Soluções em Mineração de Dados

Metodologias de projetos de Data Mining

Prof. Leandro M. Almeida
lma3@cin.ufpe.br

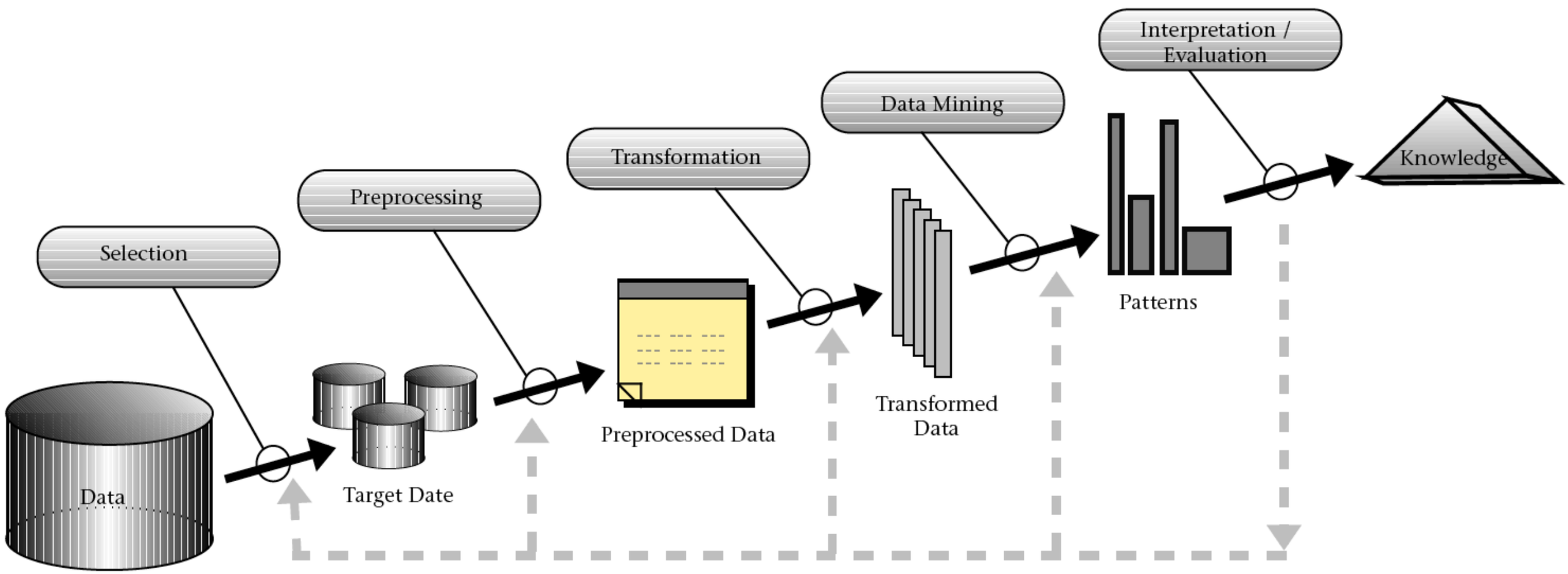
Metodologia de Projeto de DM

- Na literatura são encontradas metodologias para o desenvolvimento de projetos de mineração com o propósito de guiar os interessados

- As principais metodologias existentes são:
  - KDD (knowledge-discovery in databases)
  - CRISP-DM (Cross Industry Standard Process for Data Mining)
  - SEMMA (Sample, Explore, Modify, Model, and Assess)

KDD

- Processo estabelecido em 1989 com base para a busca por conhecimento em dados, enfatizando a aplicação em alto nível da mineração de dados



