5. ESTUDO DE CASO

Neste capítulo, foram realizados experimentos para testar os cinco algoritmos de aprendizado de máquina descritos no capítulo 3. Após uma descrição mais detalhada, os parâmetros fornecidos pela FOT foram submetidos a experimentos individuais e em conjunto. Outros testes foram feitos com aplicação de métodos como produto cruzado e seleção de variáveis. Em seguida, foram selecionadas estruturas geradas pelas Redes Bayesianas sintetizadas com Algoritmo Genético, para análise das ligações entre as variáveis usadas e suas tabelas de distribuição de probabilidade conjunta.

5.1. Descrição do Conjunto de Dados

O conjunto de dados usado neste projeto foi obtido por um sistema de oscilações forçadas (FOT), desenvolvido no Laboratório de Instrumentação Biomédica da UERJ. Os exames foram realizados em 23 indivíduos do grupo controle e 27 portadores de Fibrose Cística, que formam um grupo de teste. Em cada exame foram feitas três medidas, o que totalizou um conjunto de dados de 150 instâncias para os experimentos. As informações fornecidas pela FOT estão na tabela a seguir:

Tabela 12 – Parâmetros fornecidos pela FOT

Parâmetro	Descrição do Parâmetro
R_o	Resistência no Intercepto
R_m	Resistência Média
X_m	Reatância Média
C_{din}	Complacência Dinâmica
S	Inclinação da Curva de Resistência
Z_{4Hz}	Impedância em 4Hz
F_r	Frequência de Ressonância
E_{din}	Elastância Dinâmica

As características de indivíduos pertencentes ao grupo controle e teste, foram comparadas na Figura 22. Os gráficos com barras mostram as médias das variáveis calculadas em um intervalo de confiança de 95%, cujos valores de desvio padrão são indicados acima ou abaixo das barras. Por exemplo, o valor médio da resistência R_o no grupo controle é

2,87±0,76. Já no grupo de teste, a média de R_o sobe para 4,56±1,61. Através da Análise de Variância (ANOVA), todos os parâmetros da FOT mostraram diferença significativa nos seus respectivos valores de média ($p<0,001$).

De acordo com a Figura 22, houve um aumento na média das variáveis R_o, R_m, Z_{4Hz}, F_r e E_{din} dos indivíduos do grupo teste, se comparado aos do grupo controle. Ou seja, indivíduos portadores de fibrose cística geralmente possuem valores mais altos de resistências (R_o e R_m), impedância (Z_{4Hz}), frequência de ressonância (F_r) e elastância (E_{din}), se comparado a não portadores da doença. Já as variáveis X_m, C_{din} e S do grupo teste, apresentaram uma diminuição em sua média. Logo, conclui-se que portadores de fibrose cística possuem valores mais negativos de reatância (X_m) e inclinação da curva de resistência (*S*), e valores menores de complacência (C_{din}).

Para submeter os dados da FOT nas Redes Bayesianas, todas as amostras do conjunto de dados foram discretizadas e para cada característica da Tabela 12 foi estabelecido um ponto de corte (Tabela 13). Os valores abaixo desse ponto foram rotulados como *1*, representando valores mais baixos que a variável pode assumir. Já os valores acima do ponto de corte foram rotulados como *2*, representando os valores mais altos da variável. No caso da variável *classe*, indivíduos do grupo controle receberam o rótulo *0*, e indivíduos do grupo teste receberam o rótulo *1*. Com base nessas informações, o comportamento geral das características da FOT pode ser resumido conforme a Tabela 14.

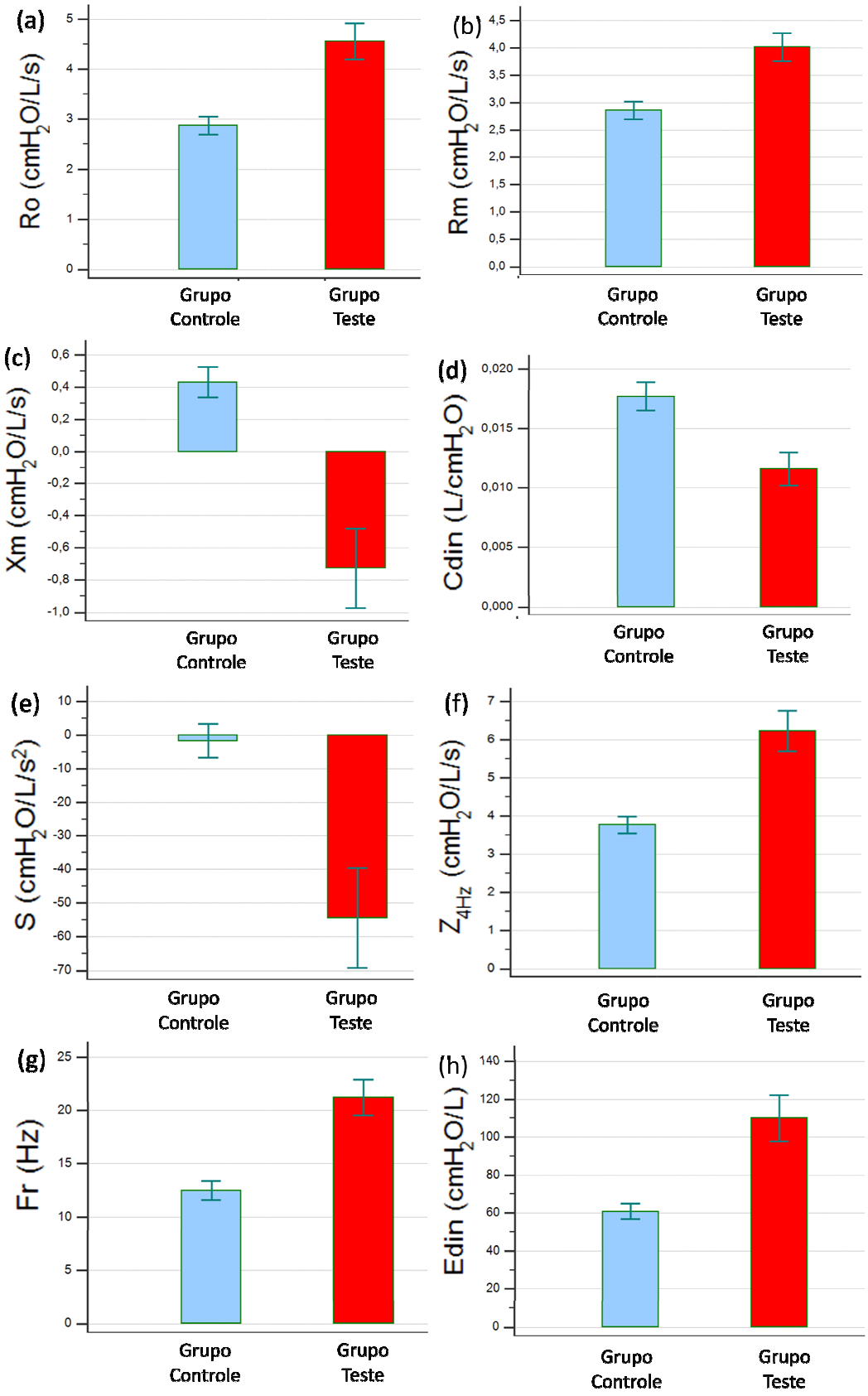

Figura 22 – Comparação dos parâmetros da FOT de indivíduos do grupo controle e do grupo teste

Tabela 13 – Pontos de corte para discretização dos parâmetros da FOT, média e desvio padrão

Parâmetro	Ponto de corte	Média (±Desvio Padrão)
R_o	3,31	3,78 ± 1,54
R_m	3,21	3,48 ± 1,13
X_m	0,18	-0,19 ± 1,04
C_{din}	0,014	0,014 ± 0,007
S	-10,25	-30,15 ± 57,61
Z_{4Hz}	4,44	5,10 ± 2,25
F_r	14,19	17,20 ± 7,50
E_{din}	72,54	87,40 ± 48,80

Tabela 14 – Comportamento geral das características do grupo controle e do grupo teste

	R_o	R_m	Z_{4Hz}	F_r	E_{din}	X_m	C_{din}	S	*Classe*
Grupo Controle	*1*	*1*	*1*	*1*	*1*	*2*	*2*	*2*	*0*
Grupo Teste	*2*	*2*	*2*	*2*	*2*	*1*	*1*	*1*	*1*

5.2. Experimento Individual dos Atributos

Cada atributo fornecido pela FOT foi submetido à análise individual para que seu desempenho em classificar pacientes fosse testado. Todos os parâmetros tiveram suas medidas de AUC, erro padrão (E_{AUC}) e intervalo de confiança de 95% (IC 95%) calculados (DELONG et al., 1988), conforme Tabela 15.

A reatância X_m e a frequência F_r foram os parâmetros que apresentaram melhor desempenho individual com valores de AUC iguais a 0,85 e 0,84, respectivamente. Os demais parâmetros apresentaram valores de AUC entre 0,76 e 0,82. Sendo assim, o desempenho de todos os atributos analisados separadamente se enquadra na faixa de acurácia moderada (0,70 a 0,90), sendo observada diferença significativa apenas entre os valores de AUC de S e F_r ($p<0,01$) e entre S e X_m ($p<0,005$).

As curvas ROC com o desempenho de cada atributo foram traçadas, mostrando que a área sob a curva é maior na faixa final do eixo x, onde uma maior quantidade de falsos positivos é aceita (Figura 23).

Tabela 15 – Desempenho individual dos parâmetros da FOT na classificação de pacientes

	AUC	E_{AUC}	IC 95%
F_r (Hz)	0,84	0,03	0,77-0,89
X_m ($cmH_2O/L/s$)	0,85	0,03	0,78-0,90
R_o ($cmH_2O/L/s$)	0,80	0,04	0,73-0,86
S ($cmH_2O/L/s^2$)	0,76	0,04	0,68-0,83
R_m ($cmH_2O/L/s$)	0,78	0,04	0,70-0,84
C_{din} (L/cmH_2O)	0,82	0,04	0,75-0,88
E_{din} (cmH_2O/L)	0,82	0,04	0,75-0,88
Z_{4Hz} ($cmH_2O/L/s$)	0,82	0,04	0,75-0,88

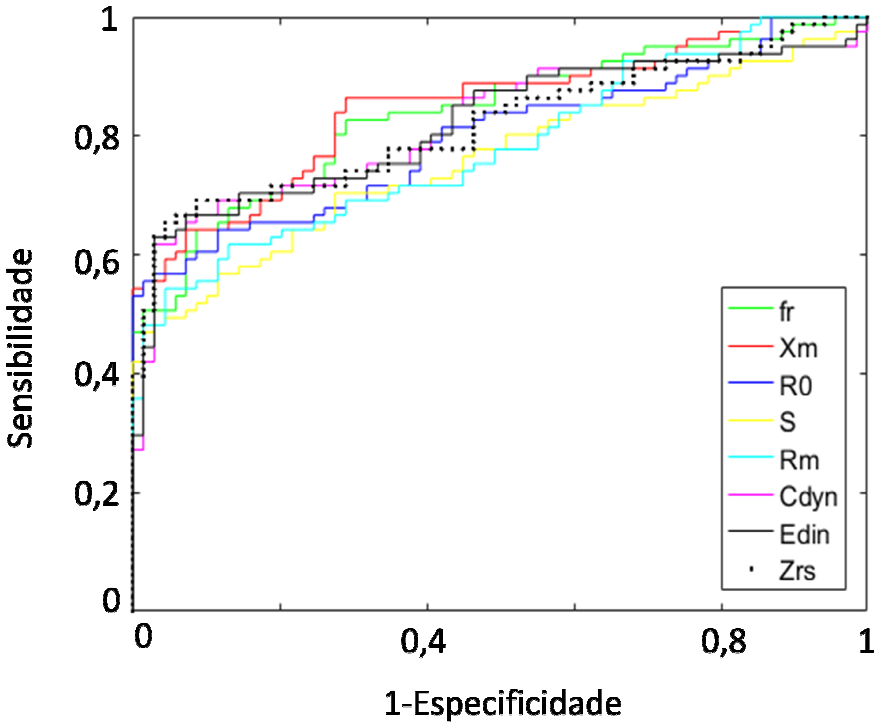

Figura 23 – Curvas ROC dos parâmetros da FOT

5.3. Experimento com Oito Atributos

Para esse experimento, os oito atributos da FOT foram usados nos seguintes algoritmos de aprendizado de máquina: 1-*Nearest Neighbor* (1-NN), *Adaboost* (ADAB), *Random Forest* (RF), *Radial Support Vector Machine* (RSVM) e Redes Bayesianas sintetizadas com Algoritmos Genéticos (RBGAOT). Esses cinco classificadores também foram comparados com o melhor parâmetro da FOT (MPF), a variável X_m. Pela Tabela 16, pode-se observar que o algoritmo RBGAOT apresentou melhor desempenho com AUC igual a 0,88. O segundo melhor resultado foi obtido pelo 1-NN com valor de AUC igual a 0,87.

Além da AUC, foram calculadas as probabilidades de ter um resultado positivo quando o indivíduo for portador da doença, denominada sensibilidade (Sens). Também foram calculadas as probabilidades de ter um resultado negativo quando o indivíduo não portar a doença, denominada especificidade (Esp). O intervalo de confiança está abaixo dos respectivos valores de Sens, Esp e AUC (DELONG et al, 1988). Na Figura 24, pode-se observar que a área sob a curva ROC dos classificadores é maior na faixa final dos eixos, onde é aceito maior quantidade de casos falsos positivos.

Tabela 16 – Resultado dos oito parâmetros da FOT submetidos aos classificadores

	Sens (%)	**Esp (%)**	**AUC**	**E_{AUC}**
MPF	86,42 (77,0-93,0)	71,01 (58,8-81,3)	0,85 (0,78-0,90)	0,03
1-NN	82,72 (72,7-90,2)	81,16 (69,9-89,6)	0,87 (0,81-0,92)	0,03
ADAB	81,48 (71,3-89,2)	78,26 (66,7-87,3)	0,86 (0,79-0,91)	0,03
RF	74,07 (63,1-83,2)	85,51 (75,0-92,8)	0,86 (0,79-0,91)	0,03
RSVM	60,49 (49,0-71,2)	100 (94,8-100,0)	0,82 (0,75-0,88)	0,04
RBGAOT	80,25 (69,9-88,3)	78,26 (66,7-87,3)	0,88 (0,82-0,93)	0,03

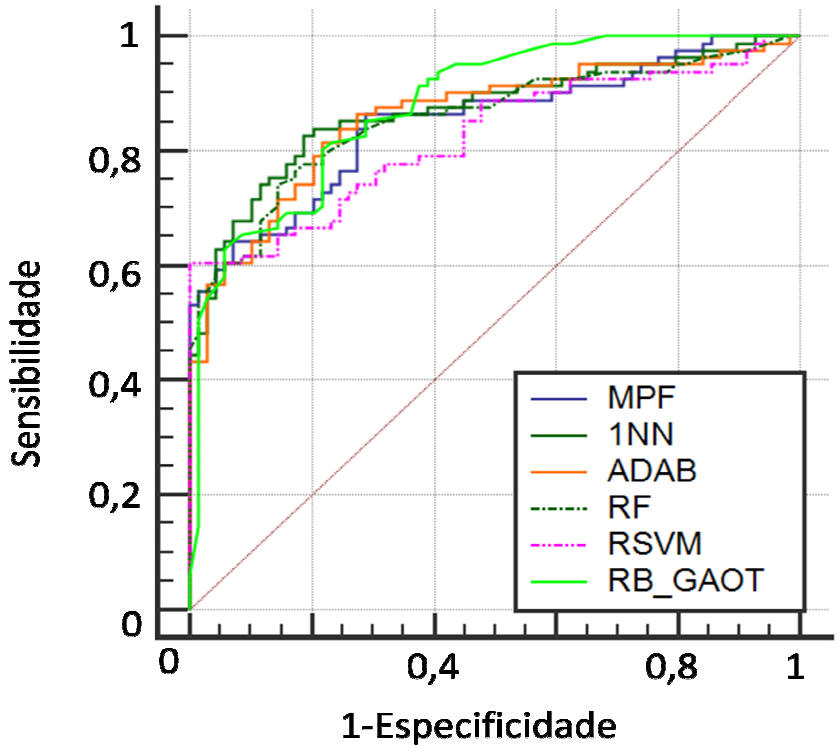

Figura 24 – Curvas ROC do experimento com todos os parâmetros da FOT

A Tabela 17 mostra, em pares, a comparação das áreas sobre as curvas ROC de todos os métodos e o erro padrão em um intervalo de confiança de 95% (DELONG et al, 1988), sendo a interseção entre linha e coluna a diferença entre dois classificadores. Nesse experimento, não foi observada diferença significativa em nenhum dos quinze pares analisados (p>0,05). Isto pode ter ocorrido devido ao tamanho do conjunto de dados usado que conta apenas com 150 instâncias, limitando assim, a quantidade de amostras usada para os testes.

