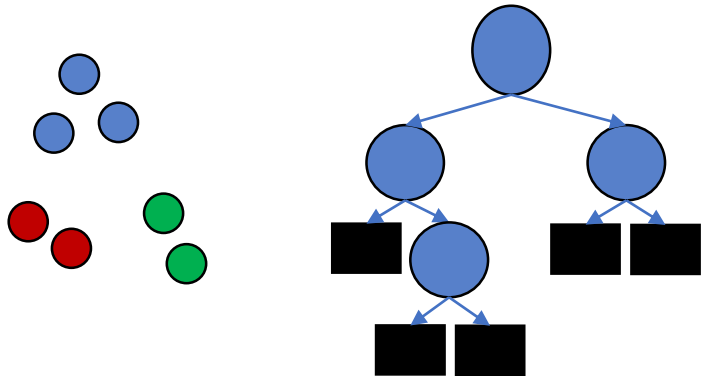
$Resultado = 1$

Soft voting

Bagging

- O algoritmo de **bagging** consiste em escolher **qualquer algoritmo de aprendizado de máquina** para compor o ensemble
- Cada modelo pode ser gerado a partir de uma **amostra de dados** do conjunto de dados original
- Assim como no algoritmo de **floresta aleatória**, o bagging assume como resultado a **soma dos votos absolutos** ou a **média dos resultados**