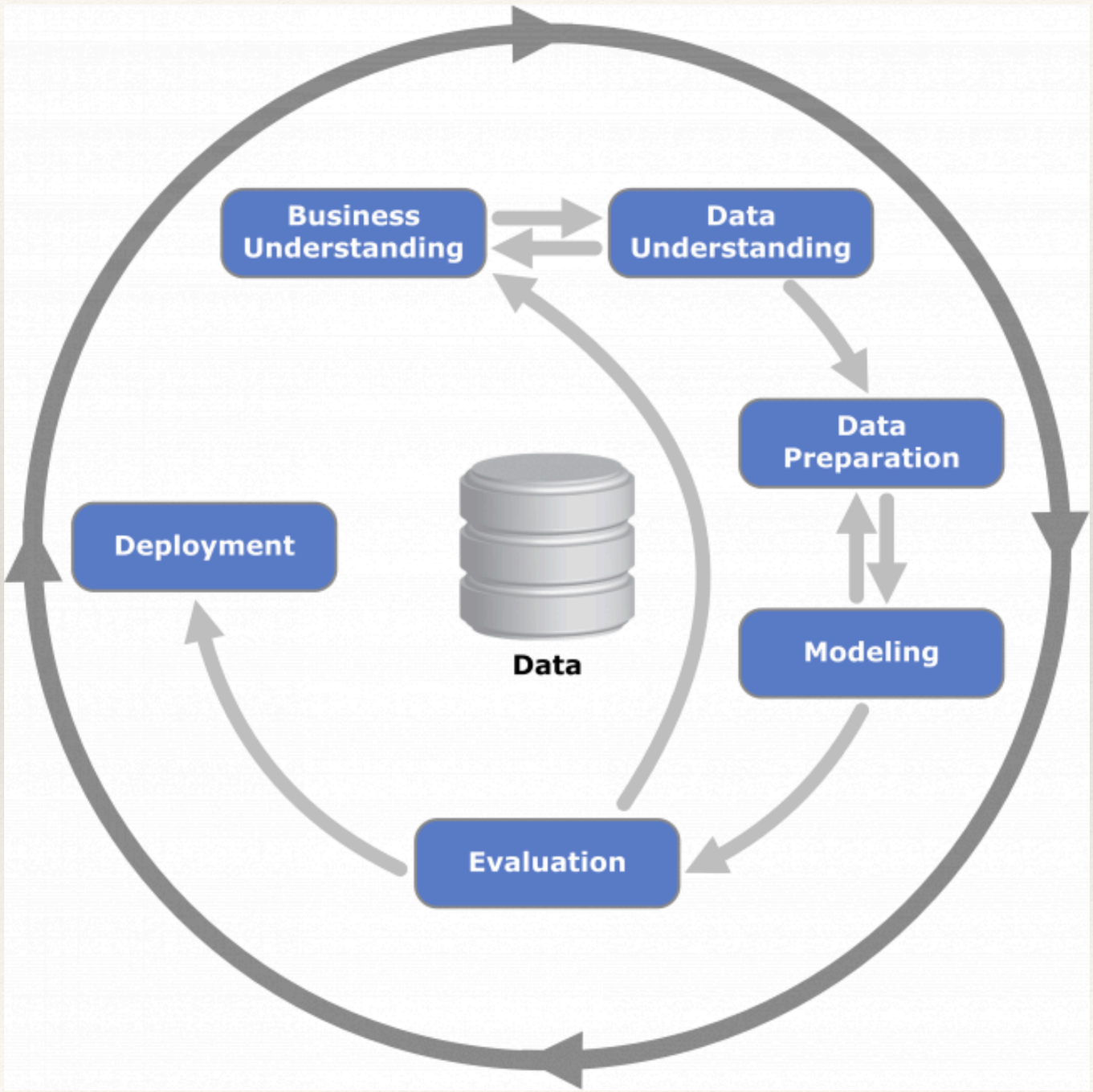
KDD

- Seleção
  - possui impacto significativo sobre a qualidade do resultado final
  - Definição do conjunto de dados contendo todas as possíveis variáveis (também chamadas de características ou atributos)
  - Normalmente essa escolha dos dados fica a critério de um especialista do domínio, ou seja, alguém que realmente entende do assunto em questão.
- Pré-processamento e Limpeza
  - Realizar tarefas que eliminem dados redundantes e inconsistentes, recuperem dados incompletos e avaliem possíveis dados discrepantes ao conjunto, chamados de outliers
- Transformação dos Dados
  - Os dados necessitam ser armazenados e formatados adequadamente para que os algoritmos possam ser aplicados (normalização, conversão de categóricos para binário, etc)
- Data Mining
- Interpretação e Avaliação

KDD

- Data Mining
  - Execução de diferentes algoritmos para a descoberta de padrões de acordo com o propósito do projeto
- Interpretação e Avaliação
  - Criar relatórios com gráficos, estatísticas e testes que corroborem o resultado obtido
  - Apresentar em linguagem não-técnica quais foram os padrões extraídos e quais as possíveis condutas a serem tomadas com o conjunto de informações/conhecimentos obtidos a partir dos dados

CRISP-DM


Business Understanding
Data Understanding
Data Preparation
Modeling
Evaluation
Deployment
Data

CRISP-DM

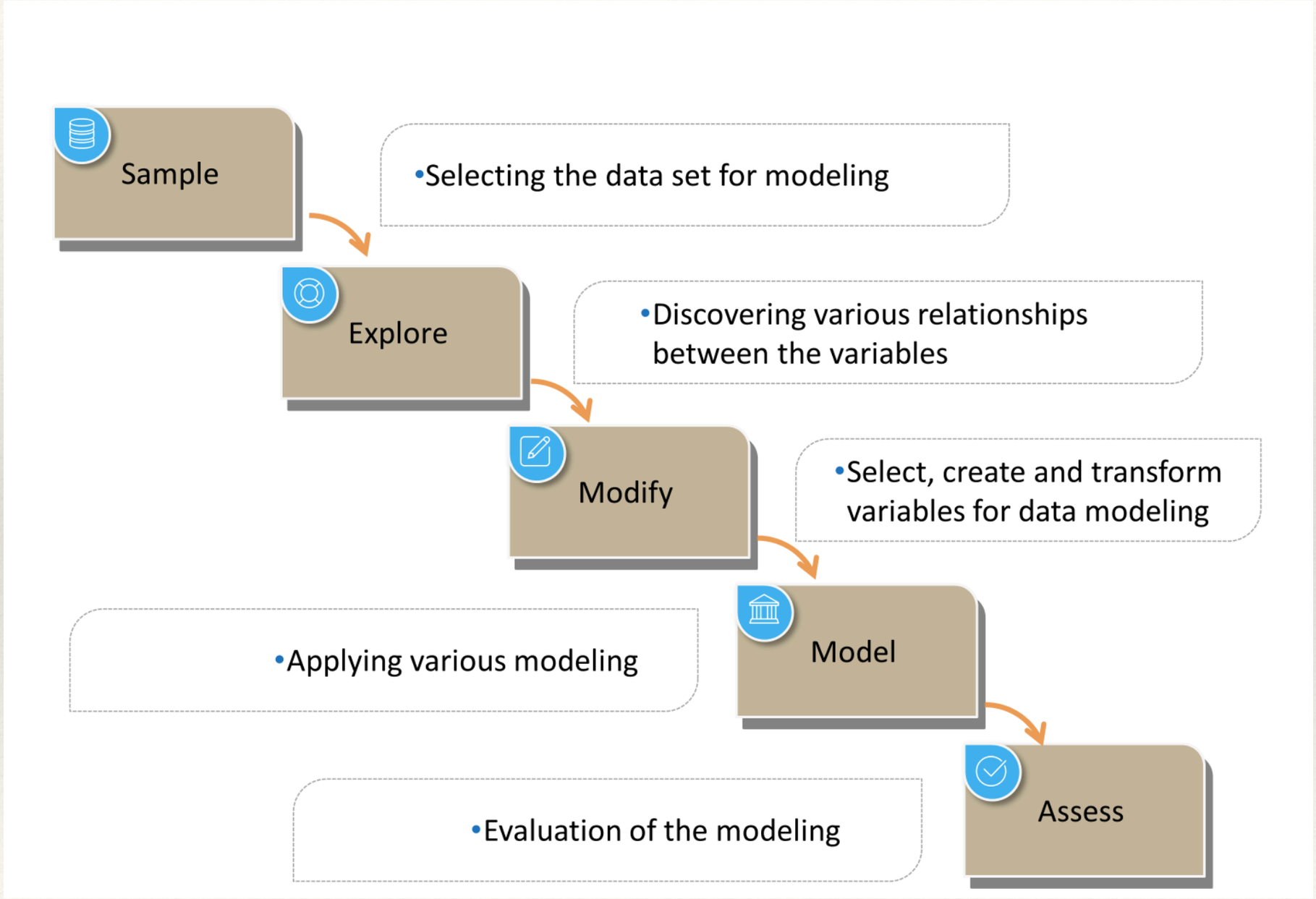
- Desenvolvida em 1996 com o objetivo de trabalhar com Big Data para descoberta de conhecimento;
- Consiste em um ciclo com 6 fases:

1. Entendimento do negócio - buscar uma compreensão adequada do problema que necessita ser resolvido
    - É preciso buscar detalhes sobre como a questão afeta a organização e quais são os principais objetivos e expectativas em relação ao trabalho como um todo.
2. Compreensão dos dados
    - Inspecionar, organizar e descrever todos os dados disponíveis
3. Preparação dos dados
    - Preparar todas as databases, definir o formato que será necessário para a análise e ajustar demais questões técnicas

CRISP-DM

3. Modelagem
    - São selecionadas e aplicadas as técnicas de mineração de dados mais apropriadas, dependendo dos objetivos identificados na primeira fase
4. Avaliação
    - Avaliação da aplicabilidade confiável dos insights e conhecimentos obtidos
5. Desenvolvimento (deploy)
    - Todo o conhecimento que for obtido por meio do trabalho de mineração e modelagem agora poderá ser aplicado de forma prática. O ideal aqui é dar uma entrega mais palpável e aplicável ao cliente a partir das análises dos dados feitas pela equipe.
    - Algumas das expectativas que se pode ter a partir deste passo é a mudança de processos da empresa ou criação de novos produtos.