Tabela 17 – Comparação dos valores de AUC dos classificadores no experimento com todos os atributos da FOT

	1-NN	ADAB	RF	RSVM	RBGAOT
MPF	0,02±0,035	0,01±0,031	0,01±0,028	0,03±0,029	0,03±0,036
1-NN	-	0,01±0,022	0,01±0,024	0,05±0,035	0,01±0,037
ADAB	-	-	0,003±0,014	0,04±0,033	0,02±0,038
RF	-	-	-	0,03±0,031	0,02±0,034
RSVM	-	-	-	-	0,06±0,039

É possível comparar a sensibilidade ao observar a especificidade com valores fixos. Essa análise limita os falsos positivos, fornecendo medidas em situações onde o algoritmo dificilmente erra a classificação dos portadores da doença. Nesse experimento, foram escolhidas especificidades com valores fixados em 75%, representando um valor moderado e 90%, representando alta especificidade (Figura 25). Esses valores limitam respectivamente, em 25 e 10% os casos de falsos positivos.

Com a especificidade fixada em 75%, observou-se melhora no desempenho dos classificadores 1-NN, ADAB, RF e RBGAOT, fazendo com que eles alcançassem valores de sensibilidade acima de 80%. De forma geral, tanto o MPF, quanto os classificadores estão dentro da faixa de sensibilidade moderada (70 a 90%).

Já com a especificidade fixada em 90%, observou-se uma diminuição nos valores da sensibilidade, fazendo com que todos os classificadores ficassem abaixo da faixa moderada. Os melhores resultados a uma especificidade de 75% ocorrem devido à maior tolerância a falsos positivos, se comparado a 90%.

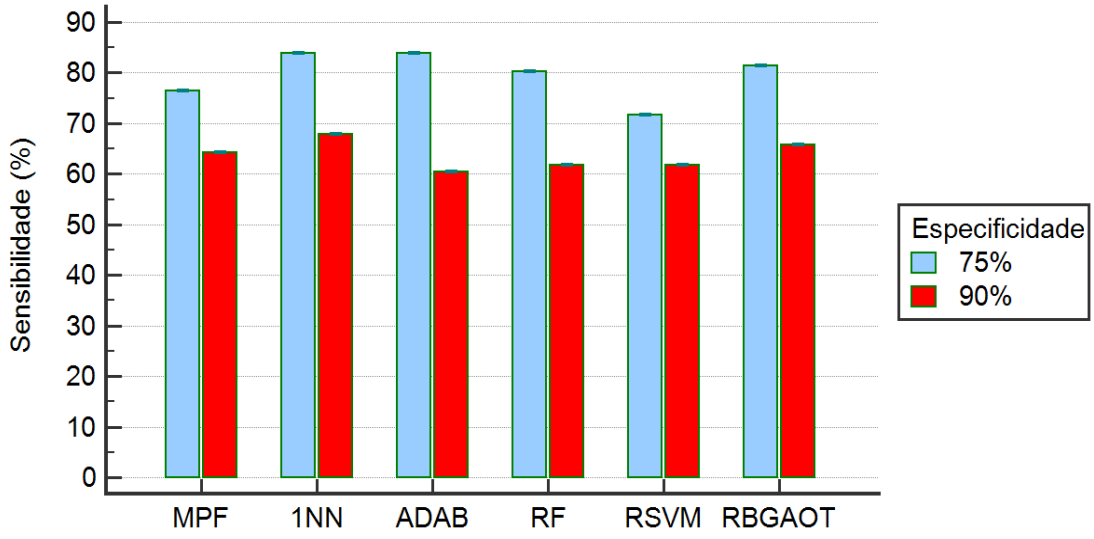

Figura 25 – Análise da sensibilidade com especificidade em 75% e 90% no experimento com oito atributos

5.4. Experimento com Oito Atributos Cruzados

Nesse experimento, o produto cruzado dos oito parâmetros da FOT foi aplicado como entrada dos classificadores. Foram geradas 36 combinações que, somadas à variável *classe*, totalizaram um conjunto com 37 atributos. Para representar possíveis soluções no RBGAOT, são necessárias matrizes de tamanho 37x37. Durante a marginalização da rede, o método *Junction Tree* (BARBER, 2003), fornecido pela *toolbox* PGM, realiza diversos processos que geram um alto custo computacional, fazendo com que o algoritmo não conseguisse convergir. No caso dos demais algoritmos, não houve falhas e o experimento pode ser realizado. As combinações geradas nesse experimento estão disponíveis no Apêndice A.

No algoritmo 1-NN houve um aumento no valor da AUC de 0,87 para 0,89 com relação ao experimento anterior. Os demais classificadores não mostraram mudança no valor da AUC, porém foram observadas alterações nos valores da sensibilidade e especificidade, conforme Tabela 18. Com base nesses valores, a curva ROC pode ser calculada para os quatro classificadores que finalizaram o experimento, além do melhor parâmetro da FOT. Por meio da Figura 26, observa-se que a curva ROC dos classificadores é maior no final dos eixos, pois esse trecho possibilita maior quantidade de falsos positivos.

Tabela 18 – Resultado do experimento com oito atributos cruzados submetidos aos classificadores

	Sens (%)	**Esp (%)**	**AUC**	**E_{AUC}**
MPF	86,42 (77,0-93,0)	71,01 (58,8-81,3)	0,85 (0,78-0,90)	0,03
1-NN	80,25 (69,9-88,3)	81,16 (69,9-89,6)	0,89 (0,83-0,94)	0,03
ADAB	67,90 (56,6-77,8)	91,30 (82,0-96,7)	0,86 (0,79-0,91)	0,03
RF	81,48 (71,3-89,2)	78,26 (66,7-87,3)	0,86 (0,80-0,91)	0,03
RSVM	58,02 (46,5-68,9)	98,55 (92,2-100,0)	0,82 (0,75-0,86)	0,03
RBGAOT	-	-	-	-

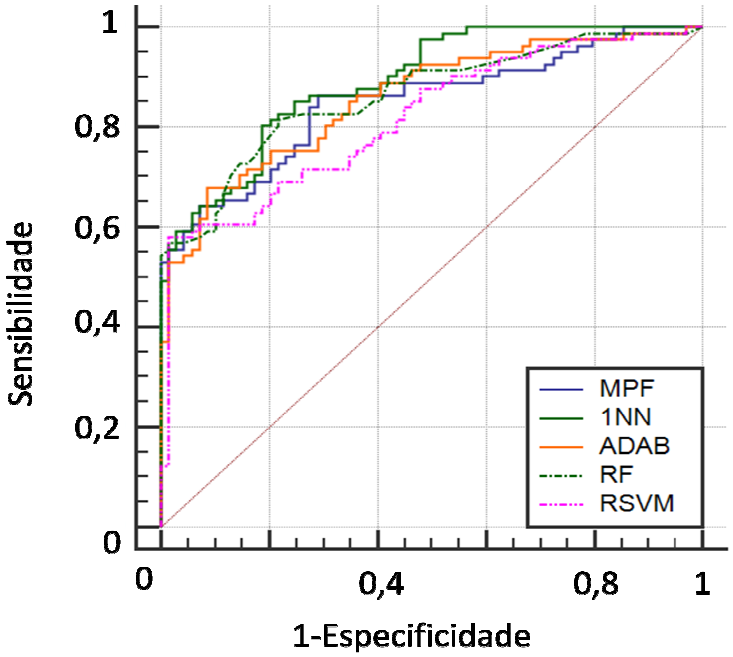

Figura 26 – Curvas ROC do experimento com oito atributos cruzados

As curvas ROC obtidas pelos quatro algoritmos foram comparadas, conforme a Tabela 19. A diferença e o erro padrão de cada par foram calculados em um intervalo de confiança de 95% (DELONG et al, 1988). Dentre os dez pares, não foram encontradas diferenças significativas (p>0,05).

Tabela 19 – Comparação dos valores da AUC dos classificadores no experimento com oito atributos cruzados

	1-NN	ADAB	RF	RSVM
MPF	0,04±0,03	0,01±0,03	0,02±0,026	0,03±0,031
1-NN	-	0,03±0,024	0,03±0,025	0,07±0,031
ADAB	-	-	0,007±0,014	0,04±0,028
RF	-	-	-	0,05±0,028

A Figura 27 mostra a comparação da sensibilidade com valores de especificidade em 75% e 90%. No primeiro caso, os classificadores 1-NN e RF apresentaram valores de sensibilidade acima de 80%, já o RSVM ficou abaixo da faixa moderada (70 a 90%). Com a especificidade a 90%, todos os classificadores apresentaram valores abaixo de 70%, assim como no experimento anterior.

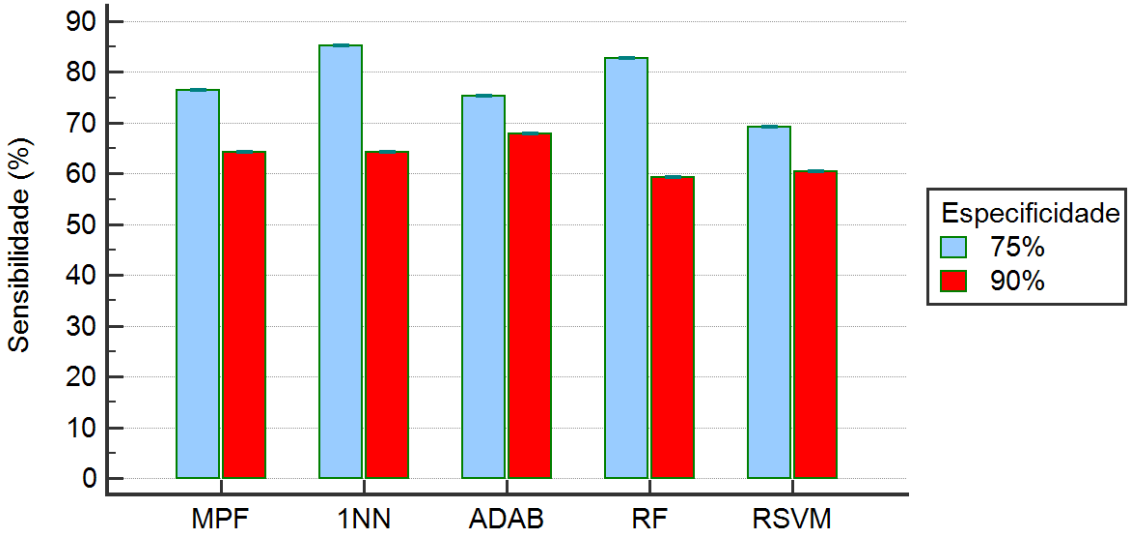

Figura 27 – Análise da sensibilidade com valores de especificidade em 75% e 90% no experimento com oito atributos cruzados

5.5. Experimento com Cinco Atributos Selecionados

Na tentativa de melhorar o desempenho dos classificadores, cinco atributos fornecidos pela FOT foram selecionados para o experimento. Essa seleção foi feita pela aplicação da função *featself* que utiliza o algoritmo 1-NN para classificar os atributos do conjunto de dados. A validação cruzada pelo método *leave-one-out* também foi usada para a seleção dos melhores atributos, sendo, a cada iteração, uma amostra escolhida para teste e as demais para treino. O critério para seleção é o resultado dessa classificação. Da mesma forma, foi testado o uso da AUC como critério de avaliação, porém as variáveis selecionadas não apresentaram resultado superior às selecionadas pelo 1-NN, que foram: R_o, R_m, X_m, C_{din} e Z_{4Hz}. Essas cinco variáveis também estão de acordo com a seleção feita por um especialista.

De acordo com a Tabela 20, o algoritmo 1-NN apresentou melhor desempenho com AUC no valor de 0,87. Já o RSVM e o algoritmo RBGAOT apresentaram desempenho inferior com AUC iguais a 0,77 e 0,79, respectivamente. Os algoritmos ADAB e RF permaneceram com AUC iguais a 0,86, porém com alterações nos valores da sensibilidade e especificidade. A Figura 28 mostra a curva ROC com o desempenho dos cinco classificadores, além do melhor parâmetro da FOT. Observa-se que na faixa final dos eixos, onde maior quantidade de falsos positivos é aceita, a curva ROC desses classificadores é maior.

Tabela 20 – Resultado do experimento com a seleção de cinco atributos submetidos aos classificadores

	Sens (%)	Esp (%)	AUC	E_{AUC}
MPF	86,42 (77,0-93,0)	71,01 (58,8-81,3)	0,85 (0,78-0,90)	0,03
1-NN	70,37 (59,2-80,0)	92,75 (83,9-97,6)	0,87 (0,81-0,92)	0,03
ADAB	74,07 (63,1-83,2)	89,86 (80,2-95,8)	0,86 (0,80-0,91)	0,03
RF	72,84 (61,8-82,1)	91,30 (82,0-96,7)	0,86 (0,79-0,91)	0,03
RSVM	60,49 (49,0-71,2)	92,75 (83,9-97,6)	0,77 (0,69-0,83)	0,04
RBGAOT	82,72 (72,7-90,2)	75,36 (63,5 - 84,9)	0,79 (0,72-0,86)	0,04

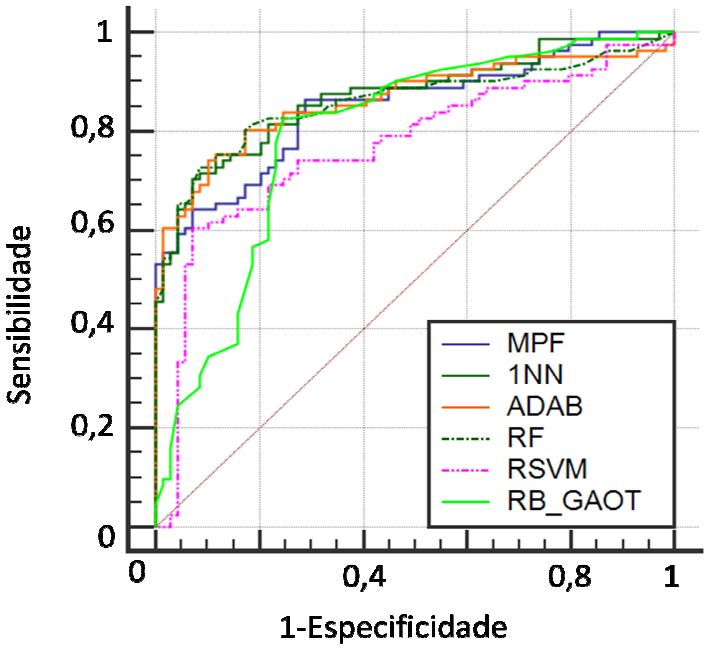

Figura 28 – Curvas ROC do experimento com seleção de atributos da FOT

A Tabela 21 mostra a diferença entre os valores das curvas ROC dos cinco algoritmos testados e o MPF. Os cálculos foram feitos em pares, com os erros padrões em um intervalo de confiança de 95% (DELONG et al, 1988). Dentre as quinze áreas comparadas, não foram encontradas diferenças significativas (p>0,05).