$A1 = 1$

$A2 = 0 \rightarrow Resultado = 0$

$A3 = 0$

Hard voting

$A1 = 1$ com 99%

$A2 = 1$ com 49% $\rightarrow Resultado = 65{,}67\%$

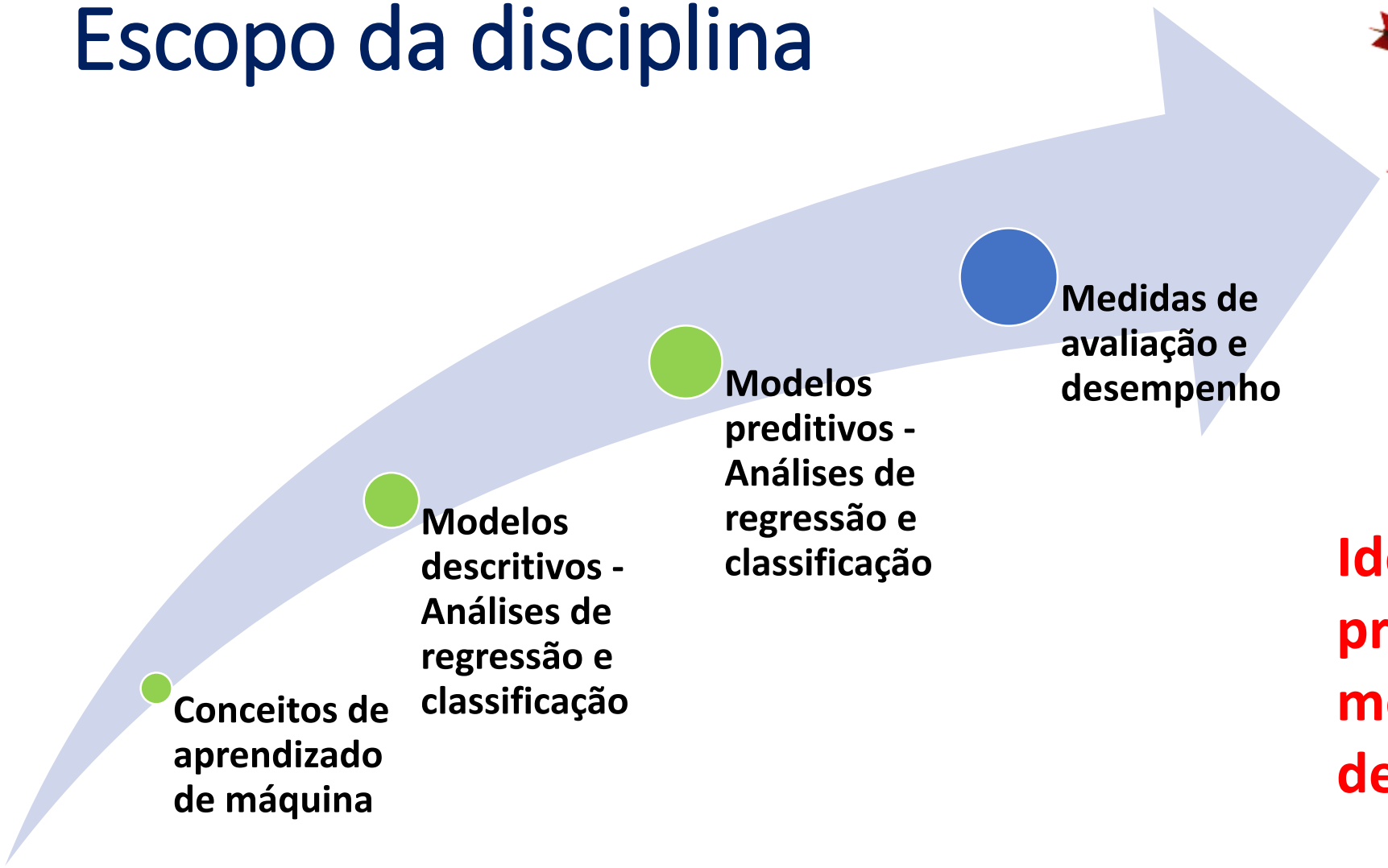
$A3 = 1$ com 49%

$$Resultado = \frac{(99 + 49 + 49)}{3}$$

$Resultado = 1$

Soft voting

Escopo da disciplina



Identificar e entender problemas tratáveis por modelos de aprendizado de máquina

O que são medidas de avaliação?

Uma medida de avaliação **quantifica a performance** de um modelo preditivo, o que envolve comparar o resultado esperado e predito

Cada tarefa de aprendizado de máquina possui suas **próprias medidas de avaliação**

Erro médio quadrático

- O **erro médio quadrático** é a média da diferença entre o valor observado e o predito, elevado ao quadrado
- Elevamos o erro ao quadrado para normalizar os sinais e penalizar os maiores erros
- Por elevar o erro ao quadrado, normalmente é uma medida usada para treinar modelos, penalizando grandes erros
- Seu grande problema é a sua **falta de interpretabilidade**

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

- n é o número de observações
- y_i é o valor da observação i
- $\hat{y}_i$ é o valor predito de i

Raiz do erro médio quadrático

- A **raiz do erro médio quadrático** aplica uma função de raiz quadrada ao **erro médio quadrático**
- Aplicando a raiz, resolvemos o problema de **interpretabilidade**, corrigindo a unidade de medida

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i))^2}$$

- n é o número de observações
- y_i é o valor da observação i
- $\hat{y}_i$ é o valor predito de i

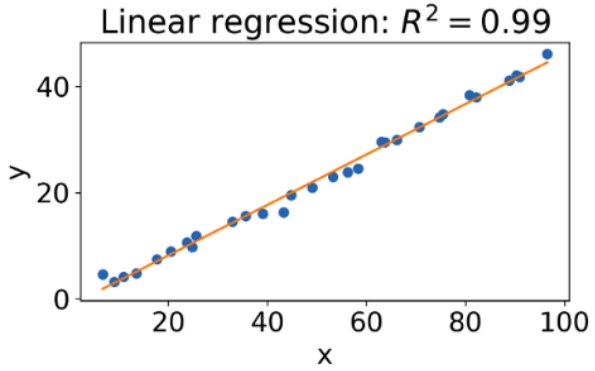
R^2

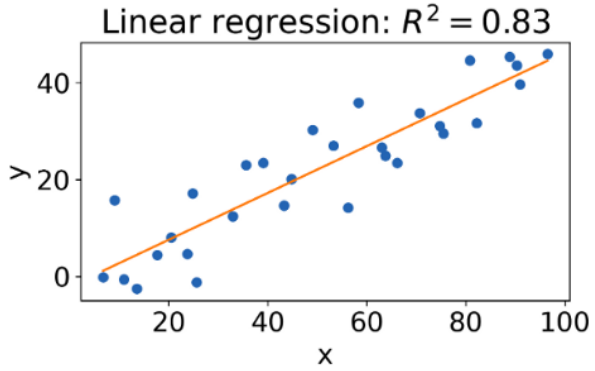
- O **coeficiente de variação** representa a quantidade de variância que é explicada pelo modelo
- Em outras palavras, a medida calcula o percentual da variância que pode ser prevista pelo modelo
- O valor varia entre 0 e 1

$$R^2 = 1 - \frac{Variância\ Residual}{Variância\ Total}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- n é o número de observações
- y_i é o valor da observação i
- $\hat{y}_i$ é o valor predito de i