SEMMA


Sample
•Selecting the data set for modeling
Explore
•Discovering various relationships between the variables
Modify
•Select, create and transform variables for data modeling
Model
•Applying various modeling
Assess
•Evaluation of the modeling

SEMMA

- Amostragem
  - Selecionar o conjunto de dados para modelagem.
- Exploração
  - Compreensão dos dados, descoberta de relações antecipadas e imprevistas entre as variáveis e também anormalidades, com a ajuda da visualização de dados.
- Modificação
  - Usa métodos para selecionar, criar e transformar variáveis na preparação para modelagem de dados.
- Modelagem
  - Aplicação de várias técnicas de extração de modelos (mineração de dados) nas variáveis preparadas
- Avaliação
  - Avaliação dos resultados da modelagem mostra a confiabilidade e utilidade dos modelos criados.

Comparação dos processos de DM

KDD	SEMMA	CRISP-DM
Pre KDD	----------	Business understanding
Selection	Sample	Data Understanding
Pre processing	Explore	
Transformation	Modify	Data preparation
Data mining	Model	Modeling
Interpretation/Evaluation	Assessment	Evaluation
Post KDD	----------	Deployment

Comparação dos processos de DM

What main methodology are you using for your analytics, data mining, or data science projects ? [200 votes total]

2014 poll 2007 poll

CRISP-DM (86)	43% 42%
My own (55)	27.5% 19%
SEMMA (17)	8.5% 13%
Other, not domain-specific (16)	8% 4%
KDD Process (15)	7.5% 7.3%
My organizations' (7)	3.5% 5.3%
A domain-specific methodology (4)	2% 4.7%
None (0)	0% 4.7%

Comparação dos processos de DM

- O SEMMA e o CRISP-DM são implementações do processo de KDD;
- CRISP-DM é mais completo que o SEMMA;
- Muitas empresas vem adotando o CRISP-DM para o desenvolvimento de soluções de mineração devido a sua completude;
- Os futuros avanços de metodologias de DM estão relacionados a linguagens baseadas em SQL e XML ainda em desenvolvimento.

Avaliação de classificadores

- A construção de classificadores de dados usa um conjunto de dados com rótulos conhecidos;
- A validação cruzada é uma técnica para avaliar a capacidade de generalização de um modelo;
- Amplamente empregada em problemas onde o objetivo da modelagem é a classificação;
  - Verificar o seu desempenho para um novo conjunto de dados

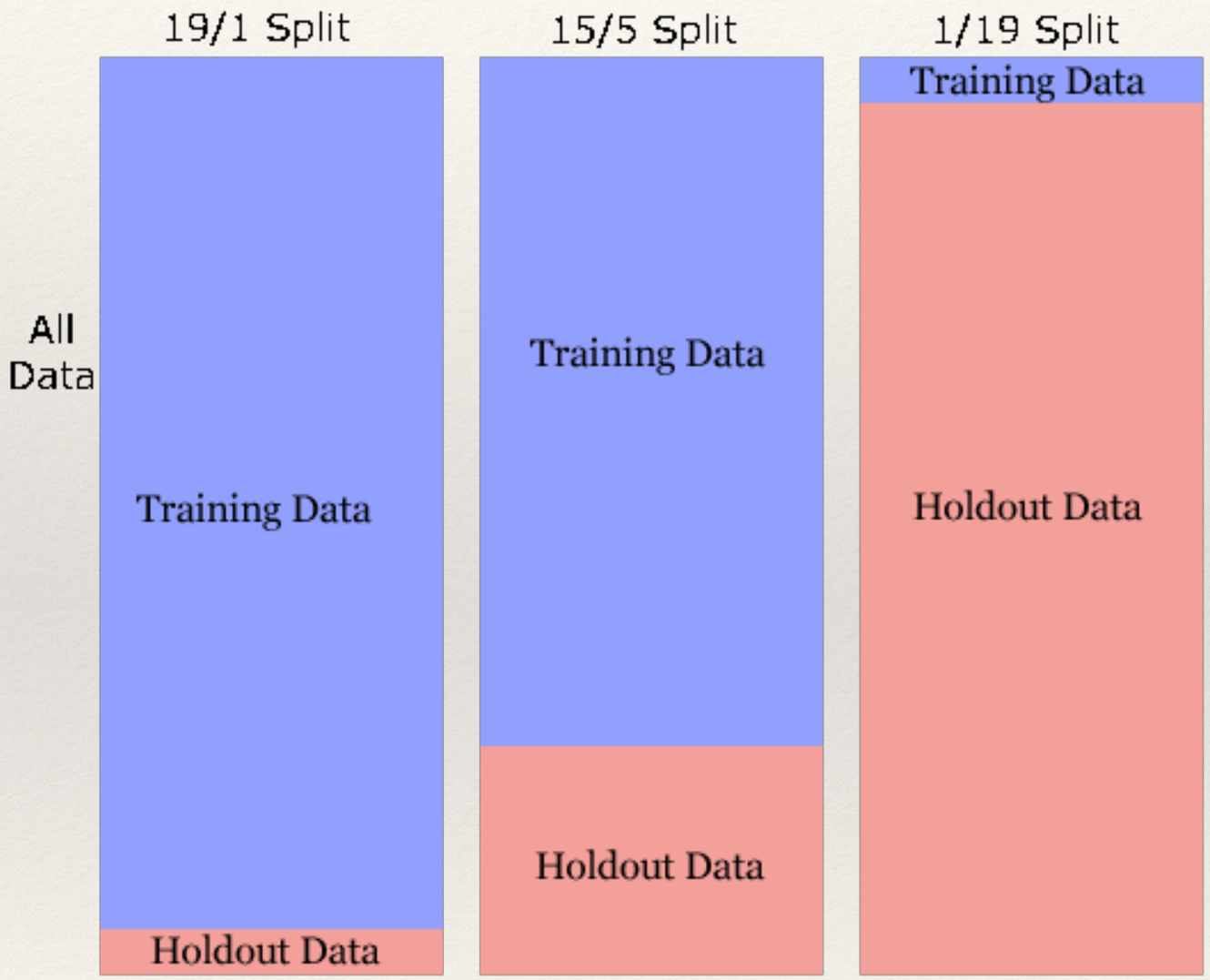
Validação cruzada

- Particiona o conjunto de dados em subconjuntos mutualmente exclusivos
- Utiliza alguns subconjuntos para a estimação dos parâmetros do modelo (dados de treinamento)
- O restante dos subconjuntos (dados de validação ou de teste) são empregados na validação do modelo.
- Diversas formas de realizar a divisão dos dados foram sugeridas, sendo as três mais utilizadas: o método *holdout*, o *k-fold* e o *leave-one-out*