Tabela 21 – Comparação dos valores de AUC dos classificadores no experimento com seleção de atributos da FOT

	1-NN	ADAB	RF	RSVM	RBGAOT
MPF	0,02±0,033	0,01±0,033	0,01±0,032	0,08±0,036	0,06±0,042
1-NN	-	0,01±0,020	0,01±0,021	0,10±0,037	0,08±0,047
ADAB	-	-	0,003±0,012	0,10±0,037	0,07±0,046
RF	-	-	-	0,09±0,035	0,07±0,045
RSVM	-	-	-	-	0,03±0,051

A análise da sensibilidade realizada com valores fixados em 75% e 90% de especificidade pode ser feita na Figura 29. Em 75%, todos os classificadores apresentaram sensibilidade acima de 80%, exceto o RSVM. Já com especificidade de 90%, os algoritmos 1-NN e RF conseguiram apresentar valores na faixa de sensibilidade de 71,60% e 72,84% respectivamente, estando os demais modelos abaixo da faixa moderada (70 a 90%).

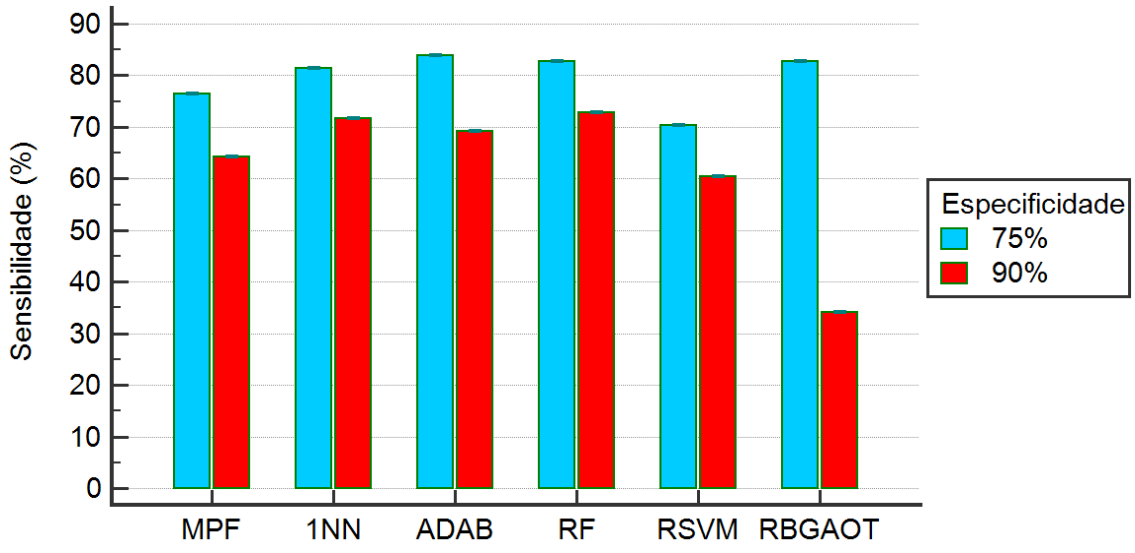

Figura 29 – Análise da sensibilidade com valores de especificidade em 75% e 90% no experimento com seleção de atributos da FOT

5.6. Experimento com Cinco Atributos Cruzados

Nesse experimento, o produto cruzado dos atributos selecionados no item 5.5, foram usados como entrada nos classificadores. Após a aplicação desse método, foram obtidas 15 combinações que, somadas a variável *classe*, totalizaram 16 variáveis de entrada para os cinco algoritmos. Nesse caso, para representar uma possível solução no RBGAOT é necessário o uso de uma matriz de tamanho 16x16. Com esse valor, foi possível passar pelo algoritmo *Junction Tree* e realizar a marginalização das redes, sem que o RBGAOT apresentasse falhas. As combinações geradas nesse experimento também estão disponíveis no Apêndice A.

Os algoritmos RF e ADAB apresentaram melhor desempenho com valores de AUC iguais a 0,89. Mesmo com a aplicação do produto cruzado, o classificador 1-NN apresentou mudança nos valores da sensibilidade e especificidade, porém não apresentou mudança no valor da AUC. No caso do RSVM e do RBGAOT, os valores da AUC aumentaram respectivamente para 0,80 e 0,88, em relação ao experimento anterior. A Figura 30 mostra as curvas ROC calculadas para os classificadores e o MPF, com base nos valores da Tabela 22. Através dessas curvas é possível observar que a curva ROC de todos os classificadores é maior na faixa final dos eixos, onde a quantidade de falsos positivos é maior.

Tabela 22 – Resultado do experimento com produto cruzado dos atributos selecionados da FOT submetidos aos classificadores

	Sens (%)	**Esp (%)**	**AUC**	**E_{AUC}**
MPF	86,42 (77,0-93,0)	71,01 (58,8-81,3)	0,85 (0,78-0,90)	0,03
1-NN	71,60 (60,5-81,1)	92,75 (83,9-97,6)	0,87 (0,81-0,92)	0,03
ADAB	69,14 (57,9-78,9)	92,75 (83,9-97,6)	0,89 (0,83-0,94)	0,03
RF	82,72 (72,7-90,2)	82,61 (71,6-90,7)	0,89 (0,83-0,94)	0,03
RSVM	83,95 (74,1-91,2)	72,46 (60,4-82,5)	0,80 (0,73-0,86)	0,03
RBGAOT	56,79 (45,3-67,8)	98,55 (92,2-100,0)	0,88 (0,81-0,92)	0,03

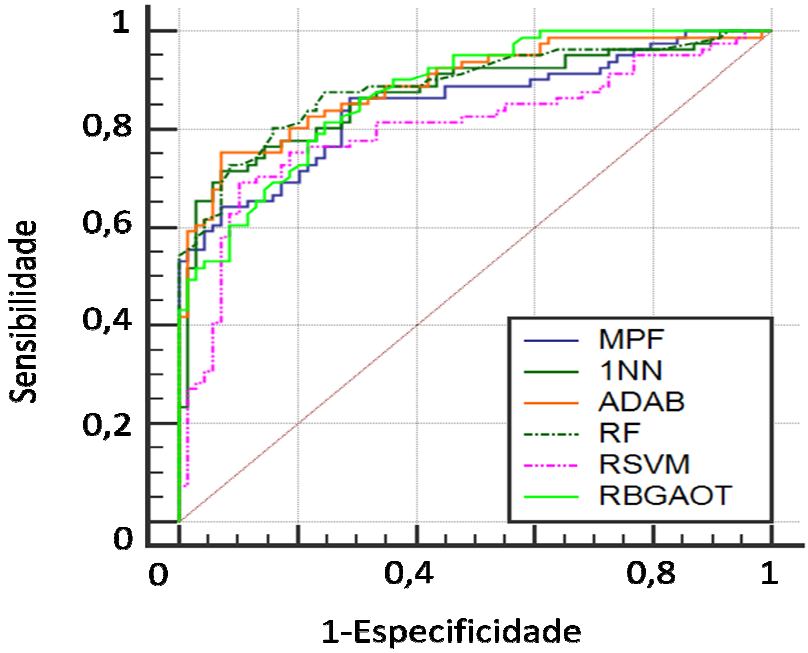

Figura 30 – Curvas ROC do experimento com cinco parâmetros da FOT cruzados

A diferença entre os valores da AUC dos classificadores e seu respectivo erro padrão calculado em um intervalo de confiança de 95% (DELONG et al, 1988), podem ser observados na Tabela 23. Dentre os valores encontrados, não foram observadas diferenças significativas (p>0,05).

Tabela 23 – Comparação dos valores de AUC dos classificadores no experimento com cinco parâmetros da FOT cruzados

	1-NN	ADAB	RF	RSVM	RBGAOT
MPF	0,02±0,041	0,04±0,038	0,04±0,039	0,05±0,049	0,03±0,041
1-NN	-	0,01±0,022	0,01±0,022	0,04±0,027	0,03±0,037
ADAB	-	-	0,02±0,017	0,03±0,031	0,01±0,040
RF	-	-	-	0,05±0,027	0,04±0,038
RSVM	-	-	-	-	0,02±0,043

A análise dos valores de sensibilidade com a especificidade fixada em 75% e 90% é mostrada na Figura 31. No primeiro caso, observou-se que a sensibilidade dos classificadores apresentou valores na faixa de sensibilidade moderada (70 a 90%), sendo que o RF chegou a 85%. No caso da especificidade em 90%, os algoritmos 1-NN, ADAB e RF conseguiram alcançar valores de sensibilidade acima de 70%, mesmo nesse caso com restrição de falsos positivos a 10%.

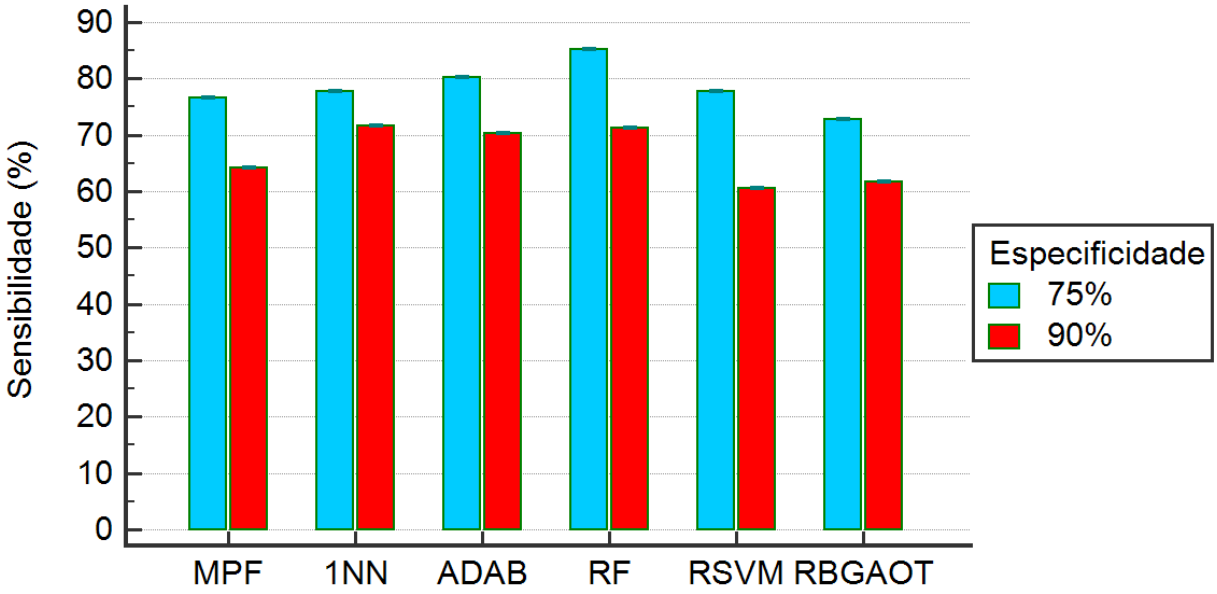

Figura 31 – Análise da sensibilidade com valores de especificidade em 75% e 90% no experimento com cinco parâmetros da FOT cruzados

5.7. Experimento com Seleção de Cinco Atributos do Produto Cruzados

Nesse experimento, a seleção de cinco atributos também foi feita pela função *featself*, como no experimento do item 5.5, com base no desempenho do algoritmo 1-NN e no método de validação cruzada *leave-one-out*. A diferença é que, nesse caso, a seleção foi feita após a aplicação do produto cruzado.

De acordo com a Tabela 24, todos os algoritmos apresentaram bom desempenho, sendo o 1-NN e ADAB com valores de AUC iguais a 0,89 e o RF e RSVM com valores de AUC iguais a 0,88. Já o RBGAOT apresentou menor desempenho com AUC igual a 0,80. O maior valor de sensibilidade foi obtido no 1-NN, já os maiores valores de especificidade foram obtidos nos classificadores ADAB e RF. A Figura 32 mostra a curva ROC com o desempenho dos classificadores, além do melhor parâmetro da FOT.

Tabela 24 – Resultado do experimento com atributos do produto cruzado selecionados e submetidos aos classificadores

	Sens (%)	Esp (%)	AUC	E_{AUC}
MPF	86,42 (77,0-93,0)	71,01 (58,8-81,3)	0,85 (0,78-0,90)	0,03
1-NN	87,65 (78,5-93,9)	81,16 (69,9-89,6)	0,89 (0,82-0,93)	0,03
ADAB	76,54 (65,8-85,2)	89,86 (80,2-95,8)	0,89 (0,83-0,94)	0,03
RF	72,84 (61,8-82,1)	94,20 (85,8-98,4)	0,88 (0,82-0,93)	0,03
RSVM	74,07 (63,1-83,2)	86,96 (76,7-93,9)	0,88 (0,83-0,93)	0,03
RBGAOT	79,01 (68,5-87,3)	71,01 (58,8-81,3)	0,80 (0,73-0,86)	0,04

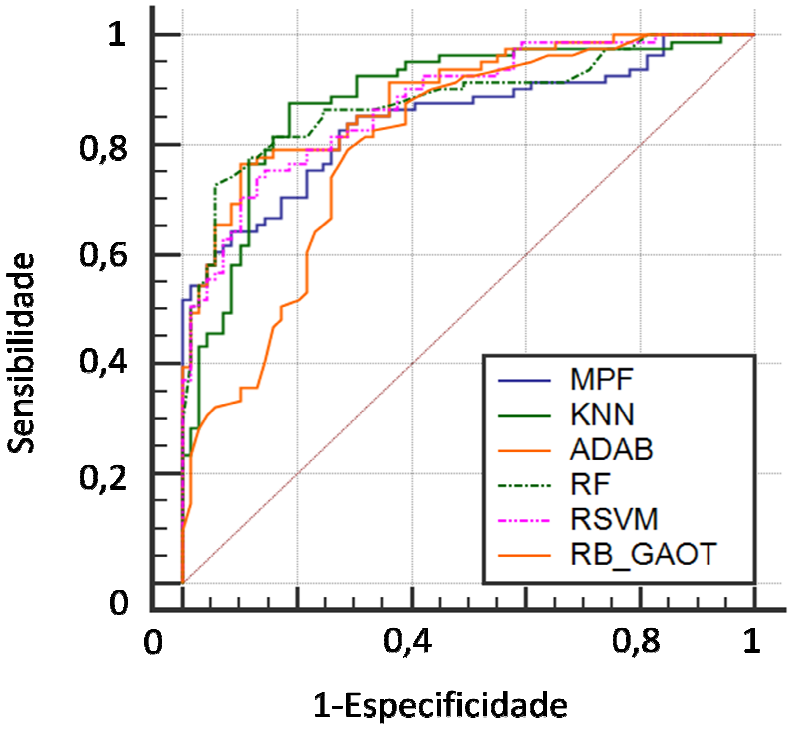

Figura 32 – Curvas ROC do experimento com atributos do produto cruzado selecionados

A Tabela 25 mostra a diferença entre os valores das curvas ROC dos algoritmos testados e o melhor parâmetro da FOT. Os cálculos foram feitos em pares, com os erros padrões em um intervalo de confiança de 95% (DELONG et al, 1988). Dentre as quinze áreas comparadas, houve diferença significativa entre o 1-NN e o RBGAOT ($p<0,05$), e entre ADAB e o RGBAOT ($p<0,05$). Nas demais combinações não foram encontradas diferenças ($p>0,05$).

