Validação cruzada

Ciência de Dados II

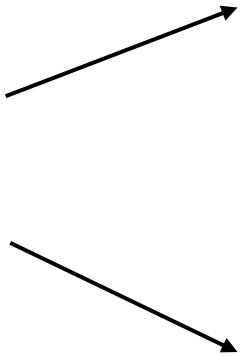
Professor: Gabriel Machado Lunardi
gabriel.lunardi@ufsm.br

Na aula passada... Avaliação holdout

Vimos como dividir um dataset em treino e teste para avaliar a performance de um modelo de aprendizado de máquina.

Conjunto de dados original

Dono(a) da casa	Status de relacionamento	Renda anual	Inadimplente
Sim	Solteiro(a)	125K	Não
Não	Casado(a)	100K	Sim
Sim	Divociado(a)	70K	Não
Não	Solteiro(a)	120K	Sim
Não	Casado(a)	95K	Não



Conjunto de treinamento (80%)

Dono(a) da casa	Status de relacionamento	Renda anual	Inadimplente
Sim	Solteiro(a)	125K	Não
Não	Casado(a)	100K	Sim
Sim	Divociado(a)	70K	Não
Não	Solteiro(a)	120K	Sim

Conjunto de testes (20%)

Dono(a) da casa	Status de relacionamento	Renda anual	Inadimplente
Não	Casado(a)	95K	?

Problemas da abordagem holdout

- **Alta variância**
  - O resultado muda muito dependendo de qual linha vai para o teste.
- **Subutilização de dados**
  - 20% dos dados (uma linha) não contribuem para o treino.
- **Datasets pequenos**
  - Com cinco linhas, perder uma linha para teste é significativo.

Validação cruzada (*Cross-validation – CV*)

Mitiga os problemas do holdout, usando todos os dados para treino e validação, mas de forma organizada.

1. *k-Fold* CV:
   - ✓Divide os dados em k partes (folds).
   - ✓Cada parte serve como **teste uma vez** e como **treino nas outras vezes**.

2. Métrica Final: Média do desempenho em todos os folds.
   - ✓Precisão, revocação, medida-F, acurácia (média dos folds)
   - ✓RMSE, MAE, R^2 (média dos folds)

Validação cruzada (*Cross-validation – CV*)

Conjunto de dados original

Dono(a) da casa	Status de relacionamento	Renda anual	Inadimplente
Sim	Solteiro(a)	125K	Não
Não	Casado(a)	100K	Sim
Sim	Divociado(a)	70K	Não
Não	Solteiro(a)	120K	Sim
Não	Casado(a)	95K	Não
Não	Casado(a)	55K	Sim
Sim	Casado(a)	58K	Sim
Sim	Divorciado(a)	120K	Não
Não	Solteiro(a)	81K	Sim

Validação cruzada (*Cross-validation – CV*)

Conjunto de dados original

Dono(a) da casa	Status de relacionamento	Renda anual	Inadimplente
Sim	Solteiro(a)	125K	Não
Não	Casado(a)	100K	Sim
Sim	Divociado(a)	70K	Não
Não	Solteiro(a)	120K	Sim
Não	Casado(a)	95K	Não
Não	Casado(a)	55K	Sim
Sim	Casado(a)	58K	Sim
Sim	Divorciado(a)	120K	Não
Não	Solteiro(a)	81K	Sim

K=3

Validação cruzada (*Cross-validation – CV*)

Conjunto de dados original

Dono(a) da casa	Status de relacionamento	Renda anual	Inadimplente	
Sim	Solteiro(a)	125K	Não	Teste
Não	Casado(a)	100K	Sim	
Sim	Divociado(a)	70K	Não	
Não	Solteiro(a)	120K	Sim	Treino
Não	Casado(a)	95K	Não	
Não	Casado(a)	55K	Sim	
Sim	Casado(a)	58K	Sim	Treino
Sim	Divorciado(a)	120K	Não	
Não	Solteiro(a)	81K	Sim	

K=3

Iteração 1

Validação cruzada (*Cross-validation – CV*)

Conjunto de dados original

Dono(a) da casa	Status de relacionamento	Renda anual	Inadimplente	
Sim	Solteiro(a)	125K	Não	Treino
Não	Casado(a)	100K	Sim	
Sim	Divociado(a)	70K	Não	
Não	Solteiro(a)	120K	Sim	Teste
Não	Casado(a)	95K	Não	
Não	Casado(a)	55K	Sim	
Sim	Casado(a)	58K	Sim	Treino
Sim	Divorciado(a)	120K	Não	
Não	Solteiro(a)	81K	Sim	

