



UNIVERSIDADE FEDERAL DE PERNAMBUCO  
CENTRO DE INFORMÁTICA  
RESIDÊNCIA EM INTELIGÊNCIA ARTIFICIAL E ROBÓTICA

GLEYCE ALVES PEREIRA DA SILVA  
JOHNNY CLEITON WILLY DA SILVA  
WILLIAMS KEIICHI SILVEIRA SATO.

ATIVIDADE FINAL - APRENDIZADO DE MÁQUINA APLICADO A  
TESTES

Recife-PE  
2023

GLEYCE ALVES PEREIRA DA SILVA  
JOHNNY CLEITON WILLY DA SILVA  
WILLIANS KEIICHI SILVEIRA SATO.

## ATIVIDADE FINAL - APRENDIZADO DE MÁQUINA APLICADO A TESTES

Relatório apresentado como requisito à obtenção de  
aprovação na disciplina de Aprendizagem de Máquinas.

Prof. Dr. Ricardo Bastos Cavalcante Prudenci

.

Recife-PE  
2023

# Sumário

	Páginas
<b>1 NoShow Dataset</b>	<b>4</b>
<b>2 Algoritmo de árvores de decisão (J48)</b>	<b>4</b>
2.1 J48 com todas as variáveis . . . . .	4
2.2 Teste 2 . . . . .	6
2.3 Teste 3 . . . . .	6
2.4 Teste 4 . . . . .	7
2.5 Teste 5 . . . . .	7
2.6 Teste 6 . . . . .	8
2.7 Teste 7 . . . . .	9
2.8 Teste 8 . . . . .	9
<b>3 Algoritmos de classificação</b>	<b>9</b>
<b>4 Experimentos de validação cruzada</b>	<b>9</b>
4.1 Apresente os resultados e comente que algoritmos foram mais bem sucedidos nos problemas analisados, considerando a métrica da área sob a curva ROC. . .	9
4.2 Comente que algoritmos foram estatisticamente superiores ao baseline (ZeroR)	9
<b>5 Curva ROC</b>	<b>9</b>

# 1 NoShow Dataset

O conjunto de dados utilizado é o "NoShow Appointments Dataset" e possui 11 variáveis. As quais você pode encontrar uma breve explicação abaixo:

1. Delay: Tempo em dias entre a marcação da consulta e a data efetiva da consulta.
2. Gender: Gênero do paciente.
3. Age: Idade do paciente.
4. Neighbourhood: Bairro onde o paciente reside.
5. Scholarship: Indica se o paciente recebe benefício do programa Bolsa Família.
6. Hypertension: Indica se o paciente possui hipertensão.
7. Diabetes: Indica se o paciente possui diabetes.
8. Alcoholism: Indica se o paciente possui problemas de alcoolismo.
9. Handicap: Indica o grau de deficiência do paciente (0: sem deficiência, 1: deficiência leve, 2: deficiência moderada, 3: deficiência grave, 4: deficiência muito grave).
10. SMS received: Indica se o paciente recebeu uma mensagem de lembrete da consulta.
11. No-show: Indica se o paciente compareceu ou não à consulta (No: compareceu, Yes: não compareceu).

## 2 Algoritmo de árvores de decisão (J48)

Aplique o algoritmo de árvores de decisão (J48) ou de regras e inspecione o conhecimento adquirido; por exemplo, identifique regras com boa cobertura e confiança e interprete o resultado; vc pode aplicar árvores de decisão com subconjuntos de variáveis de interesse.