Tabela 25 – Comparação dos valores de AUC dos classificadores no experimento com atributos do produto cruzado selecionados

	1-NN	ADAB	RF	RSVM	RBGAOT
MPF	0,04±0,035	0,04±0,028	0,04±0,028	0,03±0,031	0,05±0,045
1-NN	-	0,003±0,023	0,004±0,023	0,008±0,024	0,09±0,043
ADAB	-	-	0,008±0,013	0,01±0,021	0,09±0,043
RF	-	-	-	0,004±0,023	0,08±0,045
RSVM	-	-	-	-	0,08±0,045

A análise da sensibilidade feita com valores fixados em 75% e 90% de especificidade pode ser vista na Figura 33. Em 75%, todos os classificadores apresentaram sensibilidade na faixa moderada, exceto o RBGAOT, tendo o 1-NN e o RF alcançado valores de sensibilidade acima de 80%. Já com especificidade em 90%, apenas o algoritmo RF conseguiu apresentar valor de sensibilidade acima de 70%, estando os demais modelos abaixo da faixa moderada (70 a 90%).

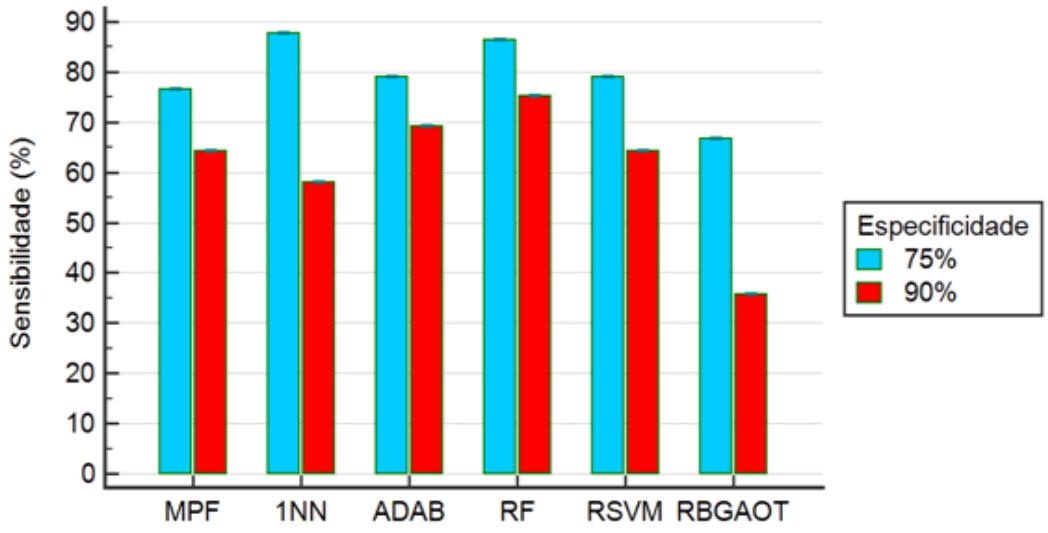

Figura 33 – Análise da sensibilidade com valores de especificidade em 75% e 90% no experimento com atributos do produto cruzado selecionado

A Figura 34 mostra um resumo com os classificadores que apresentaram melhor desempenho em cada um dos experimentos realizados, comparando-os com o melhor parâmetro da FOT. Durante os testes, os valores de AUC dos melhores algoritmos variaram entre 0,87 e 0,89 (acurácia moderada). Esses resultados mostram que o uso de algoritmos de aprendizado de máquinas teve melhor desempenho do que a classificação com a melhor característica da FOT, a reatância X_m.

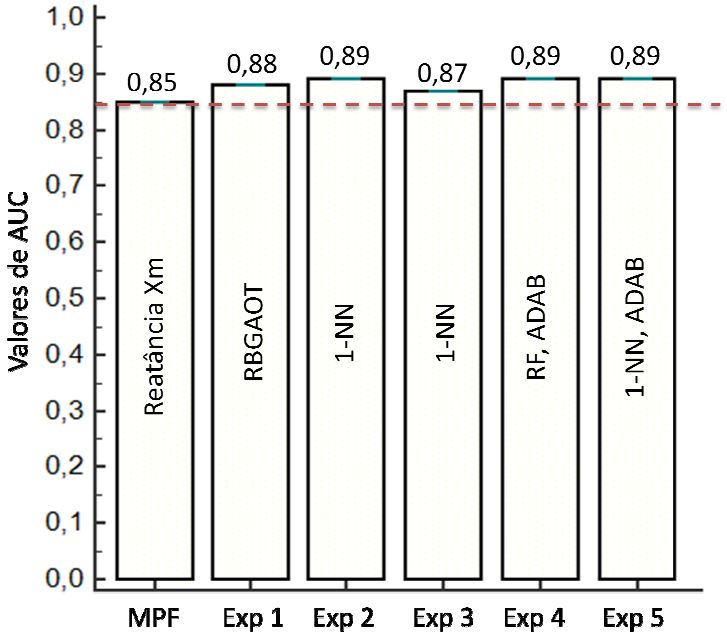

Figura 34 – Resumo dos maiores valores de AUC obtidos durante os experimentos

Os gráficos da Figura 35 e Figura 36 apresentam um resumo com o desempenho dos melhores algoritmos de cada experimento, usando como base a especificidade fixada em 75% e 90%, respectivamente. No primeiro caso, os valores de sensibilidade encontrados foram superiores, se comparados ao melhor parâmetro da FOT, variando de 83,95% a 85,19%. Já no segundo caso, onde é simulada uma situação mais restrita para o algoritmo, os valores de sensibilidade variaram entre 67,9% e 72,84%. Mesmo com a especificidade em 90%, simulando uma situação onde o algoritmo aceita apenas 10% de casos de falsos positivos, o uso de algoritmos de aprendizado de máquinas fez com que no terceiro e quarto experimentos com variáveis selecionadas, houvesse valores de sensibilidade dentro da faixa moderada.

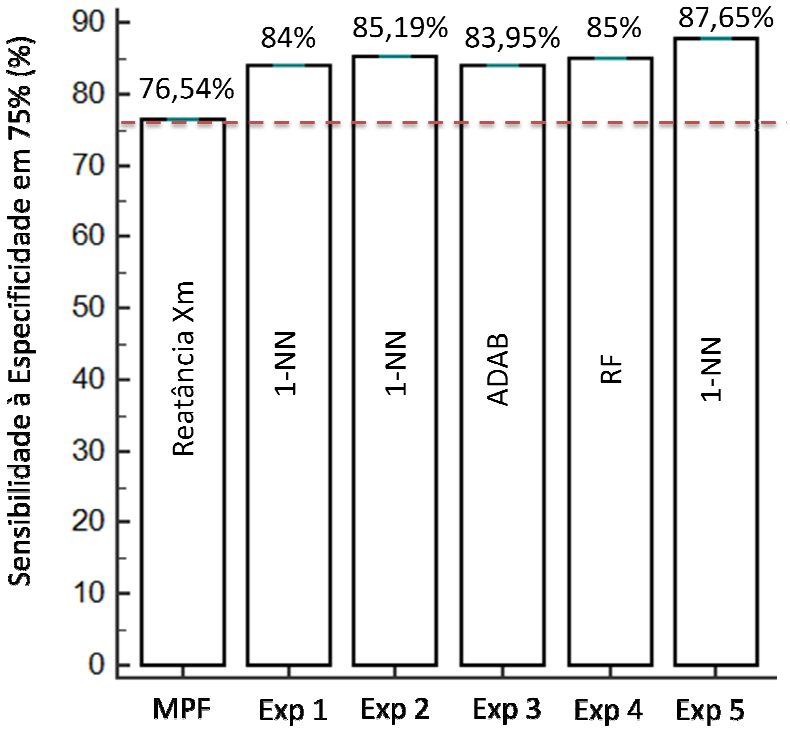

Figura 35 - Resumo dos maiores valores de sensibilidade com especificidade fixada em 75%, obtidos durante os experimentos

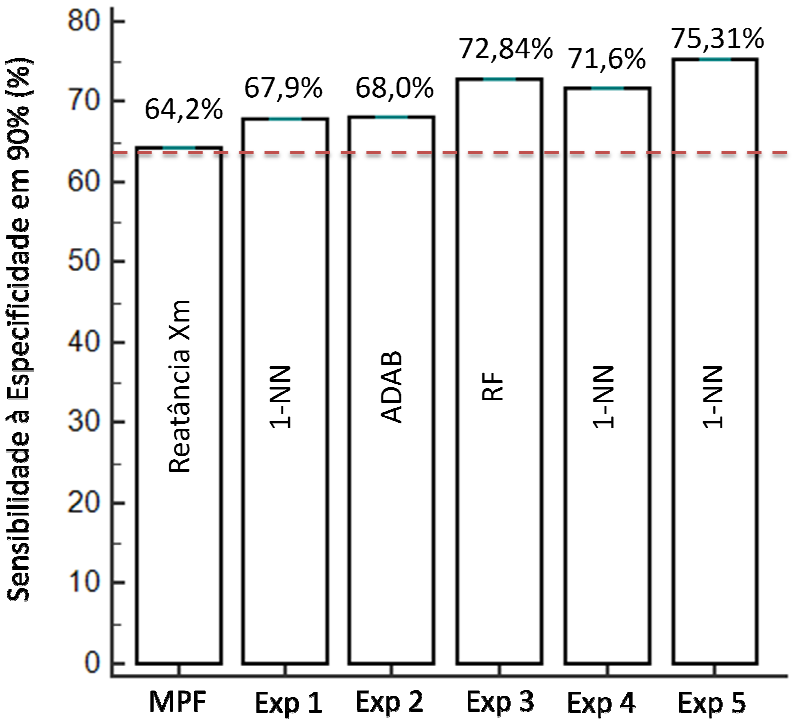

Figura 36 – Resumo dos maiores valores de sensibilidade com especificidade fixada em 90%, obtidos durante os experimentos

5.8. Inferência sobre Redes Bayesianas

As Redes Bayesianas fornecem grafos que mostram as ligações de dependência entre as variáveis do problema. Sendo assim, as redes construídas com base em matrizes, conforme item 4.6, que representam as melhores soluções geradas, são selecionadas pelo RBGAOT. Além de apresentar boa acurácia, essas estruturas devem permitir uma análise gráfica das relações existentes entre as características fornecidas pela FOT.

Algumas dessas redes foram selecionadas para análise com base na quantidade de ligações existentes entre suas variáveis. Logo, quanto menor o número de arcos existentes entre os nós da rede, mais simples será sua representação e, consequentemente, mais simples serão suas tabelas de distribuição de probabilidade conjunta (DPC). Nos itens 5.8.1 e 5.8.2 foram realizadas inferências sobre as redes selecionadas de acordo com o número de atributos de entrada usados em suas respectivas construções.

5.8.1. Rede com Oito Atributos

Com o intuito de analisar algumas das redes geradas pelo RBGAOT com todos os atributos da FOT, foi feita uma seleção com base em dois parâmetros. Primeiramente observou-se o menor número de ligações entre as variáveis, resultando em tabelas de DPC mais simples. Em seguida, foram selecionadas redes cujas tabelas de DPC apresentaram menor ocorrência de probabilidades iguais a 0,5, já que esse valor não permite inferir de forma mais precisa sobre uma situação. Como as tabelas de DPC são construídas de forma automática, elas produzem combinações que não condizem com a biomecânica. A fim de destacar apenas as combinações que descrevem situações possíveis de ocorrer, essas probabilidades foram destacadas em cada uma das tabelas.

A rede representada na Figura 37 tem estrutura composta pelos oito atributos fornecidos pela FOT: R_o, R_m, X_m, C_{din}, S, Z_{4Hz}, F_r e E_{din}, além da variável *classe*. Essa estrutura apresenta nove tabelas de DPC, cujas informações foram analisadas e comparadas com os gráficos da Figura 22.

A Tabela 26 mostra as probabilidades à priori da variável *classe*, o único nó raiz dessa rede. Por essas probabilidades, é possível observar que o conjunto de dados usado para a construção da estrutura é formado por 45% de indivíduos do grupo controle e 55% do grupo teste.

Tabela 26 – Probabilidades à priori da variável *classe* com oito atributos de entrada

P(*classe* = 0)	P(*classe* = 1)
0,45	0,55

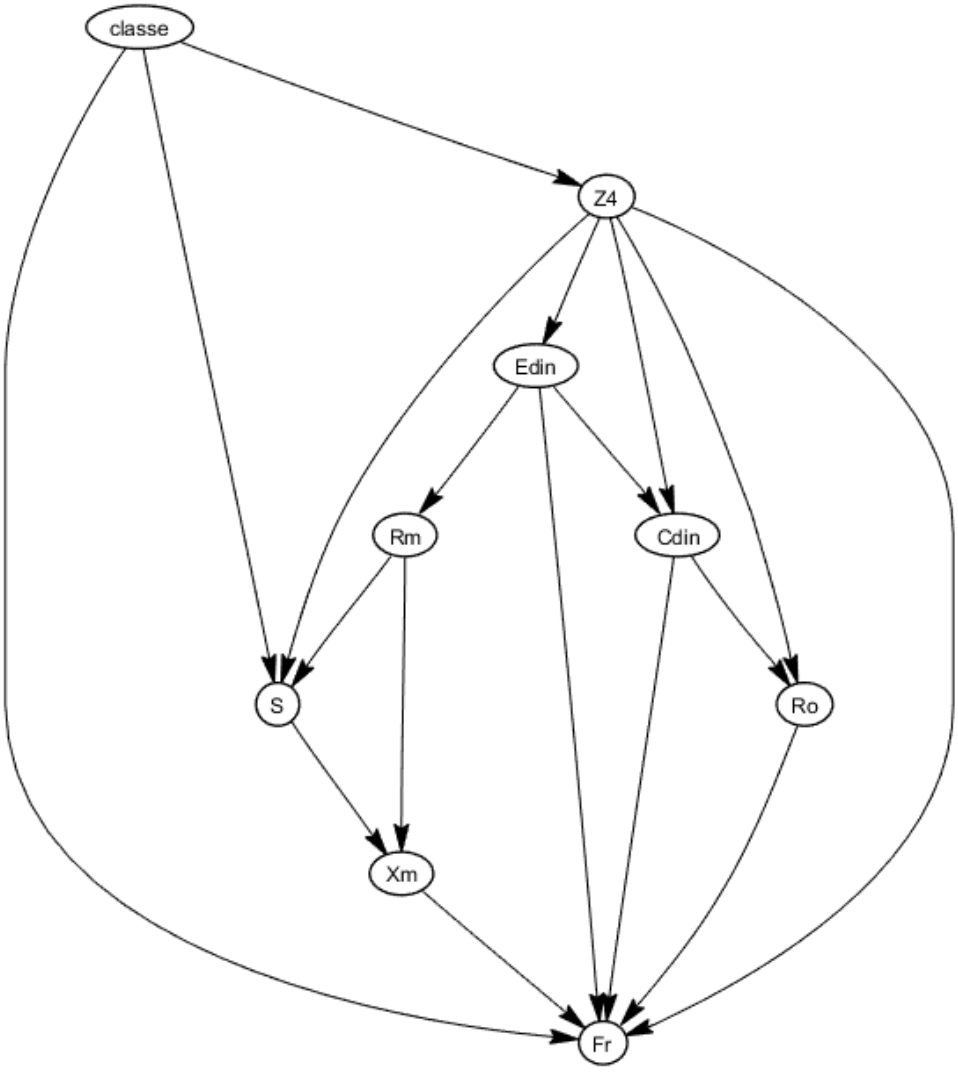

Figura 37 – Estrutura da rede com oito atributos de entrada

A variável Z_{4Hz} recebe influência apenas da variável *classe*. Pela Tabela 27, pode-se perceber que há alta probabilidade de um indivíduo ter baixa impedância (Z_{4Hz}=1) dado que não é portador da doença (*classe*=0). Já um indivíduo portador de fibrose cística (*classe*=1), tem maior probabilidade de apresentar alta impedância (Z_{4Hz} = 2). Comparando com o gráfico da Figura 22(f), em média, a impedância Z_{4Hz} é de fato menor para indivíduos que não possuem fibrose cística (grupo controle) e maior para indivíduos que possuem a doença (grupo teste).