K=3

Iteração 2

Validação cruzada (*Cross-validation – CV*)

Conjunto de dados original

Dono(a) da casa	Status de relacionamento	Renda anual	Inadimplente	
Sim	Solteiro(a)	125K	Não	Treino
Não	Casado(a)	100K	Sim	
Sim	Divociado(a)	70K	Não	
Não	Solteiro(a)	120K	Sim	Treino
Não	Casado(a)	95K	Não	
Não	Casado(a)	55K	Sim	
Sim	Casado(a)	58K	Sim	Teste
Sim	Divorciado(a)	120K	Não	
Não	Solteiro(a)	81K	Sim	

K=3

Iteração 3

Cuidado com o desbalanço de classes

E se nosso dataset contivesse um desbalanço de instâncias no atributo alvo?

- ✓Statified k-Fold
- ✓Garante a proporção de instâncias em cada fold

Cuidado com o desbalanço de classes

E se nosso dataset contivesse um desbalanço de instâncias no atributo alvo?

- ✓Statified k-Fold
- ✓Garante a proporção de instâncias em cada fold

Outros exemplos onde pode-se ter desbalanceamento

- ✓Fraudes
- ✓Doenças raras

Atenção, não se aplica a problemas de regressão!

CV é sempre melhor?

✅ Mais confiável para datasets pequenos/médios.

✅ Reduz o viés de avaliação (não depende de uma única divisão).

❌ **Custo Computacional**

Inviável para datasets muito grandes (ex.: 1M linhas).

❌ **Temporalidade**

Dados temporais exigem validação específica (ex.: Time Series Split).

```
Use validação cruzada quando puder pagar o custo
computacional e não houver dependência temporal nos dados.
```

Exemplo de dados temporais

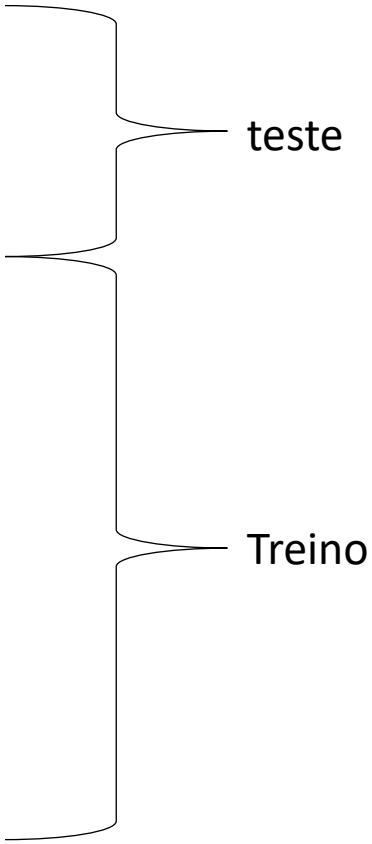
Contexto: Preços históricos da ação PETR4 (Petrobras) ao longo de 10 dias.

Data	Preço(R$)	Volume (ações)
2023-01-01	28.50	1.000.000
2023-01-02	29.10	1.200.000
2023-01-03	28.75	950.000
2023-01-04	29.30	1.100.000
2023-01-05	30.05	1.500.000
2023-01-06	29.90	1.300.000
2023-01-07	30.20	1.400.000
2023-01-08	30.50	1.600.000
2023-01-09	30.25	1.450.000
2023-01-10	30.80	1.700.000

Exemplo de dados temporais

Contexto: Preços históricos da ação PETR4 (Petrobras) ao longo de 10 dias.

Data	Preço(R$)	Volume (ações)
2023-01-01	28.50	1.000.000
2023-01-02	29.10	1.200.000
2023-01-03	28.75	950.000
2023-01-04	29.30	1.100.000
2023-01-05	30.05	1.500.000
2023-01-06	29.90	1.300.000
2023-01-07	30.20	1.400.000
2023-01-08	30.50	1.600.000
2023-01-09	30.25	1.450.000
2023-01-10	30.80	1.700.000