Validação cruzada

- Método *holdout*:
  - Divide o conjunto total de dados em dois subconjuntos mutuamente exclusivo
  - O conjunto de dados pode ser separado em quantidades iguais ou não
  - Após o particionamento, a estimação do modelo é realizada e, posteriormente, os dados de teste são aplicados
- Esta abordagem é indicada quando está disponível uma grande quantidade de dados.
- Caso o conjunto total de dados seja pequeno, o erro calculado na predição pode sofrer muita variação

Validação cruzada



Validação cruzada

- Método *k-fold*:
  - Dividir o conjunto total de dados em k subconjuntos mutuamente exclusivos do mesmo tamanho
  - Um subconjunto é utilizado para teste e os k-1 restantes são utilizados para estimação dos parâmetros (treinamento). Ao final das k iterações calcula-se a acurácia sobre os erros encontrados
  - Ao final tem-se uma medida mais confiável sobre a capacidade do modelo de representar o processo gerador dos dados

Validação cruzada

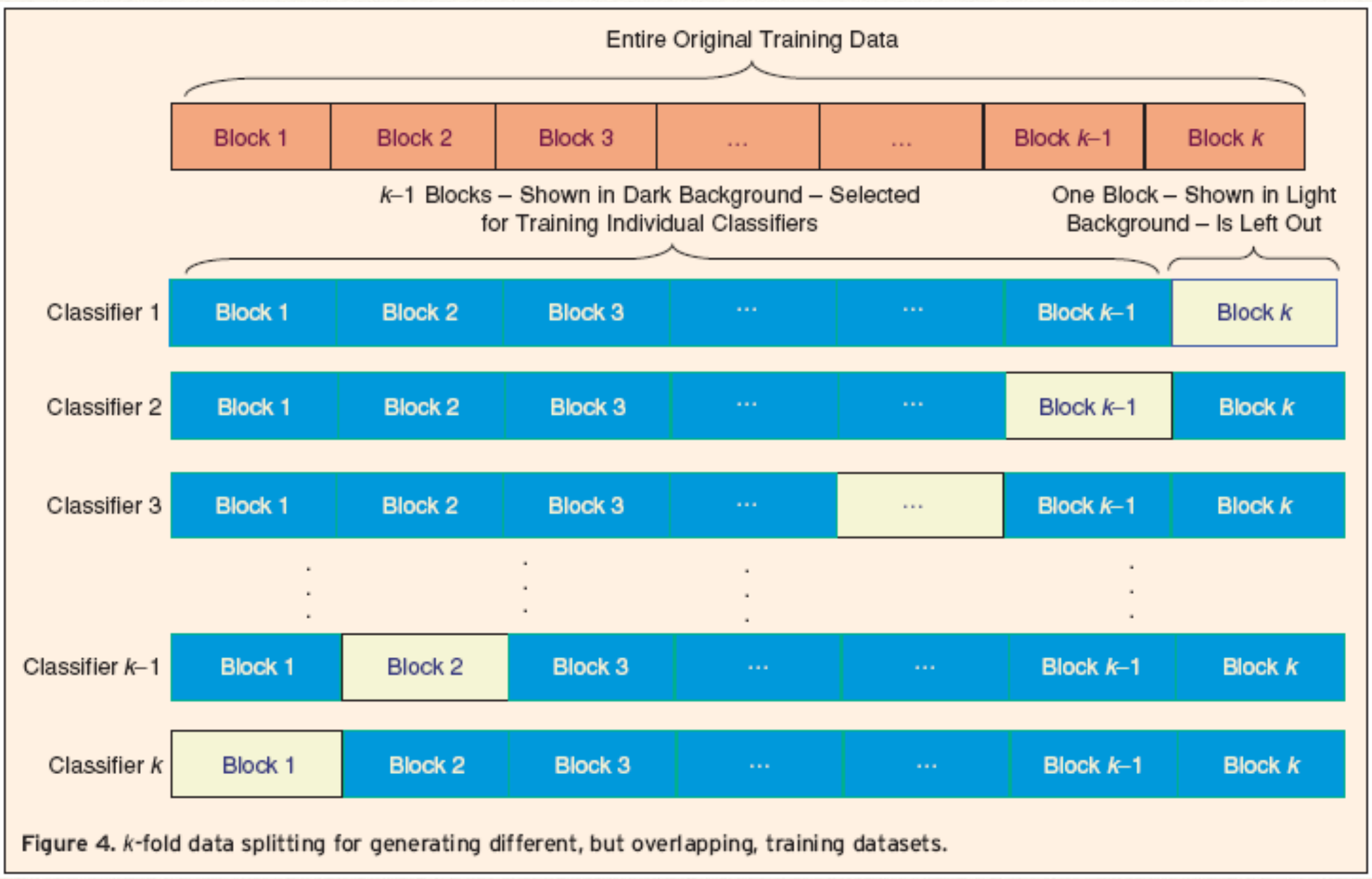
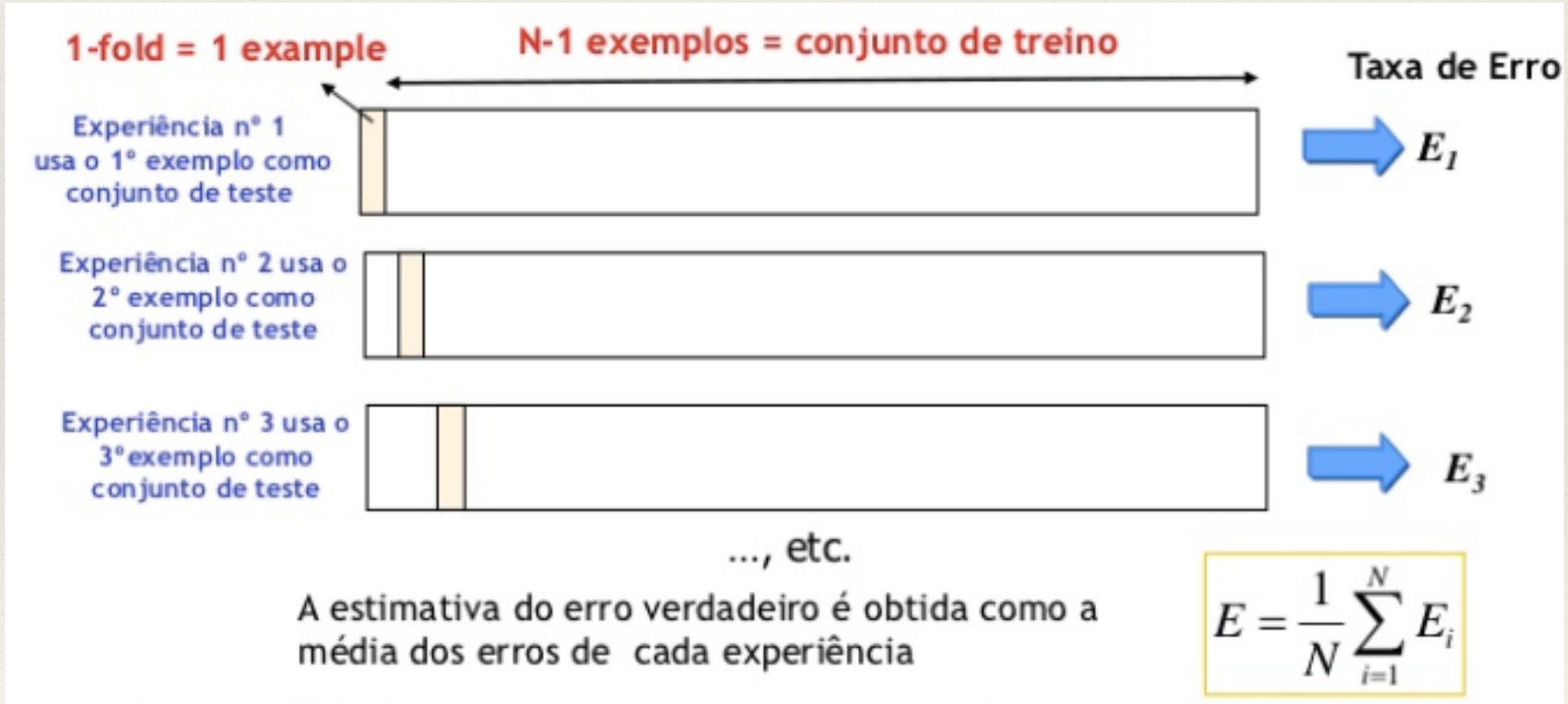