Tabela 27 – DPC para a variável Z_{4Hz} da rede com oito atributos de entrada

	$P(Z_{4Hz} = 1 \mid classe)$	$P(Z_{4Hz} = 2 \mid classe)$
classe = 0	0,93	0,07
classe = 1	0,34	0,66

A variável R_m é influenciada apenas pela elastância (E_{din}) e, de acordo com a Tabela 28, possui comportamento diretamente proporcional a ela. Ao observar um indivíduo com baixa elastância, há probabilidade de 0,91 desse indivíduo também apresentar baixa resistência R_m.

Tabela 28 – DPC para a variável R_m da rede com oito atributos de entrada

	$P(R_m = 1 \mid E_{din})$	$P(R_m = 2 \mid E_{din})$
$E_{din} = 1$	0,91	0,09
$E_{din} = 2$	0,34	0,66

Pelos gráficos da Figura 22(b) e da Figura 22(h), tanto a resistência R_m como a elastância E_{din}, possuem valores mais baixos para indivíduos do grupo controle e valores mais altos para indivíduos do grupo teste. Logo, em média, ambos os parâmetros apresentam comportamento similar, conforme indicado pela Tabela 28.

A variável E_{din} também possui comportamento diretamente proporcional à impedância Z_{4Hz}. Pela Tabela 29 pode-se observar que no caso de Z_{4Hz}=1, a elastância mostra alta probabilidade de ser baixa (E_{din}=1). De forma similar, quando um indivíduo possui Z_{4Hz} =2, há alta probabilidade de ter E_{din}=2.

Tabela 29 – DPC para a variável E_{din} da rede com oito atributos de entrada

	$P(E_{din}=1 \mid Z_{4Hz})$	$P(E_{din}=2 \mid Z_{4Hz})$
$Z_{4Hz} = 1$	0,95	0,05
$Z_{4Hz} = 2$	0,14	0,86

Os gráficos da Figura 22(f) e da Figura 22(h) mostram que a elastância E_{din} e a impedância Z_{4Hz}, em média, também apresentam um comportamento equivalente, concordando assim, com as informações da Tabela 29.

A variável X_m recebe influências de outras duas variáveis: *S* e R_m. A reatância X_m mostra comportamento diretamente proporcional à inclinação da curva de resistência *S*. Analisando as condições a seguir retiradas da Tabela 30, é possível comprovar que há alta probabilidade de um indivíduo ter X_m com valor baixo, tendo observado *S* com valor baixo:

$$P(X_m = 1|S = 1, R_m = 1) = 0{,}76$$
$$P(X_m = 1|S = 1, R_m = 2) = 0{,}87$$

De forma similar, há maior probabilidade em observar alta reatância X_m, tendo observado uma maior inclinação *S* da curva de resistência, independentemente do valor de R_m:

$$P(X_m = 2|S = 2, R_m = 1) = 0{,}96$$
$$P(X_m = 2|S = 2, R_m = 2) = 0{,}55$$

Apresentando ao algoritmo a combinação improvável onde *S*=2 (característica do grupo controle) e R_m=2 (característica do grupo teste), observa-se a dificuldade do RBGAOT calcular um valor que favoreça uma discriminação mais clara, resultando em uma probabilidade próxima a 0,5, conforme a linha 4 da Tabela 30:

Tabela 30 – DPC para a variável X_m da rede com oito atributos de entrada

	$P(X_m = 1 \mid S, R_m)$	$P(X_m = 2 \mid S, R_m)$
$S = 1, R_m = 1$	0,76	0,24
$S = 2, R_m = 1$	0,05	0,95
$S = 1, R_m = 2$	0,87	0,13
$S = 2, R_m = 2$	0,45	0,55

As variáveis X_m e *S*, que possuem valores baixos, geralmente caracterizam indivíduos do grupo teste, conforme Figura 22(c) e Figura 22(e). Esse comportamento também é observado na Tabela 30.

As variáveis C_{din} e Z_{4Hz} exercem influência sobre a variável R_o, de acordo com a rede da Figura 37. Há maior probabilidade em observar baixa resistência R_o, dado que foram observados baixos valores de Z_{4Hz}, independente do valor de C_{din}. O mesmo comportamento é observado para altas resistências, conforme Tabela 31:

Tabela 31 – DPC para a variável R_o da rede com oito atributos de entrada

	$P(R_o = 1 \mid C_{din}, Z_{4Hz})$	$P(R_o = 2 \mid C_{din}, Z_{4Hz})$
$C_{din} = 1, Z_{4Hz} = 1$	0,64	0,36
$C_{din} = 2, Z_{4Hz} = 1$	0,97	0,03
$C_{din} = 1, Z_{4Hz} = 2$	0,19	0,81
$C_{din} = 2, Z_{4Hz} = 2$	0,34	0,66

Analisando as informações dos gráficos da Figura 22(a) e Figura 22(f), a impedância Z_{4Hz} e a resistência R_o de fato apresentam valores baixos para o grupo controle e valores altos para o grupo teste. Portanto, com base na biomecânica, as linhas 1 e 4 da Tabela 31 descrevem situações difíceis de ocorrer. Isso justifica valores de probabilidades mais próximos, mostrando maior indecisão.

A variável C_{din} é influenciada por E_{din} e Z_{4Hz}. Ao analisar a Tabela 32, obtém-se probabilidades de 0,98 e 0,81 do indivíduo possuir valor de complacência maior (C_{din}=2), dado que foi observada baixa elastância (E_{din}=1). O mesmo comportamento ocorre quando C_{din}=1. Pode-se concluir que de acordo com essa rede, C_{din} é inversamente proporcional a E_{din}.

No caso das linhas 2 e 3 na Tabela 32, há duas situações improváveis de ocorrer, já que a impedância é diretamente proporcional à elastância. Porém, apenas no caso onde E_{din} = 2 e Z_{4Hz} = 1, foi encontrada uma probabilidade menor, pois no caso onde E_{din} = 1 e Z_{4Hz} = 2, o valor da probabilidade continuou alto.

Tabela 32 – DPC para a variável C_{din} da rede com oito atributos de entrada

	$P(C_{din}=1\|E_{din}, Z_{4Hz})$	$P(C_{din}=2\|E_{din}, Z_{4Hz})$
E_{din} = 1, Z_{4Hz} = 1	0,02	0,98
E_{din} = 2, Z_{4Hz} = 1	0,65	0,35
E_{din} = 1, Z_{4Hz} = 2	0,19	0,81
E_{din} = 2, Z_{4Hz} = 2	0,97	0,03

Esse comportamento inversamente proporcional da complacência e da elastância pode ser observado também nos gráficos da Figura 22(d) e Figura 22(h), comprovando assim, as informações obtidas na Tabela 32.

A variável *S* é influenciada pelas variáveis R_m, Z_{4Hz} e *classe*. De acordo com as combinações a seguir, há maior probabilidade de um indivíduo ter inclinação da curva de resistência com alto valor (*S*=2) se ele não for portador da doença (*classe*=0):

$$P(S = 2|R_m = 1, Z_{4Hz} = 1, classe = 0) = 0{,}96$$
$$P(S = 2|R_m = 2, Z_{4Hz} = 1, classe = 0) = 0{,}72$$
$$P(S = 2|R_m = 1, Z_{4Hz} = 2, classe = 0) = 0{,}72$$

Mesmo em situações improváveis, representadas nas linhas 2 e 3, as informações da Tabela 33 concordam com o comportamento da inclinação da curva de resistência, descrito na Figura 22 (e), onde o valor alto de *S* é uma característica do grupo controle. Já um valor baixo de *S*, caracteriza indivíduos portadores de fibrose cística.

Dentre as combinações improváveis contidas na Tabela 33, há duas em que o algoritmo calculou probabilidades iguais a 0,5. A primeira é a P(*S* | R_m=2, Z_{4Hz}=2, *classe*=0), que descreve um indivíduo com valores altos de resistência e impedância, e mesmo assim não é portador da doença (*classe*=0). A segunda é P(*S* | R_m=2, Z_{4Hz} =1, *classe*=1), que mostra um indivíduo com baixa impedância e portador da doença (*classe*=1). Geralmente, isto se deve ao fato do algoritmo gerar todas as combinações possíveis, até mesmo situações como essas representadas nas linhas 4 e 6, onde o RBGAOT não foi capaz de inferir e calcular as probabilidades com valores distantes.

A linha cinco da Tabela 33 descreve uma situação onde o indivíduo possui baixos valores de R_m e Z_{4Hz}, mas é doente (*classe*=1). Mesmo nessa combinação improvável, o algoritmo calculou uma alta probabilidade do indivíduo ter valor de *S* elevado, característica de quem não é portador de fibrose cística. Essa alta probabilidade pode ter sido influenciada pelo fato do indivíduo apresentar uma combinação de duas características de um não portador da doença (R_m = 1 e Z_{4Hz} = 1), contra uma característica de quem é portador da fibrose cística (*classe*=1).

Tabela 33 – DPC para a variável *S* da rede com oito atributos de entrada

	P(*S* =1|R_m, Z_{4Hz}, *classe*)	P(*S* =2|R_m, Z_{4Hz}, *classe*)
R_m = 1, Z_{4Hz} = 1, *classe* = 0	0,04	0,96
R_m = 2, Z_{4Hz} = 1, *classe* = 0	0,28	0,72
R_m = 1, Z_{4Hz} = 2, *classe* = 0	0,28	0,72
R_m = 2, Z_{4Hz} = 2, *classe* = 0	0,50	0,50
R_m = 1, Z_{4Hz} = 1, *classe* = 1	0,03	0,97
R_m = 2, Z_{4Hz} = 1, *classe* = 1	0,50	0,50
R_m = 1, Z_{4Hz} = 2, *classe* = 1	0,54	0,46
R_m = 2, Z_{4Hz} = 2, *classe* = 1	0,77	0,23

A tabela de DPC da frequência F_r que, por ser nó folha dessa rede e ser influenciada por outras seis variáveis, exige uma tabela com 2^6 linhas para representar todas as possíveis combinações ligadas a essa variável. Devido à complexidade desse caso, o algoritmo forneceu apenas probabilidades iguais a 0,5.

Pela análise dessa rede com oito atributos, observou-se que a estrutura obtida pelo algoritmo RBGAOT apresentou informações probabilísticas das relações entre os parâmetros da FOT. Em geral, os valores mais altos de probabilidades encontrados nas tabelas de DPC, foram obtidos em combinações coerentes com a biomecânica. Essas informações foram conferidas com os valores médios da Figura 22, e apresentaram o mesmo comportamento descrito por esses gráficos, confirmando a consistência do RBGAOT.

5.8.2. Rede com Seleção de Cinco Atributos

Dentre as estruturas geradas com cinco atributos de entrada, uma rede foi selecionada para inferência. Essa escolha também foi feita com base no menor número de ligações entre as variáveis e pela menor ocorrência de probabilidades iguais a 0,5. Os cinco atributos usados na construção das redes foram: R_o, R_m, X_m, C_{din} e Z_{4Hz}, além da variável *classe*.

A rede da Figura 38 gerou um total de seis tabelas de DPC, que foram analisadas e comparadas com as informações da Figura 22. A Tabela 34 apresenta as probabilidades à priori da variável *classe*, onde a probabilidade de um indivíduo não ser portador da doença é igual a 0,49. Já a probabilidade do indivíduo ser portador da doença é igual a 0,51.

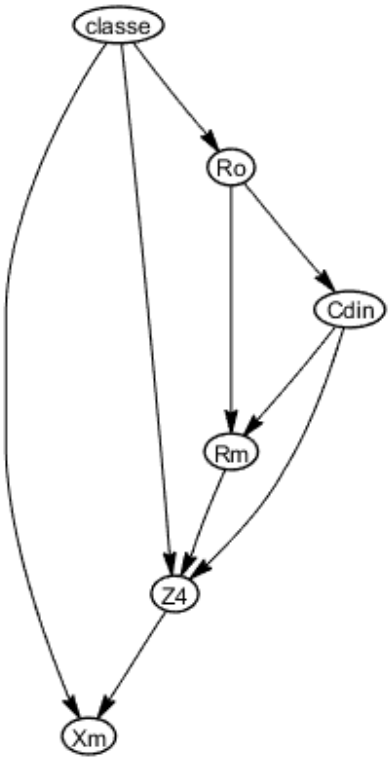

Figura 38 – Estrutura da rede com cinco atributos de entrada

Tabela 34 – Probabilidades à priori da variável *classe* da rede com cinco atributos de entrada

P(*classe* = 0)	P(*classe* = 1)
0,49	0,51

De acordo com a rede representada na Figura 38, a variável R_o é influenciada apenas pela variável *classe*. Sendo assim, a Tabela 35 mostra as probabilidades de um indivíduo ter baixa ou alta resistência (R_o = 1 ou 2), dado que foi observada a sua classe.

Tabela 35 – DPC para a variável R_o com cinco atributos de entrada

	P(R_o =1| *classe*)	P(R_o =2| *classe*)
classe = 0	0,94	0,06
classe = 1	0,47	0,53

Usando como base as condições, retiradas da Tabela 35, é possível concluir que há alta chance de um indivíduo apresentar baixo valor de R_o, dado que pertence a *classe* 0. Também há chances do indivíduo ter alto valor de R_o, dado que pertence a *classe* 1. Esse comportamento também é observado na Figura 22(a), onde a baixa resistência é característica do grupo controle, e a alta resistência é característica do grupo teste.

$$P(R_o = 1|classe = 0) = 0{,}94$$
$$P(R_o = 2|classe = 1) = 0{,}53$$

Pela Tabela 35, também foi observado que as probabilidades calculadas para *classe*=1, não possuem diferença grande entre seus valores, como ocorre na *classe*=0. Por apresentarem valores próximos a 0,5, o algoritmo mostra maior dificuldade em discriminar portadores da doença:

$$P(R_o = 1|classe = 1) = 0{,}47$$
$$P(R_o = 2|classe = 1) = 0{,}53$$

A variável C_{din} depende apenas da variável R_o. Pela Tabela 36 é possível observar que há uma probabilidade de 0,87 do indivíduo ter o valor de complacência alto (C_{din}=2), dado que foi observado um baixo valor de resistência (R_o=1). Já os indivíduos que apresentam maior valor de R_o, possuem probabilidade de 0,84 de ter complacência menor (C_{din}=1). Esse comportamento inversamente proporcional da complacência e da resistência, também é visto na Figura 22(a) e na Figura 22(d).