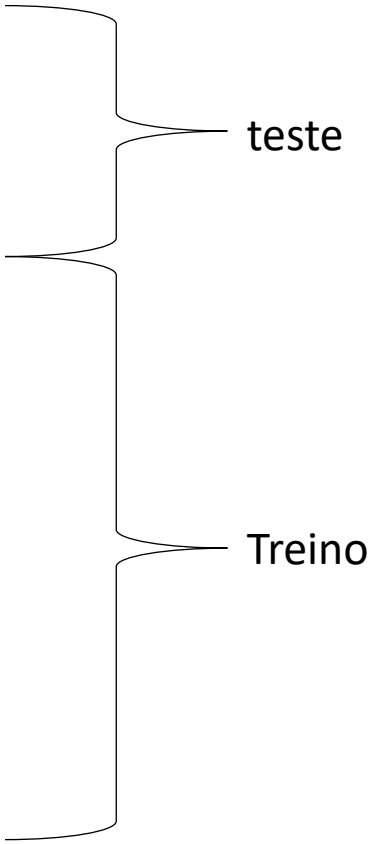


Suponha a seguinte divisão Holdout aleatória.

Exemplo de dados temporais

Contexto: Preços históricos da ação PETR4 (Petrobras) ao longo de 10 dias.

Data	Preço(R$)	Volume (ações)
2023-01-01	28.50	1.000.000
2023-01-02	29.10	1.200.000
2023-01-03	28.75	950.000
2023-01-04	29.30	1.100.000
2023-01-05	30.05	1.500.000
2023-01-06	29.90	1.300.000
2023-01-07	30.20	1.400.000
2023-01-08	30.50	1.600.000
2023-01-09	30.25	1.450.000
2023-01-10	30.80	1.700.000

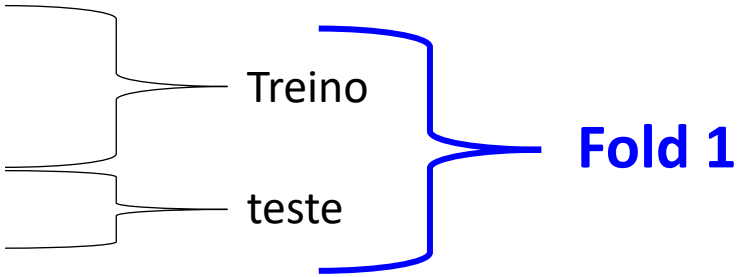


O preço de hoje depende do preço de ontem. Se mexermos na ordem, perdemos essa relação temporal.

Exemplo de dados temporais

Contexto: Preços históricos da ação PETR4 (Petrobras) ao longo de 10 dias.

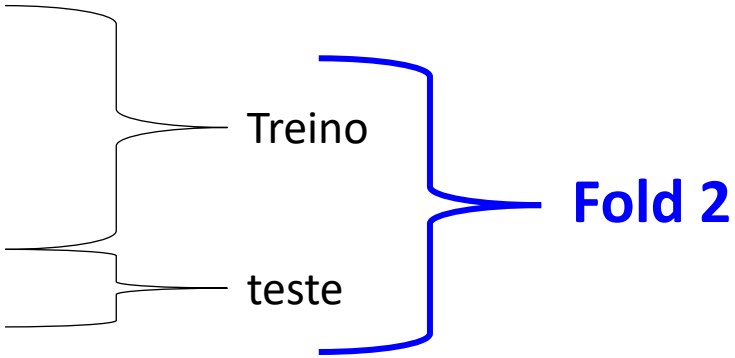
Data	Preço(R$)	Volume (ações)
2023-01-01	28.50	1.000.000
2023-01-02	29.10	1.200.000
2023-01-03	28.75	950.000
2023-01-04	29.30	1.100.000
2023-01-05	30.05	1.500.000
2023-01-06	29.90	1.300.000
2023-01-07	30.20	1.400.000
2023-01-08	30.50	1.600.000
2023-01-09	30.25	1.450.000
2023-01-10	30.80	1.700.000



Exemplo de dados temporais

Contexto: Preços históricos da ação PETR4 (Petrobras) ao longo de 10 dias.