Figure 4. *k*-fold data splitting for generating different, but overlapping, training datasets.

Validação cruzada

- Método *leave one out*:
  - Caso específico do k-fold, com k igual ao número total de dados N. Nesta abordagem são realizados N cálculos de erro, um para cada dado.
  - Alto custo computacional



Avaliação sensível à distribuição das classes e ao custo

Tomada de decisão. Podemos errar?

Um classificador permite auxiliar à tomada de decisões entre diferentes ações. Podemos permitirmos tomar decisões erradas?

Tomada de decisão numa central nuclear: Um classificador h prediz se **abrir** ou **fechar** a válvula do módulo de refrigeração num dado momento

- Avaliamos desempenho num conjunto de teste = **100 000 dados** acumulados no último mês; a classe é o resultado da decisão tomada por um operário (esperto) em cada momento
  - Número de exemplos da classe "**fechar**": **99 500**
  - Número de exemplos da classe "**abrir**": **500**
- Suponhamos h prediz sempre "**fechar**" (classe majoritária). A taxa de erro é muito pequena:

$$\text{Err} = \frac{500}{100000} \times 100 = 0.5\%$$

É h um bom clasificador?

Problema de Decisão Central Nuclear

Matriz de Confusão

	CLASSE PREDITA		
		abrir	fechar
CLASSE ATUAL	abrir	TP	FN
	fechar	FP	TN

Diagonal dos acertos

Taxa de acerto (accuracy):

$$\frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{F}P + \mathrm{FN}}$$

- ✓ **TP (true positive) – positivos verdadeiros**
  nº de exemplos classificados “abrir” que são “abrir” **(correctamente classificados)**
- ✓ **FP (false positive) – positivos falsos**
  nº de exemplos classificados “abrir” que são “fechar” **(incorrectamente classificados)**
- ✓ **TN (true negative) – negativos verdadeiros**
  nº de exemplos classificados “fechar” que são “fechar” **(correctamente classificados)**
- ✓ **FN (false negative) - negativos falsos**
  nº de exemplos classificados “fechar” que são “abrir” **(incorrectamente classificados)**

Matriz de Confusão

Problema de Classificação Binária

		CLASSE PREDITA	
		Yes (+)	No (-)
CLASSE ACTUAL	Yes (+)	TP	FN
	No (-)	FP	TN

Taxa de acerto (accuracy):

$$\frac{TP + TN}{TP + TN + FP + FN}$$

- ✓ **TP (true positive) – positivos verdadeiros**
  nº de exemplos classificados positivos que são positivos **(correctamente classificados)**
- ✓ **FP (false positive) – positivos falsos**
  nº de exemplos classificados positivos que são negativos **(incorrectamente classificados)**
- ✓ **TN (true negative) – negativos verdadeiros**
  nº de exemplos classificados negativos que são negativos **(correctamente classificados)**
- ✓ **FN (false negative) - negativos falsos**
  nº de exemplos classificados negativos que são positivos **(incorrectamente classificados)**

Medidas de Avaliação

✓ **True Positive Rate = recall (sensibility):**
proporção de positivos verdadeiros do total de positivos

$$TPR = \frac{TP}{TP + FN}$$

actual \ predita	+	-
+	TP	FN
-	FP	TN

✓ **False Positive Rate:** proporção positivos falsos (incorrectamente classificados como positivos) do total de negativos

$$FPR = \frac{FP}{FP + TN}$$

actual \ predita	+	-
+	TP	FN
-	FP	TN

✓ **True Negative Rate:**
proporção de negativos verdadeiros do total de negativos

$$TNR = \frac{TN}{FP + TN}$$

actual \ predita	+	-
+	TP	FN
-	FP	TN

✓ **False Negative Rate:** proporção de negativos falsos (incorrectamente classificados como negativos) do total de positivos

$$FNR = \frac{FN}{TP + FN}$$

actual \ predita	+	-
+	TP	FN
-	FP	TN

Precisão e Sensibilidade

- ✓ **Precision (precisão):** proporção de positivos verdadeiros do total dos exemplos classificados como positivos

$$\text{precision} = \frac{TP}{TP + FP}$$

actual \ predita	+	-
+	TP	FN
-	FP	TN

- ✓ **Recall (sensibilidade)** (true positive rate): proporção de exemplos positivos que foram correctamente classificados

$$\text{recall} = TPR = \frac{TP}{TP + FN}$$

actual \ predita	+	-
+	TP	FN
-	FP	TN

$$F - \text{measure} = \frac{2}{1/\text{precision} + 1/\text{recall}}$$

Precisão e Sensibilidade

Duas medidas de desempenho muito usadas nos sistemas de recuperação de informação (information retrieval systems). Os documentos de uma base de dados podem ser **recuperados** (classificados como relevantes) ou **rejeitados** a partir de uma **"query"** à base de dados realizado por um utilizador