Tabela 36 – DPC para a variável C_{din} da rede com cinco atributos de entrada

	$P(C_{din}=1 \mid R_o)$	$P(C_{din}=2 \mid R_o)$
$R_o = 1$	0,13	0,87
$R_o = 2$	0,84	0,16

No caso da variável R_m, há duas outras variáveis que a influenciam: C_{din} e R_o. Pela Tabela 37, é possível analisar as relações a seguir e perceber que independente do valor de C_{din}, há probabilidade da resistência R_m apresentar um valor baixo, dado que também é observado um valor baixo de R_o:

$$P(R_m = 1|C_{din} = 1, R_o = 1) = 0{,}79$$
$$P(R_m = 1|C_{din} = 2, R_o = 1) = 0{,}98$$

O mesmo ocorre com a resistência R_m, dado que R_o apresenta um valor alto:

$$P(R_m = 2|C_{din} = 1, R_o = 2) = 0{,}82$$
$$P(R_m = 2|C_{din} = 2, R_o = 2) = 0{,}77$$

A relação diretamente proporcional da resistência média (R_m) e da resistência no intercepto (R_o) descrita na Tabela 37, concordam com os gráficos da Figura 22(a) e Figura 22(b). Essa relação se mantém mesmo nas inconsistências biomecânicas C_{din}=1 e R_m=1, ou C_{din}=2 e R_o=2, como ocorre respectivamente nas linhas 1 e 4.

Tabela 37 – DPC para a variável R_m da rede com cinco atributos de entrada

	$P(R_m=1 \mid C_{din}, R_o)$	$P(R_m=2 \mid C_{din}, R_o)$
$C_{din} = 1, R_o = 1$	0,79	0,21
$C_{din} = 2, R_o = 1$	0,98	0,02
$C_{din} = 1, R_o = 2$	0,18	0,82
$C_{din} = 2, R_o = 2$	0,23	0,77

A variável X_m é influenciada pelas variáveis Z_{4Hz} e *classe*. Conforme as probabilidades da Tabela 38, há alta chance de um indivíduo apresentar alta reatância X_m, dado que Z_{4Hz} é baixo, independente do valor da variável *classe*. Esse comportamento é observado mesmo na situação improvável, onde há baixa impedância Z_{4Hz} em um portador de fibrose cística:

$$P(X_m = 2|Z_{4\text{Hz}} = 1, classe = 0) = 0{,}98$$
$$P(X_m = 2|Z_{4\text{Hz}} = 1, classe = 1) = 0{,}94$$

Do mesmo modo, há probabilidade do indivíduo apresentar baixa reatância X_m, dado que foi observado alto valor de impedância Z_{4Hz}. Mesmo com a situação improvável de um indivíduo não possuir a doença (*classe*=0) e ter alta impedância (Z_{4Hz}=2), esse comportamento se manteve, apresentando apenas, valor de probabilidade mais baixo:

$$P(X_m = 1|Z_{4\text{Hz}} = 2, classe = 0) = 0{,}64$$
$$P(X_m = 1|Z_{4\text{Hz}} = 2, classe = 1) = 0{,}80$$

Os gráficos da Figura 22 (c) e da Figura 22 (f), comprovam o comportamento inversamente proporcional da reatância X_m e da impedância Z_{4Hz}, descrito na Tabela 38.

Tabela 38 – DPC para a variável X_m da rede com cinco atributos de entrada

	$P(X_m=1 \mid Z_{4Hz}, classe)$	$P(X_m=2 \mid Z_{4Hz}, classe)$
$Z_{4Hz} = 1$, *classe* = 0	0,02	0,98
$Z_{4Hz} = 2$, *classe* = 0	0,64	0,36
$Z_{4Hz} = 1$, *classe* = 1	0,06	0,94
$Z_{4Hz} = 2$, *classe* = 1	0,80	0,20

Já a variável Z_{4Hz} é influenciada por três outras variáveis: R_m, C_{din} e *classe*. As condições a seguir, retiradas da Tabela 39, mostram que há maior probabilidade de um indivíduo apresentar alta impedância Z_{4Hz}, dado que ele é doente. Pela Figura 22(f), observa-se que de fato indivíduos portadores da doença (*classe*=1) possuem valores altos de impedância.

$$P(Z_{4\text{Hz}} = 2|R_m = 1, C_{din} = 1, classe = 1) = 0{,}77$$
$$P(Z_{4\text{Hz}} = 2|R_m = 2, C_{din} = 1, classe = 1) = 0{,}97$$
$$P(Z_{4\text{Hz}} = 2|R_m = 2, C_{din} = 2, classe = 1) = 0{,}81$$

Do ponto de vista biomecânico, as únicas situações que descrevem combinações prováveis são as linhas 3 e 6. Para esses casos, foram encontrados altos valores de probabilidade:

$$P(Z_{4\text{Hz}} = 1|R_m = 1, C_{din} = 2, classe = 0) = 0{,}99$$
$$P(Z_{4\text{Hz}} = 2|R_m = 2, C_{din} = 1, classe = 1) = 0{,}97$$

Tabela 39 – DPC para a variável Z_{4Hz} da rede com cinco atributos de entrada

	$P(Z_{4Hz}=1 \mid R_m, C_{din}, classe)$	$P(Z_{4Hz}=2 \mid R_m, C_{din}, classe)$
$R_m = 1$, $C_{din} = 1$, *classe* = 0	0,28	0,72
$R_m = 2$, $C_{din} = 1$, *classe* = 0	0,50	0,50
$R_m = 1$, $C_{din} = 2$, *classe* = 0	0,99	0,01
$R_m = 2$, $C_{din} = 2$, *classe* = 0	0,72	0,28
$R_m = 1$, $C_{din} = 1$, *classe* = 1	0,23	0,77
$R_m = 2$, $C_{din} = 1$, *classe* = 1	0,03	0,97
$R_m = 1$, $C_{din} = 2$, *classe* = 1	0,92	0,08
$R_m = 2$, $C_{din} = 2$, *classe* = 1	0,19	0,81

Para o caso de um indivíduo com as características: R_m = 2, C_{din} = 1 e *classe* = 0, o algoritmo não conseguiu definir os cálculos das probabilidades, resultando em $P(Z_{4Hz}|R_m=2,C_{din}=1,classe=0)$ igual a 0,5. Normalmente, isto ocorre em situações improváveis como essa, em que um indivíduo possui alta resistência respiratória (R_m=2) e complacência próxima à zero (C_{din}=1), porém não possui a doença (*classe*=0).

Assim como no item 5.8.1, a estrutura gerada pelo algoritmo RBGAOT apresentou probabilidades que descrevem as relações existentes entre os parâmetros da FOT. Nas combinações que representam situações coerentes com a biomecânica, também foram obtidos os maiores valores de probabilidades. Essas relações estão de acordo com os gráficos da Figura 22 e podem fornecer informações mesmo quando são submetidas a combinações improváveis, porém com menores valores de probabilidades. A inferência sobre outras redes pode ser lida no Apêndice B.

CONCLUSÃO

Este projeto seguiu a linha de pesquisa de trabalhos já realizados (AMARAL et al., 2013; AMARAL et al., 2015; AMARAL et al., 2017), com o uso dos dados fornecidos pela FOT em algoritmos de aprendizado de máquinas, mostrando que essa associação também foi eficiente na detecção de alterações respiratórias na fibrose cística. Durante os experimentos, os classificadores: 1-*Nearest Neighbor*, *Adaboost*, *Radial Support Vector Machine*, *Random Forest* e Redes Bayesianas, apresentaram valores de AUC maiores do que os valores obtidos pelo melhor parâmetro da FOT, indicando assim, maior acurácia no diagnóstico.

Dentre os testes realizados, a reatância respiratória (X_m) foi o atributo que apresentou melhor desempenho individual (AUC=0,85). No primeiro experimento, foram usados oito atributos de entrada fornecidos pela FOT, sendo o RBGAOT o algoritmo que apresentou melhor resultado (AUC=0,88).

No experimento seguinte, o produto cruzado dos oito atributos da FOT foi usado com o intuito de melhorar o desempenho dos algoritmos, fornecendo um conjunto de dados em uma dimensão mais alta. Foram geradas 36 combinações, que somadas a variável classe, totalizaram 37 atributos de entrada nos algoritmos testados. Como resultado, o 1-NN teve melhor desempenho com AUC igual a 0,89. Já o RBGAOT não convergiu nesse teste, devido aos cálculos realizados durante a marginalização da rede realizados pelo algoritmo *Junction Tree*. Essa limitação também pode ser observada em outros trabalhos com Redes Bayesianas, como no artigo (SILANDER et al., 2012), onde o número máximo de variáveis suportadas pelo modelo é 30.

No terceiro experimento foi feita a seleção de atributos de entrada com base na acurácia que, além de coincidir com a seleção realizada por um especialista, foi a técnica de seleção que indicou variáveis com melhor desempenho. Ao todo, cinco variáveis foram selecionadas: R_o, R_m, X_m, C_{din} e Z_{4Hz}, sendo o 1-NN o algoritmo que apresentou melhor desempenho, com AUC igual a 0,87. Já o algoritmo RBGAOT obteve seu desempenho mais baixo, com AUC igual a 0,79.

Durante o quarto experimento, foi aplicado o produto cruzado nos cinco atributos selecionados no teste anterior, gerando um total de 15 combinações que, somada a variável classe, totalizaram 16 atributos na entrada dos classificadores. O ADAB e RF foram os algoritmos que tiveram melhores resultados com valores de AUC iguais a 0,89.

No quinto experimento, a seleção de atributos foi feita dentre as 36 combinações resultantes da aplicação do método do produto cruzado nos oito atributos fornecidos pela FOT. Os classificadores que apresentaram melhor desempenho foram 1-NN e ADAB, com valores de AUC iguais a 0,89, seguidos do RF e RSVM, com valores de AUC iguais a 0,88.

Em todos os experimentos, pelo menos um algoritmo de aprendizado de máquina apresentou sensibilidade acima de 80%, ao observar uma especificidade fixada em 75%. Em uma situação mais restrita para o algoritmo, com especificidade fixada em 90%, pelo menos dois algoritmos alcançaram a faixa da sensibilidade moderada (70 a 90%) nos experimentos com seleção de atributos. Esse é um bom resultado já que essa análise representa uma situação onde o algoritmo é limitado a 10% de falsos positivos. Em ambos os casos, os resultados obtidos superaram a sensibilidade obtida pelo melhor parâmetro da FOT.

Além da acurácia no diagnóstico, a interpretabilidade também foi analisada pelas Redes Bayesianas, construídas e selecionadas por Algoritmos Genéticos. Mesmo sendo treinada com um conjunto de dados limitado, essa técnica se mostrou eficiente, apresentando probabilidades condicionais capazes de descrever o comportamento das características do sistema respiratório de um indivíduo portador de fibrose cística. Vale ressaltar que as redes geradas com cinco atributos da FOT (terceiro experimento), apresentaram maior facilidade em sua inferência, porém menor valor de AUC (AUC=0,79). Já as redes geradas com oito atributos (primeiro experimento), apresentaram melhor desempenho (AUC=0,88), porém possuem uma análise mais difícil devido à maior quantidade de variáveis e ligações entre elas.

O presente trabalho mostrou que o uso de Redes Bayesianas fornece interpretabilidade ao resultado obtido, mostrando as relações existentes entre as variáveis que descrevem a biomecânica do sistema respiratório. Por meio das estruturas geradas, é possível quantificar e compreender melhor como essas variáveis se relacionam, mantendo ainda boa acurácia na detecção de alterações respiratórias em portadores de fibrose cística. Sendo assim, novas informações são geradas e, somadas aos métodos atuais, podem ser usadas para auxílio da equipe médica no estudo da doença.

Uma das limitações observadas no algoritmo RBGAOT foi sua sensibilidade durante o experimento com 37 atributos de entrada, fazendo com que ele não convergisse. Uma melhoria proposta para esse problema é o uso de outro método que realize a marginalização da rede e tenha menor custo computacional. Outra limitação está no descarte das estruturas que não sejam DAG ou que não possuam a variável classe. Uma possível solução para estes casos é a elaboração de uma rotina que realize o reparo dessas redes geradas, transformando-

as em estruturas válidas para o problema. Outras metaheurísticas também podem ser testadas para criação e seleção de estruturas de Redes Bayesianas, além do algoritmo genético.

Este trabalho foi desenvolvido em ambiente Matlab por se tratar de um protótipo. Entretanto, como melhoria futura, será feita a implementação do algoritmo RBGAOT no *software* Python, devido à facilidade em usar um ambiente aberto e gratuito. Há outras duas propostas que visam melhorar a forma de analisar o resultado obtido pelas Redes Bayesianas. Uma delas é a disponibilização de uma plataforma na internet, onde outros pesquisadores possam obter modelos ao inserir seus dados. Outra proposta é o desenvolvimento de um algoritmo que forneça os valores de probabilidade obtidos pela inferência diagnóstica nas redes de forma automática.