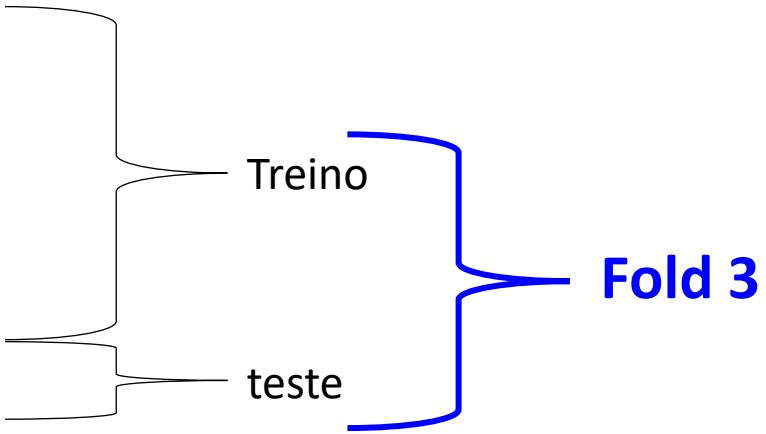
Data	Preço(R$)	Volume (ações)
2023-01-01	28.50	1.000.000
2023-01-02	29.10	1.200.000
2023-01-03	28.75	950.000
2023-01-04	29.30	1.100.000
2023-01-05	30.05	1.500.000
2023-01-06	29.90	1.300.000
2023-01-07	30.20	1.400.000
2023-01-08	30.50	1.600.000
2023-01-09	30.25	1.450.000
2023-01-10	30.80	1.700.000



Exemplo de dados temporais

Contexto: Preços históricos da ação PETR4 (Petrobras) ao longo de 10 dias.

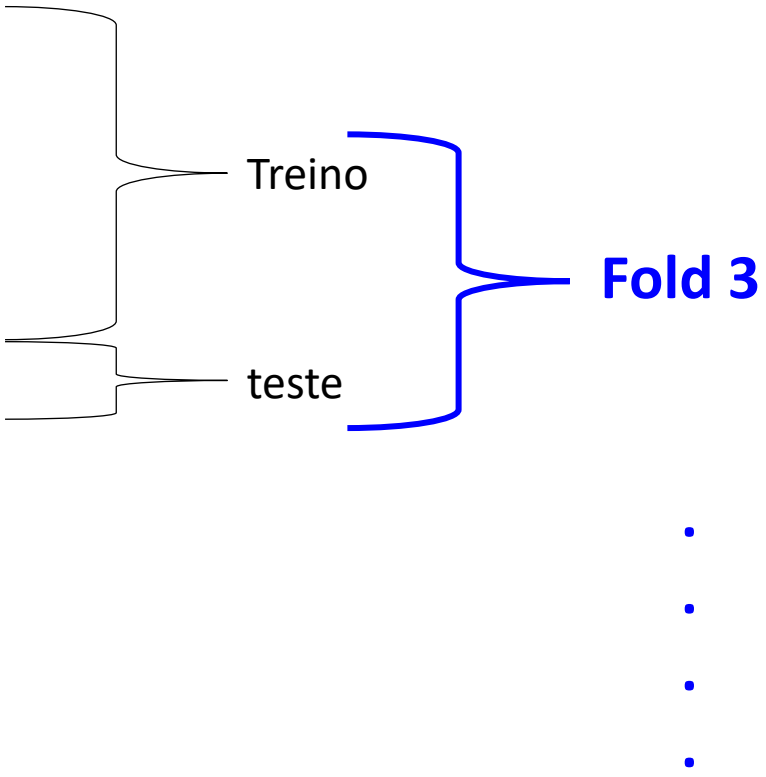
Data	Preço(R$)	Volume (ações)
2023-01-01	28.50	1.000.000
2023-01-02	29.10	1.200.000
2023-01-03	28.75	950.000
2023-01-04	29.30	1.100.000
2023-01-05	30.05	1.500.000
2023-01-06	29.90	1.300.000
2023-01-07	30.20	1.400.000
2023-01-08	30.50	1.600.000
2023-01-09	30.25	1.450.000
2023-01-10	30.80	1.700.000



Exemplo de dados temporais

Contexto: Preços históricos da ação PETR4 (Petrobras) ao longo de 10 dias.

Data	Preço(R$)	Volume (ações)
2023-01-01	28.50	1.000.000
2023-01-02	29.10	1.200.000
2023-01-03	28.75	950.000
2023-01-04	29.30	1.100.000
2023-01-05	30.05	1.500.000
2023-01-06	29.90	1.300.000
2023-01-07	30.20	1.400.000
2023-01-08	30.50	1.600.000
2023-01-09	30.25	1.450.000
2023-01-10	30.80	1.700.000



Dúvidas?

Ciência de Dados II

Professor: Gabriel Machado Lunardi
gabriel.lunardi@ufsm.br