Avaliação de problemas de classificação

- Uma forma de extrair as medidas de desempenho de problemas de classificação é através de uma **matriz de confusão**
- Dado um problema de duas classes (positiva P e negativa N), a matriz de confusão é definida da forma ao lado, onde:
  - **Verdadeiro Positivo** é o número de exemplos positivos classificados como positivos
  - **Falso Positivo** é o número de exemplos negativos classificados como positivos
  - **Verdadeiro Negativo** é o número de exemplos negativos classificados como negativos
  - **Falso Negativo** é o número de exemplos positivos classificados como negativos

Classe	Predita P	Predita N
Observada P	VP	FN
Observada N	FP	VN

Acurácia

- A **acurácia** é a proporção de exemplos classificados corretamente dentre o total de observações avaliadas
- É uma medida global bastante comum para problemas de classificação, mas só é uma escolha válida para problemas em que as **classes são balanceadas**
- Em um exemplo de avaliação com 200 observações (ao lado), temos uma acurácia de **0,85**

$$accuracy = \frac{VP + VN}{VP + FN + FP + VN}$$

Classe	Predita P	Predita N
Observada P	VP = 90	FN = 10
Observada N	FP = 20	VN = 80

Sensibilidade / Recuperação / TPR

- A **sensibilidade** é a proporção de exemplos classificados corretamente dentre as observações positivas
- É uma escolha válida para os casos que queremos capturar o maior número possível de casos positivos
- Em um exemplo de avaliação com 200 observações (ao lado), temos uma sensibilidade de **0,90**

$$sensibility = \frac{VP}{VP + FN}$$

Classe	Predita P	Predita N
Observada P	VP = 90	FN = 10
Observada N	FP = 20	VN = 80

Especificidade / TNR

- A **especificidade** é a proporção de exemplos classificados corretamente dentre as observações negativas
- É uma escolha válida para os casos que queremos capturar o maior número possível de casos negativos
- Em um exemplo de avaliação com 200 observações (ao lado), temos uma especificidade de **0,80**

$$specificity = \frac{VN}{VN + FP}$$

Classe	Predita P	Predita N
Observada P	VP = 90	FN = 10
Observada N	FP = 20	VN = 80

Precisão

- A **precisão** é a proporção de exemplos classificados corretamente dentre as observações **classificadas como positivas**
- É uma escolha válida para os casos que que o custo de um **falso positivo** é maior que um **falso negativo**
- Em um exemplo de avaliação com 200 observações (ao lado), temos uma precisão de **0,82**

$$precision = \frac{VP}{VP + FP}$$

Classe	Predita P	Predita N
Observada P	VP = 90	FN = 10
Observada N	FP = 20	VN = 80

Medida F

- A **medida F** é a média entre a **precisão** e a **recuperação**
- É uma escolha válida para os casos em que o custo de um **falso positivo** é igual ao de um **falso negativo**
- Em um exemplo de avaliação com 200 observações (ao lado), temos uma medida f de **0,86**

Classe	Predita P	Predita N
Observada P	VP = 90	FN = 10
Observada N	FP = 20	VN = 80

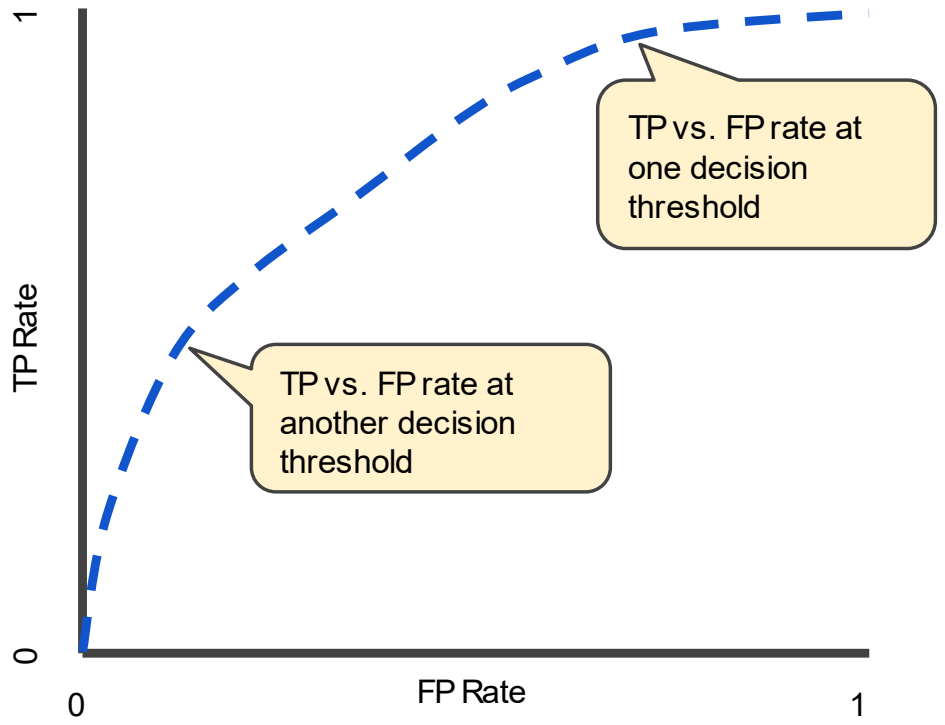
Medidas de posição para problemas com múltiplas classes