REFERÊNCIAS

AMARAL JLM, LOPES AJ, FARIA ACD, MELO PL. *Machine learning algorithms and forced oscillation measurements to categorise the airway obstruction severity in chronic obstructive pulmonary disease*, *Computer Methods and Programs in Biomedicine*, Elsevier 118, p. 186-197, 2015;

AMARAL JLM, LOPES AJ, JANSEN JM, FARIA ACD, MELO PL. *An improved method of early diagnosis of smoking-induced respiratory changes using machine learning algorithms*, *Computer Methods and Programs in Biomedicine*, Elsevier 112, p. 441-454, 2013;

AMARAL JLM, LOPES AJ, VEIGA J, FARIA ACD, MELO PL. *High-accuracy detection of airway obstruction in asthma using machine learning algorithms and forced oscillation measurements*, *Computer Methods and Programs in Biomedicine*, Elsevier 144, p. 113–125, 2017;

ANDERSEN DH, *Cystic fibrosis of the pancreas and its relation to celiac disease clinical and pathologic study*, *American Journal of Diseases of Children* 56(2):344-399, 1938;

ARA-SOUZA AL. Redes Bayesianas: uma introdução aplicada a *Credit Scoring*, Centro de Ciências Exatas e Tecnológicas – Universidade Federal de São Carlos, 2010;
BARBER D. *Probabilistic Modelling and Reasoning The Junction Tree Algorithm*, Universidade de Edinburgh, 2003;

BRATKO I. *Machine Learning: Between Accuracy and Interpretability*, *Learning Networks and Statistics, International Centre for Mechanical Sciences*, Vol. 382, p. 164-177, Springer, Vienna, 1997;

BREIMAN L. *Random Forests*, *Kluwer Academic Publishers*, *Machine Learning*, v.45, p. 5–32, 2001;

BROWN LE; TSAMARDINOS I; ALIFERIS CF. *A Novel Algorithm for Scalable and Accurate Bayesian Network Learning*; Departamento de Informática Biomédica da Universidade de Vanderbilt, 2004;

CARVALHO MA. Discretização de Atributos Contínuos em Sistemas de Informação Utilizando Algoritmos Genéticos para a Aplicação da Teoria dos Conjuntos Aproximados, Universidade Federal de Itajubá, 2010;

CASTELLANI C., CUPPENS H., MACEK M.J., CASSIMAN J.J., KEREM E., DURIE P., TULLIS E., ASSAEL B.M., BOMBIERI C., BROWN A., CASALS T., CLAUSTRES M., CUTTING G.R., DEQUEKER E., DODGE J., DOULL I., FARRELL P., FEREC C., GIRODON E., JOHANNESSOM M., KEREM B., KNOWLES M., MUNCK A., PIGNATTI P.F., RADOJKOVIC D., RIZZOTTI P., SCHWARZ M., STUHRMANN M., TZETIS M., ZIELENSKI J., ELBORN J.S; Consensus on the use and interpretation of cystic fibrosis mutation analysis in clinical practice; *Journal of Cystic Fibrosis* 7, p.179–196; 2008;

CASTILLO E; GUTIÉRREZ JM; ALI SH. *Sistemas expertos y modelos de redes probabilísticas*, *Academia Española de Ingeniería*, Madrid, 1996;

CHAVES BB. Estudo do Algoritmo Adaboost de Aprendizagem de Máquina Aplicado a Sensores e Sistemas Embarcados, Escola Politécnica da Universidade de São Paulo, 2012;

CIVICIOGLU P. *Backtracking Search Optimization Algorithm for numerical optimization problems*, *Applied Mathematics and Computation*, Vol. 219, p. 8121–8144, Elsevier *Science*, 2013;

COLLINS M; SCHAPIRE RE; SINGER Y. *Logistic Regression, AdaBoost and Bregman Distances*, *Machine Learning*, v. 48, p. 253-285, *Kluwer Academic Publishers*, 2002;

COOPER G; HERSKOVITS E. *A Bayesian method for the induction of probabilistic networks from data*, *Technical Report* SMI-91-1, *Section on Medical Informatics*, Universidade de Stanford, 1991;

COSTA HSRM. Estudo comparativo de abordagens ao problema de débito de transações bancárias em contas com saldo insuficiente, Departamento de Matemática Aplicada Faculdade de Ciências da Universidade do Porto, 2012;

Cystic Fibrosis Foundation Patient Registry, *Annual Data Report*, Bethesda, Maryland, 2014;

DALCIN PLT, SILVA FAA; Fibrose cística no adulto: aspectos diagnósticos e terapêuticos; Jornal Brasileiro de Pneumologia, p. 107-117; 2008;

DUBOIS AB, BRODY AW, LEWIS DH, BURGESS BF. *Oscillation mechanics of lungs and chest in man*, *Journal of applied physiology*, 8:587-594, 1956;

DUIN RPW; JUSZCZAK P; PACLIK P; PEKALSKA E; RIDDER DMJ; TAX DMJ; VERZAKOV S. PRTools 4.1, *A Matlab Toolbox for Pattern Recognition*, Universidade de Tecnologia Delft, Holland, 2007;

FACELI K; LORENA AC; GAMA J; CARVALHO ACPLF. Inteligência Artificial: uma Abordagem de Aprendizado de Máquina, LTC, 2011;

FARRELL PM, WHITE TB, REN CL, HEMPSTEAD SE, ACCURSO F, DERICHS N, HOWENSTINE M, MCCOLLEY SA, ROCK M, ROSENFELD M, SERMET-GAUDELUS I, SOUTHERN KW, MARSHALL BC, SOSNAY PR, *Diagnosis of Cystic Fibrosis: Consensus Guidelines from the Cystic Fibrosis Foundation*, *The Journal of Pediatrics* Volume 181S, 2017;

FAWCETT T. *An Introduction to ROC Analysis*, *Pattern Recognition Letters*, V. 27, N. 8, p. 861–874, 2006;

FAYYAD UM; IRANI KB. *Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning*, Machine Learning, p. 1022-1027, 1993;

GABRIEL PHR; DELBEM ACB. Fundamentos de Algoritmos Evolutivos, Notas Didáticas do ICMC-USP, N. 75, p. 35, 2008;

GACTO MJ; ALCALÁ R; HERRERA F. *Interpretability of linguistic fuzzy rule-based systems: An overview of interpretability measures*, *Information Sciences*, Elsevier, V. 181, p. 4340–4360, DOI:10.1016, 2011;

GONÇALVES AR. Redes Bayesianas; Universidade de Campinas; Disponível em: <http://www-users.cs.umn.edu/~andre/arquivos/pdfs/bayesianas.pdf> Acessado em: 03/01/2018;

GUYON I; ELISSEEFF A. *An Introduction to Variable and Feature Selection*, *Journal of Machine Learning Research*, Vol 3, 1157-1182, 2003;

HANLEY JA; MCNEIL BJ, *The Meaning and Use of the Area under a Receiver Operating Characteristic*; *Diagnostic Radiology*; Vol. 143, N. 1; 1982;

HASTIE T; TIBSHIRANI R; FRIEDMAN J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, 2008;

HERSKOVITS E; COOPER G. *Kutató: An Entropy-Driven System for Construction of Probabilistic Expert Systems from Databases*, Knowledge Systems Laboratory, Medical Computer Science - Stanford University, 1991;

HORTA RAM; CARVALHO FA; ALVES FJS; JORGE MJ. Comparação de Técnicas de Seleção de Atributos para Previsão de Insolvência de Empresas Brasileiras no Período 2005-2007. 34º Encontro da ANPAD, 2010;

HOUCK CR; JOINES JA; KAY MG. *A Genetic Algorithm for Function Optimization: A Matlab Implementation*, Universidade do Estado da Carolina do Norte, NCSU-IE, TR 95–09, 1995;

HUANG J; LING CX. *Using AUC and Accuracy in Evaluating Learning Algorithms*, IEEE *Transaction Knowledge and Data Engineering*, Vol. 17, N. 3, p. 299–310, 2005;

KORB KB; NICHOLSON AE. *Introducing Bayesian Network*, *Bayesian Artificial Intelligence*, 2ª Edição, Cap. 2, p. 29-54, CRC Press, 2010;

KUNCHEVA LI. *Combining Pattern Classifiers: Methods and Algorithms*, Wiley Interscience, New Jersey, 2014;

LACERDA EGM; CARVALHO ACPLF. Introdução aos Algoritmos Genéticos, Sistemas Inteligentes: Aplicações a Recursos Hídricos e Ciências Ambientais, Capítulo 3, p. 87-148, Editora da Universidade Federal do Rio Grande do Sul, 1999;

LACERDA LS; LOPES AJ; CARVALHO ARS; GUIMARÃES ARM; FIRMIDA MC; CASTRO MCS; MOGAMI R; MELO PL. The Role of Multidetector Computed Tomography and the Forced Oscillation Technique in Assessing Lung Damage in Adults With Cystic Fibrosis, Respiratory Care, Vol 63 Issue 3, PubMed: 29208759, 2017;

LARRAÑAGA P; POZA M; YURRAMENDI Y; MURGA RH; KUIJPERS CMH. *Structure Learning of Bayesian Networks by Genetic Algorithms: Performance Analysis of Control Parameters*; IEEE *Transactions on pattern analysis and machine intelligence*, Vol. 18, N. 9, p. 912-926, 1996;

LIAW A; WIENER M. *Classification and Regression by Random Forest*; *R. News*, Vol. 2/3, p.18–22, 2002;
LIMA AN, FARIA ACD, LOPES AJ et al. Técnica de oscilações forcadas na avaliação funcional de pacientes com fibrose cística com idade superior a 18 anos, Pulmão RJ 2010;

LIMA AN, FARIA CDF, LOPES AJ, JANSEN JM, MELO PL. *Forced oscillations and respiratory system modeling in adults with cystic fibrosis*, *BioMedical Engineering OnLine*, DOI 10.1186/s12938-015-0007-7, 2015;

LORENA AC; CARVALHO ACPLF. Uma Introdução às *Support Vector Machines*, Revista de Informática Teórica e Aplicada, Volume 14 - Número 2, 2007;

MACLEOD D, BIRCH M. *Respiratory input impedance measurement: forced oscillation methods*, *Medical & Biological Engineering & Computing*, Vol 39 p. 505-516, 2001;

MANDAL S; SINHA RK; MITTAL K. *Comparative Analysis of Backtrack Search Optimization Algorithm with other Evolutionary Algorithms for Global Continuous Optimization,* International Journal of Computer Science and Information Technologies, V. 6, N. 3, p. 3237-3241, 2015;

MARGINEANTU DD; DIETTERICH TG. *Pruning Adaptive Boosting, Machine Learning: Proceedings of the Fourteenth International Conference*, p. 211-218, 1997;

MARINHO CL; MAIOLI MCP; AMARAL JLM; LOPES AJ; MELO PL. *Respiratory resistance and reactance in adults with sickle cell anemia: Correlation with functional exercise capacity and diagnostic use*, PLoS One 12 (12): e0187833, 2017;

MARQUES RL; DUTRA I. Redes Bayesianas: o que são, para que servem, algoritmos e exemplos de aplicações; Coppe Sistemas – UFRJ, 2003;

MELO PL, GIANELLA-NETO WM A. Avaliação da mecânica ventilatória por oscilações forçadas: Fundamentos e aplicações clínicas. Jornal brasileiro de pneumologia: publicação oficial da Sociedade Brasileira de Pneumologia e Tisilogia; 26:194-206; 2000;

MELO PL. Técnica de oscilações forçadas na prática pneumológica: Princípios e exemplos de potenciais aplicações, Pulmão RJ, Vol 24 p. 42-48, 2015;

MENSXMACHINA, *Toolbox Probabilistic Graphical Model* 9.2.3, Universidade de Creta, Departamento de Ciência da Computação, Campus Voutes, 2011, Disponível em: <http://mensxmachina.org/en/software/pgm-toolbox/>. Acessado em: 08/02/2018;

MERSCHMANN LHC. Classificação Probabilística Baseada em Análise de Padrões, Universidade Federal Fluminense, 2007;

METZ CE. *Basic Principles of ROC Analysis*, *Seminars in Nuclear Medicine,* Vol. 8, N. 4, 1978;

MIRANDA IA, FARIA ACD, LOPES AJ, JANSEN JM, MELO PL. *On the Respiratory Mechanics Measured by Forced Oscillation Technique in Patients with Systemic Sclerosis*; Plos One, Vol. 8(4): e61657, doi:10.1371/journal.pone.0061657, 2013;

MITCHELL TM. *Machine Learning*, McGraw-Hill; 1997;
MOTA LR, SOUZA EL, ROCHA PHSA, FONSECA MJ, SANTOS JF, LAGE VMGB, LIMA RLLF. Estudos genéticos sobre a Fibrose Cística no Brasil: uma revisão sistemática, Revista de Ciências Médicas e Biológicas, 2015;

MURUZÁBAL J; COTTA C. *A Study on the Evolution of Bayesian Network Graph Structures, Advances in Probabilistic Graphical Models*, p. 193-213, 2007;

MYERS J, LASKEY K, DEJONG K. *Learning Bayesian Networks from Incomplete Data using Evolutionary Algorithms*, 1st *Annual Conference on Genetic and Evolutionary Computation*, V. 1, p. 458-465, 1999;

NEAPOLITAN, RE. *Learning Bayesian Networks*, *Prentice Hall Series in Artificial Intelligence*, *Northeastern Illinois University*, 2003;

ONISKO A; DRUZDZEL MJ; WASYLUK H. *Learning Bayesian Network Parameters from Small Data Sets: Application of Noisy-OR Gates*, *International Journal of Approximate Reasoning*, Elsevier, Vol. 27, p. 165-182, 2001;

PINHO AF; MONTEVECHI JAB; MARINS FAS; MIRANDA RC. Algoritmos Genéticos: Fundamentos e Aplicações, Meta-Heurísticas em Pesquisa Operacional, Capítulo 2, p. 21-32, DOI: 10.7436/2013.mhpo.02, 2013;

PGM – *Probabilistic Graphical Model toolbox*, Mens x Machina, Departamento da Ciência da Computação, Universidade de Creta, Grécia, Disponível em: <http://mensxmachina.org/en/software/pgm-toolbox> Acessado em: 10/02/2018;

PIFER AC. Estudo comparativo de métricas de pontuação para aprendizagem estrutural de Redes Bayesianas, Centro de Tecnologia – Programa de Pós-Graduação em Engenharia Elétrica, Universidade Federal do Rio Grande do Norte, 2006;

REIS FJC, DAMACENO N. Fibrose Cística, Jornal de Pediatria, Sociedade Brasileira de Pediatria Vol 74 Supl 1, 1998;

RIBEIRO FCV; LOPES AJ; MELO PL. *Reference values for respiratory impedance measured by the Forced Oscillation Technique in adult men and women*, The Clinical Respiratory Journal, DOI: 10.1111/crj.12783, PMID: 29470844, 2018;

RIBEIRO JD, RIBEIRO MAGO, RIBEIRO AF, Controvérsias na fibrose cística – do pediatra ao especialista, Jornal de Pediatria Vol 78 Supl 2, 2002;

RODRIGUES YE; MANICA E; ZIMMER ER; PASCOAL TA; MATHOTAARACHCHI SS; ROSA-NETO P. *Wrappers Feature Selection in Alzheimer's Biomarkers Using kNN and*

SMOTE Oversampling. Sociedade Brasileira de Matemática Aplicada e Computacional, Tendências em Matemática Aplicada e Computacional, Vol. 18, N. 1, p. 15-34, doi: 10.5540, tema 2017.018.01.0015, 2017;

ROSA TO; LUZ HS. Conceitos Básicos de Algoritmos Genéticos: Teoria e Prática, Anais do XI Encontro de Estudantes de Informática do Tocantins, p. 27-37, 2009;

SANTANA AL; REGO LP; FRANCÊS CRL; CARVALHO SV; VIJAYKUMAR NL. Aplicação de Modelos Markovianos para a Análise Temporal e Melhoria da Interpretabilidade de Redes Bayesianas, 39º SBPO – A pesquisa Operacional e o Desenvolvimento, p. 456 -465, 2007;

SCHAPIRE RE. *Explaining AdaBoost*, *Empirical Inference*, p. 37–52, *Springer*, 2013;

SILANDER T; MYLLYMÄKI P. A Simple Approach for Finding the Globally Optimal Bayesian Network Structure, *Proceedings of the Twenty-second Annual, Conference on Uncertainty in Artificial Intelligence,* 2012;

SILVA WT; LADEIRA M. Mineração de Dados em Redes Bayesianas, Universidade de Brasília, Capítulo 6, 2016;

SMOLA A; VISHWANATHAN SVN. *Introduction to Machine Learning*, Cambridge *University Press*, 2008;

TONDA A; LUTTON E; REUOLLION R; SQUILLERO G; WUILLEMIN PH. *Bayesian Network Structure Learning from Limited Datasets through Graph Evolution*, 15º *European Conference on Genetic Programming*, EuroGP 2012, Malaga – Spain, 7244, p. 254-265, 2012;

ZHU J; ZOU H; SAHARON R; HASTIE T. *Multi-class AdaBoost*, *Statistics and Its Interface*, v. 2, p. 349–360, 2009;

APÊNDICE A – Combinações do Produto Cruzado

A Tabela 40 mostra as 36 combinações obtidas ao aplicar o método do produto cruzado nos oito parâmetros fornecidos pela FOT. Essas combinações foram usadas como atributos de entrada nos experimentos realizados nos itens 5.4 e 5.7. Já a Tabela 41 mostra as 15 combinações geradas a partir dos cinco atributos da FOT selecionados durante o experimento do item 5.6.