✓ **Precision (precisão):** mede a proporção dos documentos recuperados que são realmente relevantes do total de documentos recuperados.

$$\text{precisão} = \frac{\text{documentos relevantes recuperados}}{\text{documentos recuperados}}$$

✓ **Recall (sensibilidade)** (true positive rate): reflete a probabilidade de que um documento realmente relevante seja recuperado pelo sistema

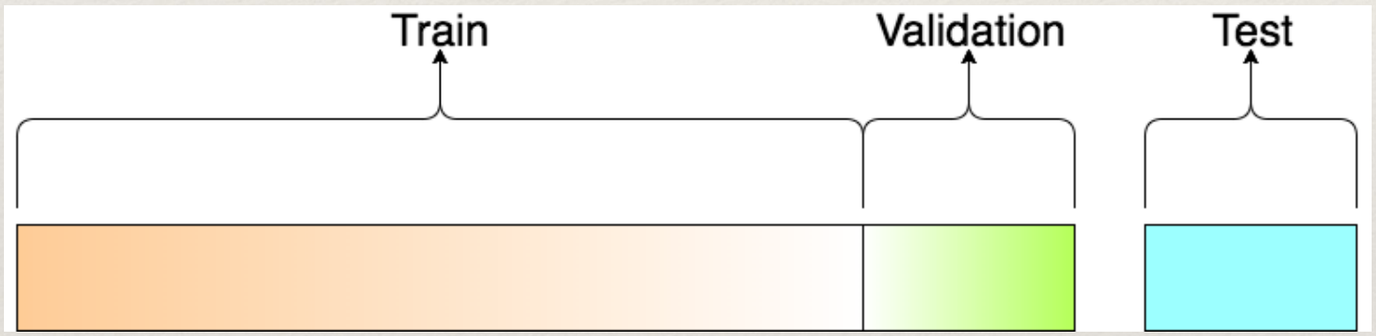
$$\text{sensibilidade} = \frac{\text{documentos relevantes recuperados}}{\text{documentos relevantes}}$$

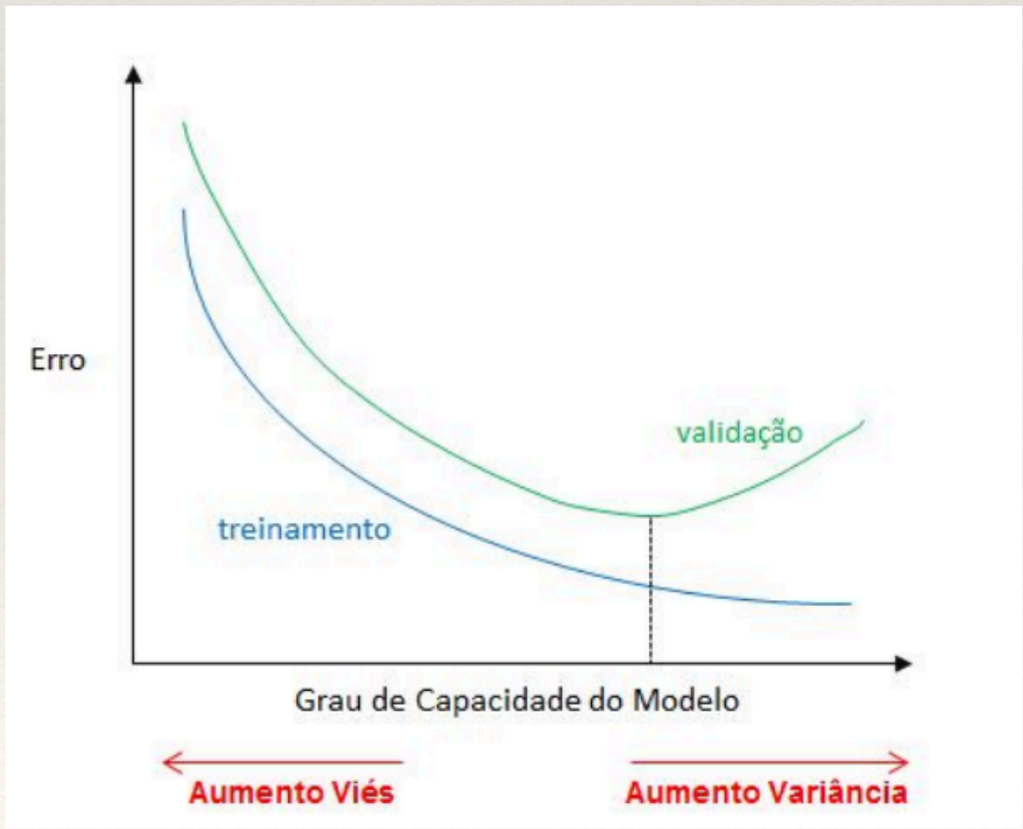
Comparação de Classificadores

- Como realizar corretamente a comparação de classificadores?
  - Realizar o pré-processamento dos dados
  - Utilizar a mesma semente para geração de números aleatórios
  - Separar os conjuntos de Treinamento, Validação e Teste
  - Utilizar a validação cruzada com k-10 usando o Treinamento
- Calcular a significância estatística com base nos desempenhos dos classificadores

Conjunto de Validação

- Utilizado para avaliar o nível de generalização do modelo
- Importante para definir um critério de parada da busca
- Não guiar a busca apenas pelo desempenho na validação





Comparação de Classificadores

- Supondo que os dados não são gerados a partir de uma normal
  - Utilizar um teste não-paramétrico para verificar a significância estatística
    - Probabilidade de que os dados não são de uma mesma distribuição
    - Necessário trabalhar com dados do tipo real
- Se os dados seguem uma normal, usar o teste *t-Student*
  - Realizar o teste Shapiro-Wilk

Comparação de Classificadores

- O teste de Shapiro-Wilk é utilizado para verificar a normalidade dos dados, antes de se definir o teste de hipótese que deve ser utilizado para verificar diferença entre amostras

1. Elaboração das hipóteses:

   H_0 : *a amostra provém de uma população normal*

   H_1 : *a amostra não provém de uma população normal*

2. Definir o nível de significância do teste:

   $\alpha = 0.05$, ou seja, o teste terá um nível de confiança de 95%.