- Quando trabalhamos com **classificação binária**, normalmente estamos mais interessados em medidas **referentes a classe positiva**
- Quando trabalhamos com **múltiplas classes** ou queremos **representar ambas as classes**, precisamos aplicar as medidas locais por classes e agregar os seus resultados
- É comum utilizar **médias aritméticas** ou **médias ponderadas**
- Em um exemplo de avaliação com 200 observações (ao lado), temos uma precisão de **0,82 + 0,88 / 2 = 0,85**

Classe	Predita P	Predita N
Observada P	VP = 90	FN = 10
Observada N	FP = 20	VN = 80

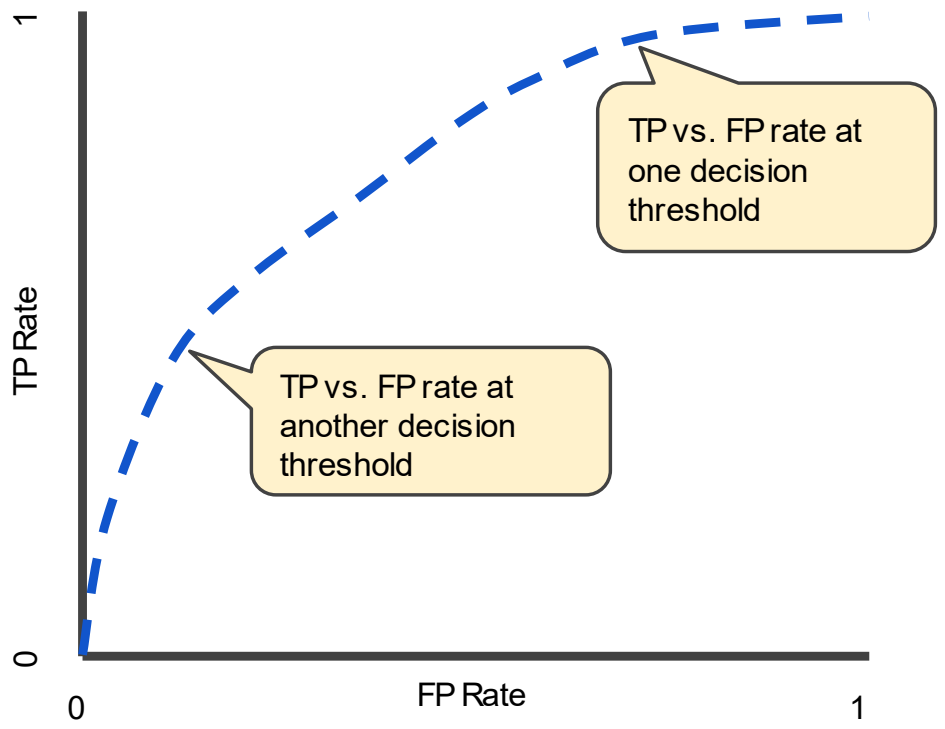
Receiver Characteristic Operation (ROC)

- Quando trabalhamos com problemas de classificação, aplicamos **limiares** sobre um valor de probabilidade para decidir se uma predição é **positiva** ou **negativa**
- Ao invés de testar vários limiares, gerar suas matrizes de confusão e calcularmos suas métricas, podemos resumir esse processo em um gráfico da curva ROC
- O gráfico ROC nos mostra a performance de um modelo de classificação para **todos os possíveis limiares**