### 2.1 J48 com todas as variáveis

Primeiro, rodamos o algoritmo J48 selecionando toda as 11 variáveis e o resultado foi:  
Time taken to build model: 0.14 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances	4410	79.819
Incorrectly Classified Instances	1115	20.181%
Kappa statistic	0	
Mean absolute error	0.3222	
Root mean squared error	0.4014	
Relative absolute error	99.9778%	
Root relative squared error	100%	
Total Number of Instances	5525	

**=== Detailed Accuracy By Class ===**

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
1,000	1,000	0,798	1,000	0,888	?	0,499	0,798
0,000	0,000	?	0,000	?	?	0,499	0,201
Weighted Avg.	0,798	0,798	?	0,798	?	0,499	0,677

**Tabela 1.** *Parte 1 da tabela*

Class
No
Yes

**Tabela 2.** *Parte 2 da tabela*

**=== Confusion Matrix ===**

Classified as		Actual Class
a	b	
4410	0	a = No
1115	0	b = Yes

**Tabela 3.** *Classificação*

No primeiro teste utilizando a árvore de decisão J48 sem nenhuma alteração, obtemos os seguintes resultados:

- Acurácia (Correctly Classified Instances): 79.819%
- Precisão (Precision): 0.798
- Cobertura (Recall): 1.0

- Utilizamos a técnica de validação cruzada com 300 folds.

Esses resultados indicam que o modelo obteve uma acurácia de 79.819%, ou seja, classificou corretamente aproximadamente 79.819% das instâncias. A precisão foi de 0.798, o que significa que, entre as instâncias classificadas como positivas, cerca de 79.8% foram classificadas corretamente. A cobertura foi de 1.0, indicando que o modelo identificou corretamente todas as instâncias positivas. Utilizamos a validação cruzada com 300 folds para avaliar o desempenho do modelo de forma mais robusta.

## **2.2 Teste 2**

No segundo teste, utilizamos o filtro Resample para realizar o balanceamento dos dados. Os resultados obtidos foram os seguintes:

- Acurácia (Correctly Classified Instances): 87.819%
- Precisão (Precision): 0.949
- Cobertura (Recall): 0.926
- Utilizamos a técnica de validação cruzada com 300 folds.

Isso significa que, no segundo teste, o modelo alcançou uma acurácia de 87.819%, ou seja, aproximadamente 87.819% das instâncias foram classificadas corretamente. A precisão foi de 0.949, indicando que, entre as instâncias classificadas como positivas, cerca de 94.9% foram classificadas corretamente. A cobertura foi de 0.926, o que significa que o modelo identificou corretamente cerca de 92.6% das instâncias positivas.

Utilizamos a validação cruzada com 300 folds para avaliar o desempenho do modelo de forma mais robusta e garantir resultados confiáveis.

## **2.3 Teste 3**

No terceiro teste, utilizamos o filtro Resample mantendo-o, porém removemos o atributo "Delay" do conjunto de dados. Os resultados obtidos foram os seguintes:

- Acurácia (Correctly Classified Instances): 86.371%
- Precisão (Precision): 0.888
- Cobertura (Recall): 0.949
- Utilizamos a técnica de validação cruzada com 300 folds.

Comparado aos testes anteriores, observamos que a acurácia teve uma diminuição em relação ao teste anterior, atingindo 86.371%. Isso significa que aproximadamente 86.371% das instâncias foram classificadas corretamente. A precisão também apresentou uma queda, alcançando o valor de 0.888, indicando que, entre as instâncias classificadas como positivas, cerca de 88.8% foram classificadas corretamente.

Por outro lado, a cobertura (Recall) apresentou um aumento para 0.949, indicando que o modelo foi capaz de identificar corretamente aproximadamente 94.9% das instâncias positivas.

Esses resultados foram obtidos usando o filtro Resample e removendo o atributo "Delay" do conjunto de dados. O processo de validação cruzada com 300 folds foi aplicado para avaliar o desempenho do modelo de forma mais robusta e confiável.

## **2.4 Teste 4**

No quarto teste, utilizamos o filtro Resample mantendo-o, porém removemos os atributos "Gender" e "Neighborhood" do conjunto de dados. Os resultados obtidos foram os seguintes:

- Acurácia (Correctly Classified Instances): 80.9231
- Precisão (Precision): 0.818
- Cobertura (Recall): 0.981
- Utilizamos a técnica de validação cruzada com 300 folds.