Tabela 40 – 36 Combinações geradas pelo produto cruzado no experimento dos itens 5.4 e 5.7 com seleção de cinco atributos

	$\mathbf{F_r}$	$\mathbf{X_m}$	$\mathbf{R_o}$	**S**	$\mathbf{R_m}$	$\mathbf{C_{din}}$	$\mathbf{E_{din}}$	$\mathbf{Z_{4Hz}}$
$\mathbf{F_r}$	$F_r.F_r$	$F_r.X_m$	$F_r.R_o$	$F_r.S$	$F_r.R_m$	$F_r.C_{din}$	$F_r.E_{din}$	$F_r.Z_{4Hz}$
$\mathbf{X_m}$	-	$X_m.X_m$	$X_m.R_o$	$X_m.S$	$X_m.R_m$	$X_m.C_{din}$	$X_m.E_{din}$	$X_m.Z_{4Hz}$
$\mathbf{R_o}$	-	-	$R_o.R_o$	$R_o.S$	$R_o.R_m$	$R_o.C_{din}$	$R_o.E_{din}$	$R_o.Z_{4Hz}$
S	-	-	-	S.S	$S.R_m$	$S.C_{din}$	$S.E_{din}$	$S.Z_{4Hz}$
$\mathbf{R_m}$	-	-	-	-	$R_m.R_m$	$R_m.C_{din}$	$R_m.E_{din}$	$R_m.Z_{4Hz}$
$\mathbf{C_{din}}$	-	-	-	-	-	$C_{din}.C_{din}$	$C_{din}.E_{din}$	$C_{din}.Z_{4Hz}$
$\mathbf{E_{din}}$	-	-	-	-	-	-	$E_{din}.E_{din}$	$E_{din}.Z_{4Hz}$
$\mathbf{Z_{4Hz}}$	-	-	-	-	-	-	-	$Z_{4Hz}.Z_{4Hz}$

Tabela 41 – 15 Combinações geradas pelo produto cruzado no experimento do item 5.6 com seleção de cinco atributos

	$\mathbf{X_m}$	$\mathbf{Z_{4Hz}}$	$\mathbf{R_m}$	$\mathbf{C_{din}}$	$\mathbf{R_o}$
$\mathbf{X_m}$	$X_m.X_m$	$X_m.Z_{4Hz}$	$X_m.R_m$	$X_m.C_{din}$	$X_m.R_o$
$\mathbf{Z_{4Hz}}$	-	$Z_{4Hz}.Z_{4Hz}$	$Z_{4Hz}.R_m$	$Z_{4Hz}.C_{din}$	$Z_{4Hz}.R_o$
$\mathbf{R_m}$	-	-	$R_m.R_m$	$R_m.C_{din}$	$R_m.R_o$
$\mathbf{C_{din}}$	-	-	-	$C_{din}.C_{din}$	$C_{din}.R_o$
$\mathbf{R_o}$	-	-	-	-	$R_o.R_o$

APÊNDICE B – Inferência sobre estruturas de Redes Bayesianas

1. Inferência sobre a Rede 1 com cinco atributos de entrada

A rede da Figura 39 possui uma característica diferente das outras redes apresentadas. Além da variável *classe* (Tabela 42), a variável R_o também foi usada como nó raiz, e portanto, também possui uma tabela de probabilidade à priori (Tabela 43). Isso ocorre devido à aleatoriedade dos indivíduos gerados na população inicial ou criados através dos operadores de mutação e *crossover* do algoritmo genético. Esses indivíduos são apresentados como possíveis soluções ao problema e selecionados de acordo com o valor da AUC que possuem, ou seja, o algoritmo genético não leva em consideração a presença de um ou mais nós raízes, desde que a estrutura gerada tenha um bom desempenho. Mesmo com essa particularidade, é possível inferir sobre as tabelas de DPC obtidas por essa rede.

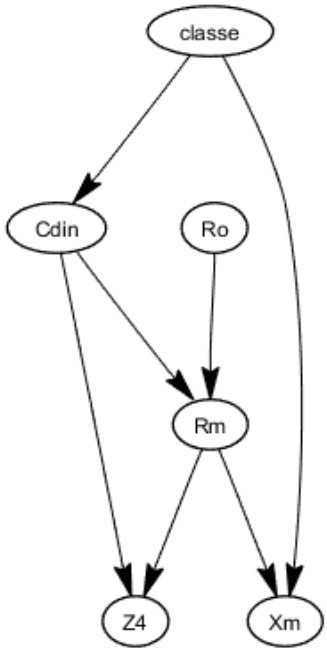

Figura 39 – Estrutura da rede 1 gerada com cinco atributos de entrada

Tabela 42 – Probabilidades à priori da variável *classe* da rede 1 com cinco atributos de entrada

P(*classe* = 0)	P(*classe* = 1)
0,50	0,50

Tabela 43 – Probabilidades à priori da variável R_o da rede 1 com cinco atributos de entrada

$P(R_o = 1)$	$P(R_o = 2)$
0,69	0,31

A variável C_{din} é influenciada apenas pela variável *classe*. Pela Tabela 44, pode-se observar que há probabilidade de 0,92 do indivíduo ter complacência maior, dado que não possui a doença (*classe*=0). Para o caso de valores mais baixos (C_{din}=1), há uma possibilidade de 0,61 do paciente pertencer à classe *1*.

Tabela 44 – DPC da variável C_{din} da rede 1 com cinco atributos de entrada

	$P(C_{din} = 1 \mid classe)$	$P(C_{din} = 2 \mid classe)$
$classe = 0$	0,08	0,92
$classe = 1$	0,61	0,39

De acordo com a rede 1, a variável X_m recebe influencias das variáveis R_m e *classe*. Pela Tabela 45, são retiradas as condições a seguir mostrando que na maioria das combinações há menor probabilidade de um indivíduo apresentar X_m=1, dado que foi observado um baixo valor da resistência (R_m=1), independente do valor da classe:

$$P(X_m = 1 | R_m = 1, classe = 0) = 0{,}04$$
$$P(X_m = 1 | R_m = 1, classe = 1) = 0{,}25$$

Outra observação pode ser feita levando em consideração a variável *classe*. Também há menor chance de um indivíduo apresentar baixa reatância, dado que não é portador da doença (*classe*=0):

$$P(X_m = 1 | R_m = 1, classe = 0) = 0{,}04$$
$$P(X_m = 1 | R_m = 2, classe = 0) = 0{,}36$$

Tabela 45 – DPC da variável X_m da rede 1 com cinco atributos de entrada

	$P(X_m = 1 \mid R_m, classe)$	$P(X_m = 2 \mid R_m, classe)$
$R_m = 1$, *classe* = 0	0,04	0,96
$R_m = 2$, *classe* = 0	0,36	0,64
$R_m = 1$, *classe* = 1	0,25	0,75
$R_m = 2$, *classe* = 1	0,78	0,22

A variável Z_{4Hz} é influenciada por R_m e C_{din}. Pela Tabela 46 pode-se observar que há maior probabilidade em ter alta impedância Z_{4Hz}, dado que foi observada complacência igual a 1. Outra observação a ser feita é sobre a combinação R_m=2 e C_{din}=2. Trata-se de uma situação difícil de ocorrer, pois normalmente um indivíduo com alta resistência R_m é portador da doença e possui complacência C_{din}=1. Mesmo assim o algoritmo calcula uma probabilidade de 0,65 para um paciente ter alta impedância Z_{4Hz}, dado que foram observados R_m e C_{din} iguais a 2.

Tabela 46 – DPC da variável Z_{4Hz} da rede 1 gerada com cinco atributos de entrada

	$P(Z_{4Hz} = 1 \mid R_m, C_{din})$	$P(Z_{4Hz} = 2 \mid R_m, C_{din})$
$R_m = 1$, $C_{din} = 1$	0,22	0,78
$R_m = 2$, $C_{din} = 1$	0,05	0,95
$R_m = 1$, $C_{din} = 2$	0,97	0,03
$R_m = 2$, $C_{din} = 2$	0,35	0,65

A Tabela 47 mostra as relações de R_m, C_{din} e R_o, de acordo com a rede 1. As probabilidades contidas nessa tabela reafirmam a relação direta entre R_m e R_o, mostrando a alta probabilidade de R_m ser baixo, dado que R_o também é baixo. Da mesma forma, observa-se alta probabilidade de um paciente ter R_m=2, dado que foi observado R_o=2.

Tabela 47 – DPC da variável R_o da rede 1 gerada com cinco atributos de entrada

	$P(R_m = 1 \mid C_{din}, R_o)$	$P(R_m = 2 \mid C_{din}, R_o)$
$C_{din} = 1$, $R_o = 1$	0.88	0.12
$C_{din} = 2$, $R_o = 1$	0.98	0.02
$C_{din} = 1$, $R_o = 2$	0.18	0.82
$C_{din} = 2$, $R_o = 2$	0.19	0.81

2. Inferência sobre a Rede 2 com cinco atributos de entrada

A rede da Figura 40 possui seis tabelas de DPC. As probabilidades à priori da variável *classe* estão na Tabela 48 e mostram a probabilidade de um indivíduo pertencer à classe *0* ou a classe *1*.

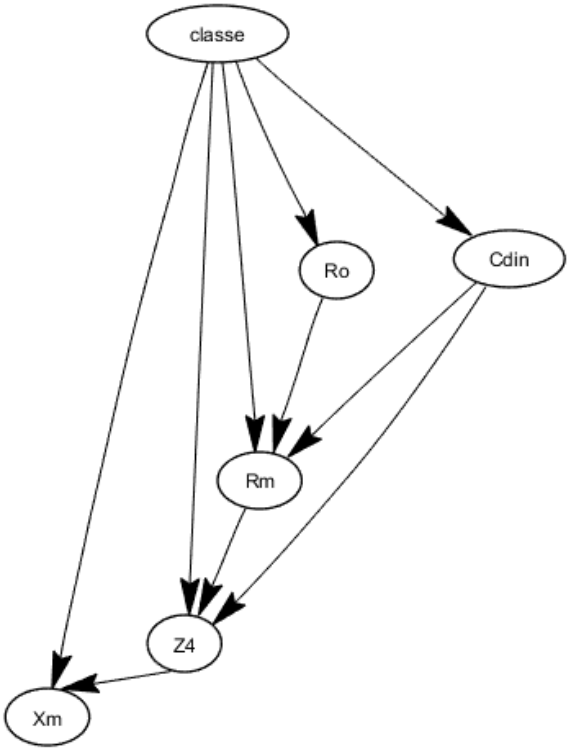

Figura 40 – Estrutura da rede 2 gerada com cinco atributos de entrada

Tabela 48 – Probabilidades à priori da variável *classe* da rede 2 com cinco atributos de entrada

P(*classe* = 0)	P(*classe* = 1)
0,50	0,50

Conforme a rede da Figura 40, a variável R_o é influenciada apenas pela variável *classe*. Pela Tabela 49 é possível concluir que há probabilidade de um indivíduo ter baixa resistência R_o, dado que pertence a *classe 0*. Da mesma forma, há chances do indivíduo ter alto valor de R_o, dado que pertence a *classe 1*.

Tabela 49 – DPC para a variável R_o da rede 2 gerada com cinco atributos de entrada

	$P(R_o = 1 \mid classe)$	$P(R_o = 2 \mid classe)$
classe = 0	0,94	0,06
classe = 1	0,47	0,53

Assim como observado na rede 1, a variável C_{din} é dependente apenas da variável *classe*. De acordo com as condições a seguir, retiradas da Tabela 50, há alta probabilidade de um indivíduo ter complacência alta (C_{din}=2), dado que não possui a doença (*classe*=0). Também há maior probabilidade do paciente ter baixa complacência (C_{din}=1), dado que possui a doença (*classe*=1).

Tabela 50 – DPC para a variável C_{din} da rede 2 com cinco atributos de entrada

	$P(C_{din} = 1 \mid classe)$	$P(C_{din} = 2 \mid classe)$
classe = 0	0,06	0,94
classe = 1	0,61	0,39

As variáveis Z_{4Hz} e *classe* exercem influência sobre a variável X_m. Conforme as probabilidades da Tabela 51, independente da classe, há alta probabilidade de um indivíduo ter alta reatância (X_m=2), dado que Z_{4Hz} é baixo:

Tabela 51 – DPC para a variável X_m da rede 2 com cinco atributos de entrada

	$P(X_m = 1 \mid Z_{4Hz}, classe)$	$P(X_m = 2 \mid Z_{4Hz}, classe)$
Z_{4Hz} = 1, *classe* = 0	0,02	0,98
Z_{4Hz} = 2, *classe* = 0	0,64	0,36
Z_{4Hz} = 1, *classe* = 1	0,06	0,94
Z_{4Hz} = 2, *classe* = 1	0,80	0,20

A variável Z_{4Hz} é influenciada pelas variáveis: R_m, C_{din} e *classe*. De acordo com a Tabela 52, há maior probabilidade de um indivíduo apresentar alta impedância Z_{4Hz}, dado que é portador da doença:

Tabela 52 – DPC para a variável Z_{4Hz} da rede 2 com cinco atributos de entrada

	$P(Z_{4Hz} = 1 \mid R_m, C_{din}, classe)$	$P(Z_{4Hz} = 2 \mid R_m, C_{din}, classe)$
$R_m = 1, C_{din} = 1, classe = 0$	0,28	0,72
$R_m = 2, C_{din} = 1, classe = 0$	0,50	0,50
$R_m = 1, C_{din} = 2, classe = 0$	0,99	0,01
$R_m = 2, C_{din} = 2, classe = 0$	0,72	0,28
$R_m = 1, C_{din} = 1, classe = 1$	0,23	0,77
$R_m = 2, C_{din} = 1, classe = 1$	0,03	0,97
$R_m = 1, C_{din} = 2, classe = 1$	0,92	0,08
$R_m = 2, C_{din} = 2, classe = 1$	0,19	0,81

As variáveis C_{din}, R_o e *classe* influenciam R_m. As probabilidades da Tabela 53 mostram que R_m é diretamente proporcional a R_o, independente dos valores assumidos pelas variáveis *classe* e C_{din}.

O algoritmo apresentou probabilidade igual a 0,5 para o caso de um indivíduo com as seguintes características: C_{din}=1, R_o=2 e *classe*=0. Normalmente, isto ocorre devido à combinação difícil de um indivíduo ter complacência baixa (C_{din}=1), alta resistência (R_o=2) e não possuir a doença (*classe*=0). Em geral, essas são características de portadores de fibrose cística (*classe*=1).

Tabela 53 – DPC para a variável R_m da rede 2 com cinco atributos de entrada

	$P(R_m = 1 \mid C_{din}, R_o, classe)$	$P(R_m = 2 \mid C_{din}, R_o, classe)$
$C_{din} = 1, R_o = 1, classe = 0$	0,72	0,28
$C_{din} = 2, R_o = 1, classe = 0$	0,99	0,01
$C_{din} = 1, R_o = 2, classe = 0$	0,50	0,50
$C_{din} = 2, R_o = 2, classe = 0$	0,28	0,72
$C_{din} = 1, R_o = 1, classe = 1$	0,80	0,20
$C_{din} = 2, R_o = 1, classe = 1$	0,97	0,03
$C_{din} = 1, R_o = 2, classe = 1$	0,17	0,83
$C_{din} = 2, R_o = 2, classe = 1$	0,19	0,81