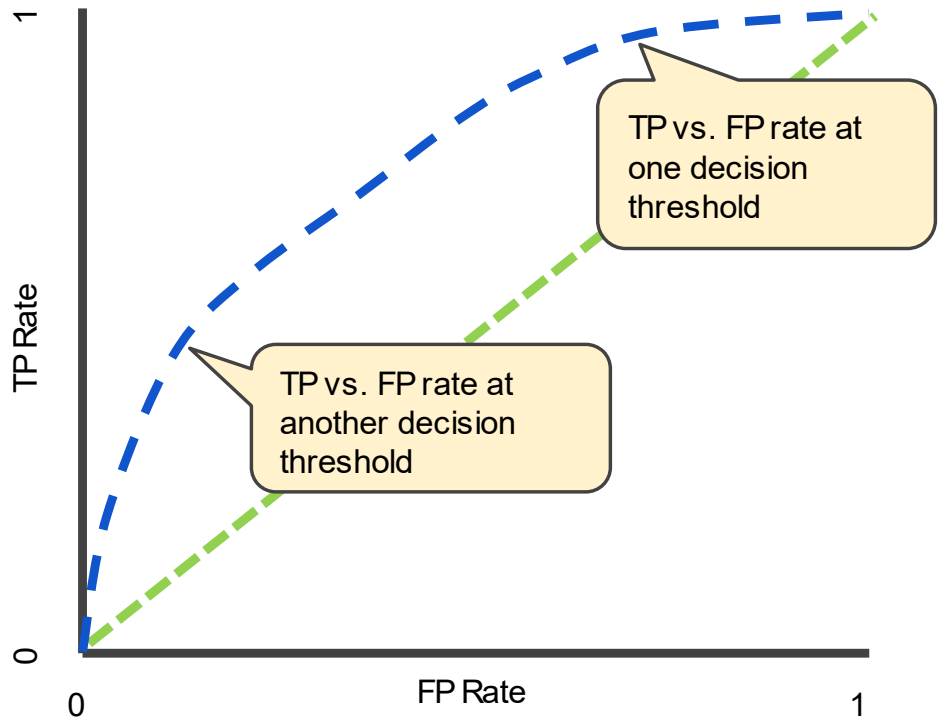
Receiver Characteristic Operation (ROC)

- Para cada **limiar** é associado um valor de **TPR** (sensibilidade) e **FPR** (1 – especificidade)
- **Limiares baixos** fazem com que uma quantidade maior de observações seja classificada como positiva, por conta disso, a taxa de falsos positivos também será alta
- **Limiares altos** fazem com que uma quantidade menor de observações seja classificada como positiva, por conta disso, a taxa de falsos positivos também será baixa



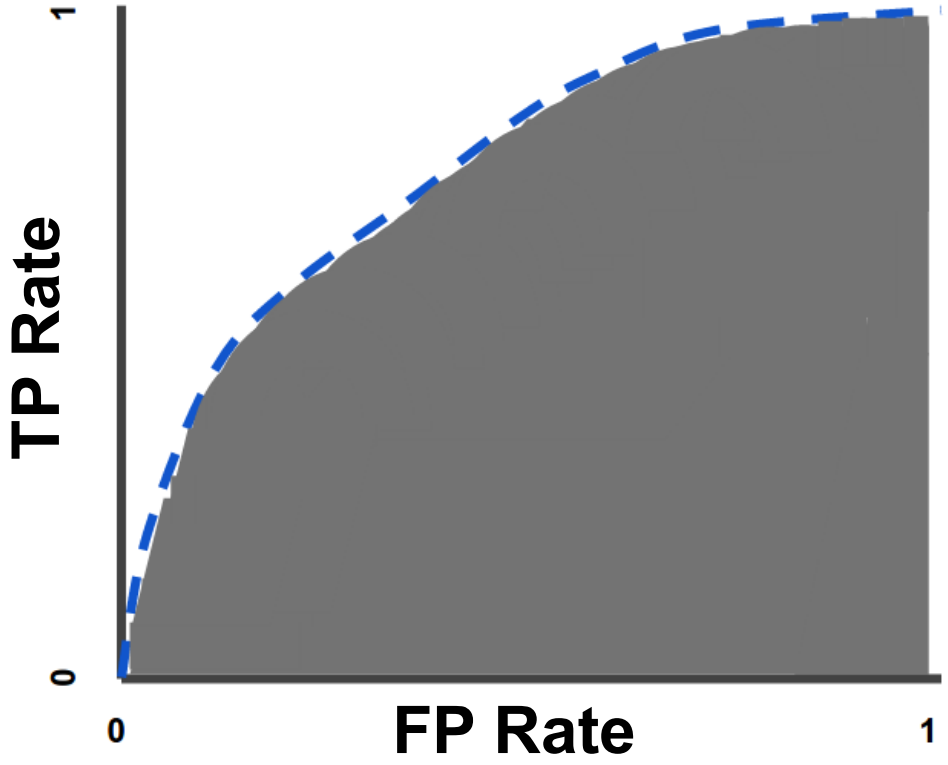
Receiver Characteristic Operation (ROC)