3. Cálculo do $p - value$ e tomada de decisão:

   Rejeita H_0 se $p\text{-}value < \alpha$;

   Não é possível rejeitar H_0 se $p\text{-}value > \alpha$.

```
>>> from scipy import stats
>>> np.random.seed(12345678)
>>> x = stats.norm.rvs(loc=5, scale=3, size=100)
>>> stats.shapiro(x)
(0.9772805571556091, 0.08144091814756393)
```

```
>>> a = [2,3,4,3,5,6] # amostra a
>>> b = [5,7,8,8,5,6] # amostra b
>>> stats.ttest_ind(a,b)
(array(-3.2391053207156637), 0.008882866991368814
```

Comparação de Classificadores

- Estatística não paramétrica: abrange técnicas que não dependem de dados pertencentes a nenhuma distribuição particular

- Testes de Significância Estatística não paramétricos
  - Como verificar se dois conjuntos de dados são diferentes?
  - Defini-se a hipótese nula H0 = mesma distribuição
    - Mesma média, mediana. i.e. parâmetros.

Testes de Significância Estatística

- Com o cálculo do teste de significância é possível:
  - Rejeitar H0.
    - Os dados tem origem de diferentes distribuições
  - Não Rejeitar H0
    - Os dados não possuem origens diferentes
- Testes utilizados com dados de acurácia de classificadores para determinar as diferenças ou não

Testes de Significância Estatística

- Usualmente os testes requerem um conhecimento estatístico para a uma melhor compreensão/uso/interpretação
- Geralmente os testes retornam ***p-value*** que ajuda a interpretar o resultado. Para ***alpha = 0.05***:
  - **p <= alpha**: reject H0, different distribution.
  - **p > alpha**: fail to reject H0, same distribution.

Testes de Significância Estatística

- Mann-Whitney U Test
  - Também conhecido como Wilcoxon-Mann-Whitney
  - Requer 20 observações de cada conjunto de dados
  - Teste para amostras pareadas
  - Retorna o cálculo da estatística e o p-value

Testes de Significância Estatística

- Wilcoxon Signed-Rank Test
  - Equivalente ao Student t-test
  - Requer 20 observações de cada conjunto de dados
  - Teste para amostras pareadas
  - Retorna o cálculo da estatística e o p-value

Testes de Significância Estatística

- Kruskal-Wallis H Test
  - Teste para amostras não pareadas
  - Requer 5 observações de cada conjunto de dados
  - Retorna o cálculo da estatística e o p-value
  - Eficiente para amostras com muitos dados

Testes de Significância Estatística

- Friedman Test
- Apropriado para amostras pareadas relativas a repetidas medidas
- Requer 10 observações de cada conjunto de dados
- Retorna o cálculo da estatística e o p-value

Estrutura do código

1. Bibliotecas
    1. Necessárias para a utilização dos diferentes métodos, modelos, técnicas.
2. Semente para números aleatórios
    1. Importante para comparações da evolução da busca por parâmetros
    2. Toda função que utilizar geração de número aleatórios possuirá o random_state (checar a documentação)
3. Leitura dos dados
    1. Utilizar o Pandas para a manipulação de dados
4. Análise exploratória dos dados
    1. Muitas funções são encontradas nas bibliotecas Pandas e Seaborn
5. Preparação dos dados
    1. As principais funções são oriundas do Pandas, porém existem outras bibliotecas disponíveis
    2. Verificar antes a necessidade de dividir o conjunto em treinamento e teste

Estrutura do código

6. Divisão da base de dados em treino e teste
    1. Realizar a divisão da base em treino e teste
    2. Utilizar apenas o conjunto de dados de treino para buscar parâmetros do modelo
    3. SEMPRE deixar o conjunto de teste fora do processo de busca
    4. Verificar a necessidade de definir random_state
7. Busca por parâmetros de modelos com base no conjunto de treinamento
    1. Dividir o conjunto de treino para ter o conjunto de validação
    2. Executar diferentes variações de parâmetros
    3. Verificar a necessidade de definir random_state
8. Definição dos modelos de classificação
    1. Criar instâncias dos modelos de classificação
    2. Verificar a necessidade de definir random_state para cada modelo

Estrutura do código

9. Definição do modelo experimental
    1. Utilização a priori do modelo k-fold visando a comparação entre o modelos de classificação
    2. Verificar a necessidade de definir random_state
10. Execução do modelo experimental
    1. De posse das melhores configurações de modelos, executar o k-fold
    2. Verificar a necessidade de definir random_state
    3. Coletar o vetor com os desempenhos de cada modelo nos folds
11. Comparação de modelos usando teste de significância
    1. Aplicar testes de significância estatística usando o vetor de desempenhos no folds
    2. Verificar a necessidade de definir random_state
12. Apresentação de resultados
    1. Utilizar gráficos e relatórios para apresentar os resultados e apontar o método mais apropriado para a base de dados

Hiperparâmetros

- Hiperparâmetros: são os dados que controlam o próprio processo de treinamento
- Categorias de dados durante a construção de modelos:
  - **Dados de entrada** ou dados de treinamento
  - **Parâmetros do modelo** para a tarefa de mineração que são ajustados durante o treinamento
  - Os **hiperparâmetros** controlam o processo de treinamento e não mudam durante o mesmo