Comparado aos testes anteriores, observamos que a acurácia teve uma diminuição em relação ao teste anterior, atingindo 80.9231%. Isso significa que aproximadamente 80.9231% das instâncias foram classificadas corretamente. A precisão também apresentou uma queda, alcançando o valor de 0.818, indicando que, entre as instâncias classificadas como positivas, cerca de 81.8% foram classificadas corretamente.

Por outro lado, a cobertura (Recall) apresentou um aumento para 0.981, indicando que o modelo foi capaz de identificar corretamente aproximadamente 98.1% das instâncias positivas.

Esses resultados foram obtidos usando o filtro Resample e removendo os atributos "Gender" e "Neighborhood" do conjunto de dados. O processo de validação cruzada com 300 folds foi aplicado para avaliar o desempenho do modelo de forma mais robusta e confiável.

## **2.5 Teste 5**

No quinto teste, utilizamos o filtro Resample mantendo-o, porém removemos os atributos "Diabetes" e "Alcoholism" do conjunto de dados. Os resultados obtidos foram os seguintes:

- Acurácia (Correctly Classified Instances): 79.819%
- Precisão (Precision): 0.798

- Cobertura (Recall): 1.0
- Utilizamos a técnica de validação cruzada com 300 folds.

Comparado aos testes anteriores, observamos que a acurácia teve uma diminuição em relação ao teste anterior, atingindo 79.819%. Isso significa que aproximadamente 79.819% das instâncias foram classificadas corretamente. A precisão também apresentou uma queda, alcançando o valor de 0.798, indicando que, entre as instâncias classificadas como positivas, cerca de 79.8% foram classificadas corretamente.

Por outro lado, a cobertura (Recall) apresentou um aumento para 1.0, indicando que o modelo foi capaz de identificar corretamente todas as instâncias positivas.

Esses resultados foram obtidos usando o filtro Resample e removendo os atributos "Diabetes" e "Alcoholism" do conjunto de dados. O processo de validação cruzada com 300 folds foi aplicado para avaliar o desempenho do modelo de forma mais robusta e confiável.

## 2.6 Teste 6

No sexto teste, utilizamos o filtro Class Ordem e obtivemos os seguintes resultados:

- Acurácia (Correctly Classified Instances): 79.819%
- Precisão (Precision): ?
- Cobertura (Recall): 0.0
- Utilizamos a técnica de validação cruzada com 300 folds.

Neste teste, a acurácia foi de 79.819%, indicando que aproximadamente 79.819% das instâncias foram classificadas corretamente. No entanto, o valor da precisão não foi fornecido (?), o que significa que o valor é desconhecido ou indefinido para esta métrica. A cobertura (Recall) apresentou um valor de 0.0, indicando que o modelo não conseguiu identificar corretamente nenhuma das instâncias positivas.

Esses resultados foram obtidos usando o filtro Class Ordem e o processo de validação cruzada com 300 folds. É importante ressaltar que a falta de informações sobre a precisão e a baixa cobertura podem indicar problemas ou limitações com o modelo ou com a configuração utilizada.



## **2.7 Teste 7**

## **2.8 Teste 8**

# **3 Algoritmos de classificação**

Escolha um algoritmo de classificação (por exemplo, IBk) e use técnicas de seleção de atributos para identificar as variáveis consideradas mais relevantes para o problema de classificação escolhido. Analise o impacto da seleção de atributos, avaliando o algoritmo escolhido usando (a) todos os atributos do problema, ou seja, sem seleção (b) e apenas usando os atributos selecionados.

# **4 Experimentos de validação cruzada**

Usando o ambiente Experimenter do WEKA, realize experimentos de validação cruzada com os algoritmos ZeroR, Regressão Logística, IBk, J48, Naive Bayes e Random Forest.

**4.1 Apresente os resultados e comente que algoritmos foram mais bem sucedidos nos problemas analisados, considerando a métrica da área sob a curva ROC.**

**4.2 Comente que algoritmos foram estatisticamente superiores ao baseline (ZeroR)**

# **5 Curva ROC**

## **Referências**