- A diagonal principal indica que a proporção das observações classificadas corretamente da classe **positiva** é igual a proporção das observações classificadas incorretamente da classe **negativa**, ou seja, **o resultado é aleatório**
- Dessa forma, os **melhores limiares** ficam na parte **esquerda superior do gráfico**, indicando a maior taxa de verdadeiros positivos para a menor taxa de falsos positivos



Area Under the Curve (AUC)

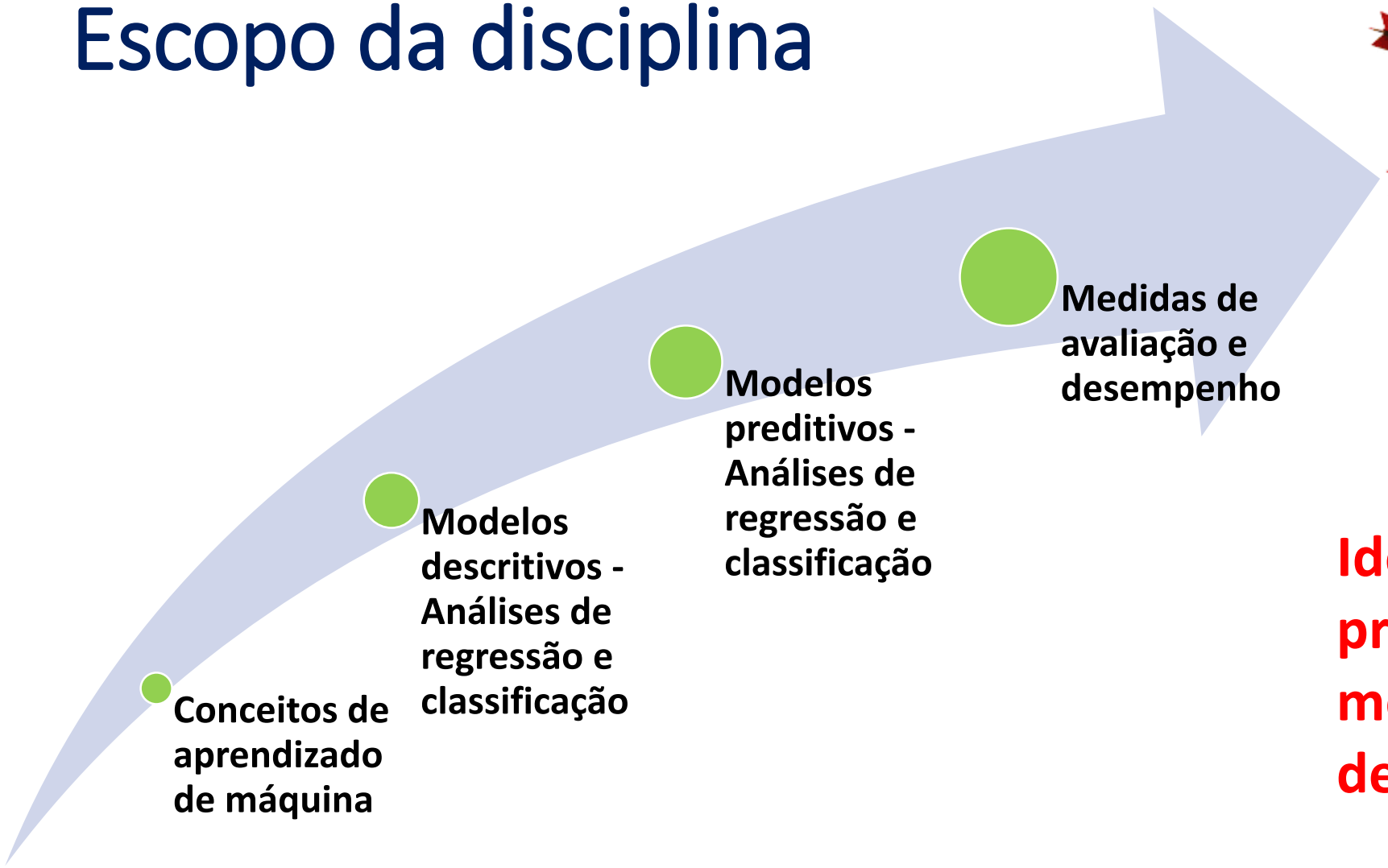
- A **AUC** nos fornece uma medida de performance agregada de **todos os possíveis limiares**, seu valor está dentro do intervalo [0,1]
- A AUC pode ser interpretada como a **capacidade do modelo distinguir as classes**
- Maior o valor da AUC, melhor é o modelo em distinguir as classes