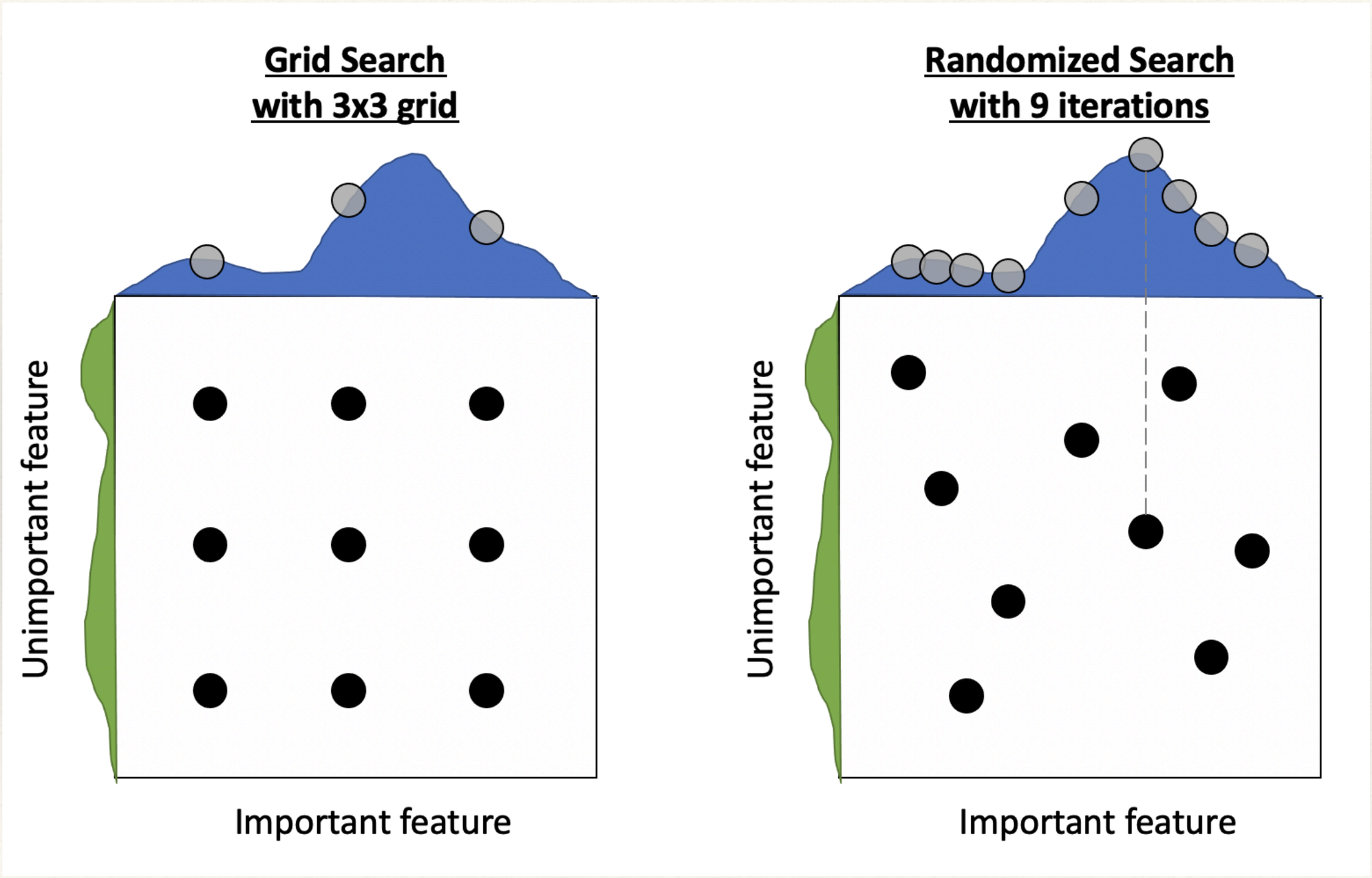
Ajuste de hiperparâmetro

- Normalmente realizado com base em testes de várias configurações
- Valores para os testes definidos dentro de limites previamente especificados
- Necessário definição de uma métrica relacionado o domínio do problema
- Existem diferentes técnicas para ajuste/otimização de hiperparâmetros

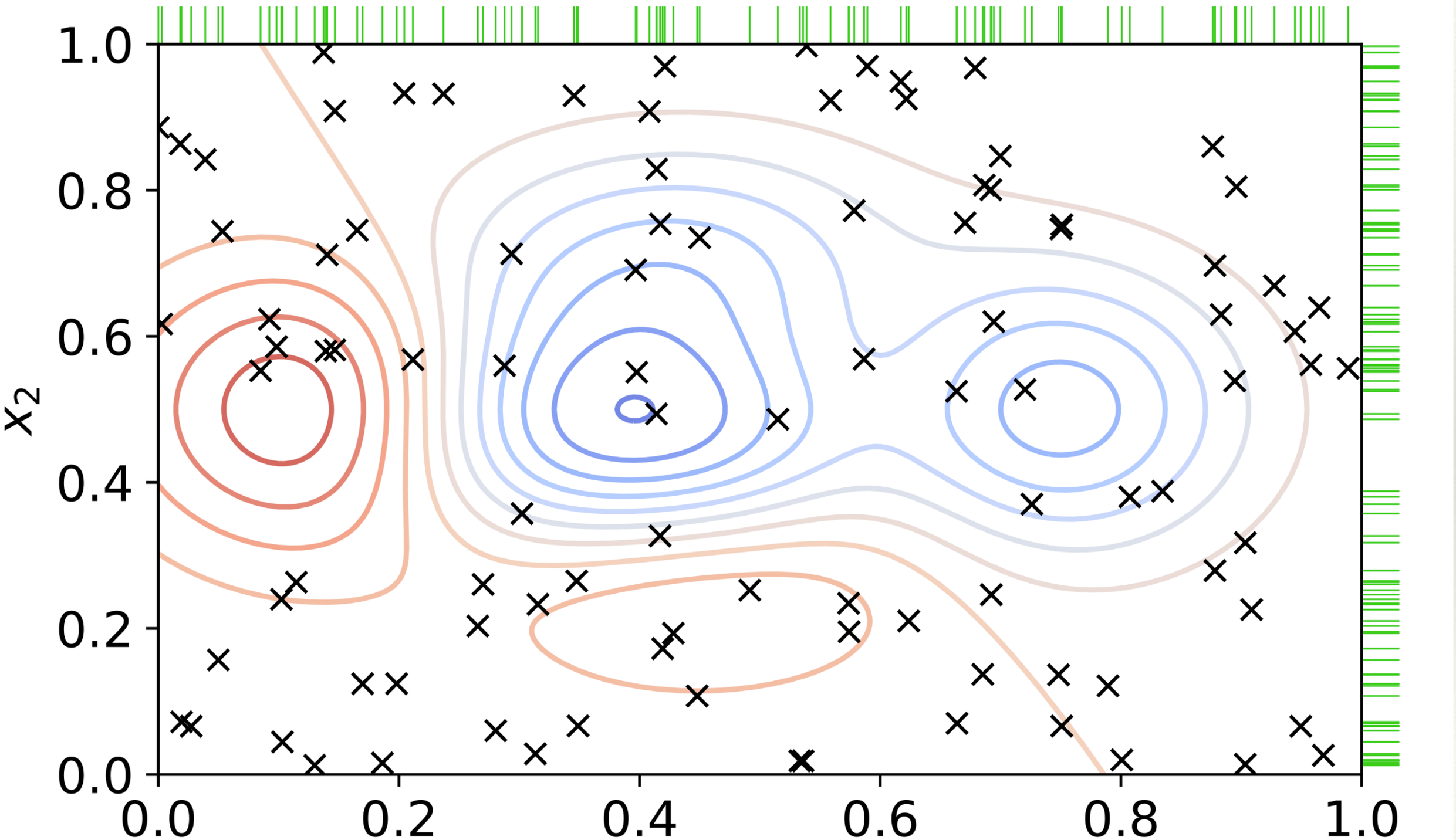
Métodos de busca

- Os métodos de busca por hiperparâmetros exigem a definição de limites para o espaço de busca
- Muitas vezes tratado com problema de otimização
- Principais métodos:
  - **Busca aleatória**: espaço de busca definido com base nos hiperparâmetros e pontos aleatórios
  - **Busca em grade**: espaço de busca em grade para teste de cada possibilidade.

Métodos de busca



Random Search



Atividade

- Considerando base de dados do Bank:
  - Ajustar o script para realizar busca por hiperparâmetros para todos os modelos
  - Realizar a comparação de significância estatística
  - Apresentar os resultados e discussões.