Notas sobre outros tipos de medidas

- Além das medidas que usamos para avaliarmos os nossos modelos, também temos outros dois conjuntos de medidas: **medidas de justiça** e **indicadores de negócio**
- **Medidas de justiça** buscam quantificar as diferenças em métricas em grupos **protegidos** e **não protegidos**, definidos por alguma regra de estratificação (sem diferença, sem discriminação)
- **Indicadores de negócio** são específicos de cada área de atuação e quantificam o desempenho de um determinado processo da empresa

Escopo da disciplina



Identificar e entender problemas tratáveis por modelos de aprendizado de máquina

Estudo de caso

Using machine learning for predicting intensive care unit resource use during the COVID-19 pandemic in Denmark

The COVID-19 pandemic has put massive strains on hospitals, and tools to guide hospital planners in resource allocation during the ebbs and flows of the pandemic are urgently needed. We investigate whether machine learning (ML) can be used for predictions of intensive care requirements a fixed number of days into the future. Retrospective design where health Records from 42,526 SARS-CoV-2 positive patients in Denmark was extracted. Random Forest (RF) models were trained to predict risk of ICU admission and use of mechanical ventilation after *n* days (*n* = 1, 2, …, 15). An extended analysis was provided for *n* = 5 and *n* = 10. Models predicted *n*-day risk of ICU admission with an area under the receiver operator characteristic curve (ROC-AUC) between 0.981 and 0.995, and *n*-day risk of use of ventilation with an ROC-AUC between 0.982 and 0.997. The corresponding *n*-day forecasting models predicted the needed ICU capacity with a coefficient of determination (R^2) between 0.334 and 0.989 and use of ventilation with an R^2 between 0.446 and 0.973. The forecasting models performed worst, when forecasting many days into the future (for large *n*). For *n* = 5, ICU capacity was predicted with ROC-AUC 0.990 and R^2 0.928, and use of ventilator was predicted with ROC-AUC 0.994 and R^2 0.854. Random Forest-based modelling can be used for accurate *n*-day forecasting predictions of ICU resource requirements, when *n* is not too large.

Fonte: https://www.nature.com/articles/s41598-021-98617-1

Estudo de caso

Using machine learning for predicting intensive care unit resource use during the COVID-19 pandemic in Denmark

- 42526 pacientes que testaram positivo para SARS-CoV-2
- Dados coletados de 03/2020 até 05/2021
- Variáveis incluindo idade, imc, sexo, fumante e diagnósticos
- Variáveis temporais com as datas dos testes, admissão no hospital, uti, uso de ventilação mecânica, medicamentos, resultados de testes e sinais vitais

- Cada paciente teve uma timeline construída a partir do diagnóstico positivo da doença, até serem censurados
- A censura foi aplicada no momento da morte, um teste PCR negativo, 30 dias sem hospitalização ou a data de extração dos dados
- Para cada timeline, foram extraídos snapshots a cada 24 horas, resultando em 1246019 snapshots
- Cada snapshot é composto das variáveis do paciente e variáveis temporais até o momento do snapshot

Fonte: https://www.nature.com/articles/s41598-021-98617-1

Estudo de caso

Using machine learning for predicting intensive care unit resource use during the COVID-19 pandemic in Denmark

- Para cada snapshot, foram definidas variáveis dependentes de **n-dias**, tanto para **admissão em UTI** quanto para **uso de ventilação mecânica**
- Para uma janela de previsão de **10 dias**, seriam criadas **20 variáveis** e treinados **20 modelos diferentes**
- Dessa forma, os autores decidiram resolver a previsão de recursos do hospital como **problemas de classificação** e **agregar os resultados** da classificação para estimar a quantidade de leitos e ventiladores mecânicos necessários para os **próximos n-dias**

Fonte: https://www.nature.com/articles/s41598-021-98617-1

Estudo de caso

Using machine learning for predicting intensive care unit resource use during the COVID-19 pandemic in Denmark

- Os autores escolheram o algoritmo de **florestas aleatórias** (ensemble de árvores de decisão) para resolver os problemas de classificação
- Foram usadas **500 árvores de decisão para cada ensemble**
- Cada ensemble teve otimizado o parâmetro de **profundidade das árvores** usando a estratégia de **validação cruzada** e avaliando os modelos pela medida **AUC-ROC**
- Para as prever a **quantidade** de **pacientes internados** e com uso de **ventilação mecânica**, os pacientes foram avaliados pelos modelos de classificação para cada dia (snapshot) e tiveram a quantidade de positivos somadas para o dia

Fonte: https://www.nature.com/articles/s41598-021-98617-1

Estudo de caso

Using machine learning for predicting intensive care unit resource use during the COVID-19 pandemic in Denmark

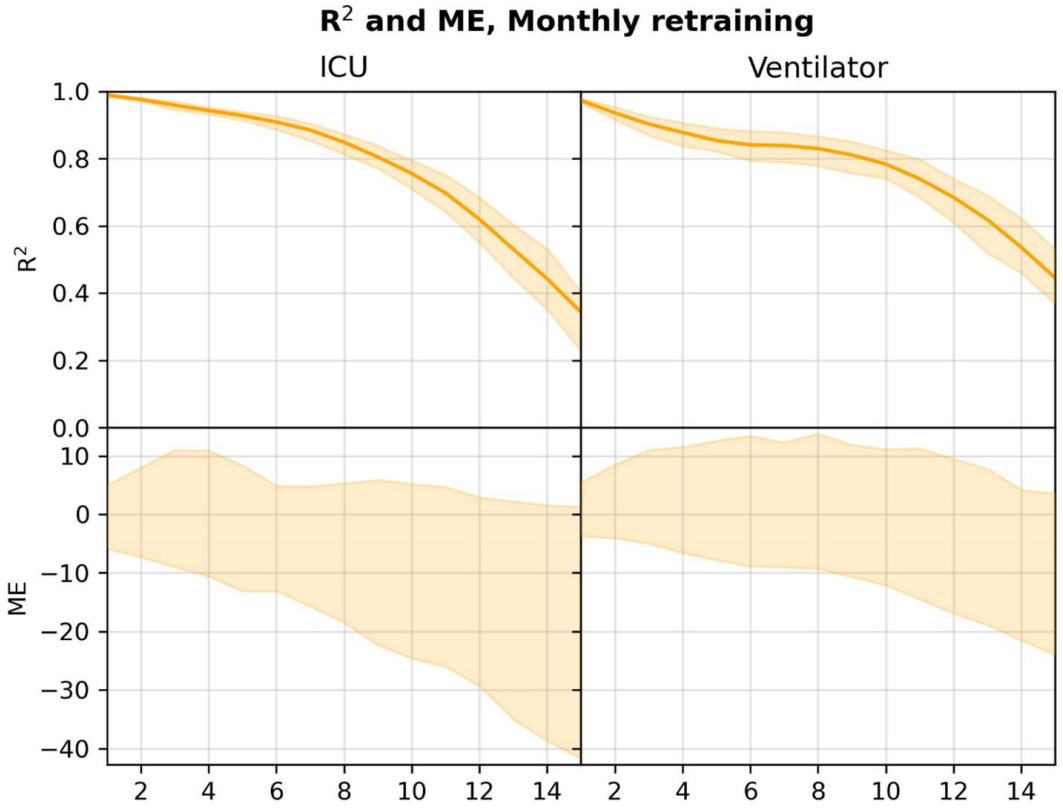
- Além de avaliar os modelos de classificação com a medida AUC-ROC, os autores avaliaram o resultados agregados dos modelos com as medidas R^2 e EM, tratando como um como um **problema de regressão**
- Os modelos gerados com essa abordagem foram comparados com **baselines de regressão, dia replicado** e **regressão logística** no lugar de **floresta aleatória**
- Os autores também calcularam os intervalos de confiança a 95% para as medidas de AUC-ROC e R^2

Fonte: https://www.nature.com/articles/s41598-021-98617-1

Estudo de caso

Using machine learning for predicting intensive care unit resource use during the COVID-19 pandemic in Denmark

- Para replicar um **cenário real**, foram avaliadas estratégias de **retreino mensal** e **primeira onda**
- Os autores avaliaram que as **florestas aleatórias** obtiveram melhores resultados com o **retreino mensal,** quando o número de dias futuros não é grande (n <= 5)
- O resultado de R^2 foi de 0,928 e 0,854 para n = 5



Fonte: https://www.nature.com/articles/s41598-021-98617-1

Estudo de caso

Using machine learning for predicting intensive care unit resource use during the COVID-19 pandemic in Denmark

O estudo demonstra que aplicar predições a nível de pacientes para uma população com o propósito de prever os recursos necessários para os próximos n-dias é possível e deve ser considerada para os eventos pandêmicos atuais e futuros

Fonte: https://www.nature.com/articles/s41598-021-98617-1

Modelos de aprendizado de máquina aplicados a dados