



Universidade do Estado do Rio de Janeiro

Centro de Tecnologia e Ciências

Faculdade de Engenharia

Domingos Sávio Mattos de Andrade

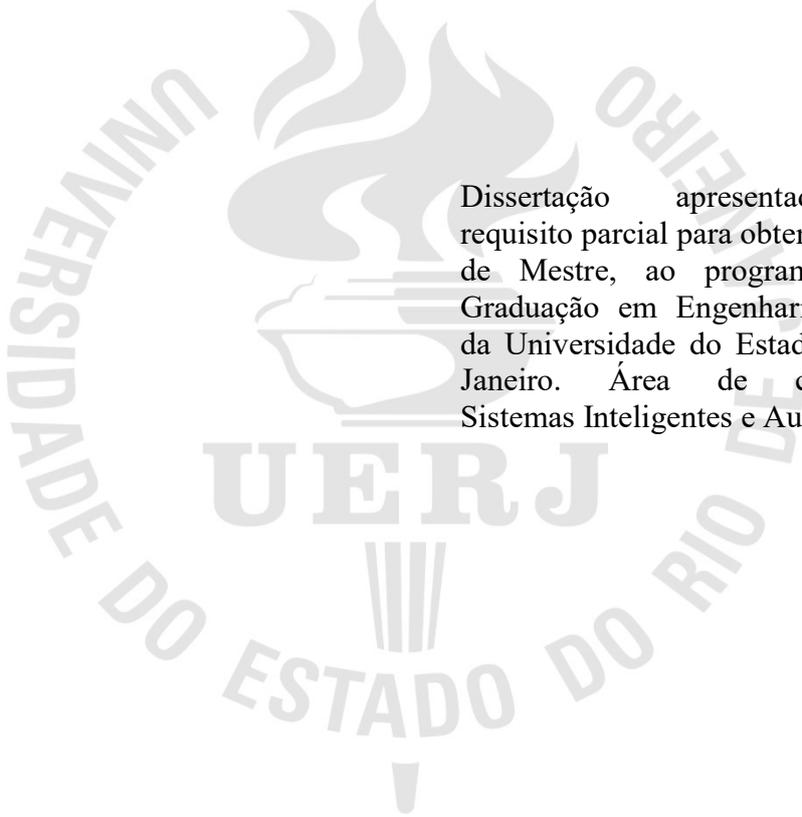
**Detecção de Alterações Respiratórias na Esclerose Sistêmica Através da
Técnica de Oscilações Forçadas e Algoritmos de Aprendizado de Máquinas**

Rio de Janeiro

2018

Domingos Sávio Mattos de Andrade

**Detecção de Alterações Respiratórias na Esclerose Sistêmica Através da Técnica
de Oscilações Forçadas e Algoritmos de Aprendizado de Máquinas**



Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao programa de Pós-Graduação em Engenharia Eletrônica da Universidade do Estado do Rio de Janeiro. Área de concentração: Sistemas Inteligentes e Automação.

Orientadores: Prof. Dr. Jorge Luís Machado do Amaral

Prof. Dr. Pedro Lopes de Melo

Rio de Janeiro

2018

CATALOGAÇÃO NA FONTE
UERJ / REDE SIRIUS / BIBLIOTECA CTC/B

A553 Andrade, Domingos Sávio Mattos de.
Detecção de alterações respiratórias na esclerose sistêmica através da técnica de oscilações forçadas e algoritmos de aprendizado de máquinas / Domingos Sávio Mattos de Andrade. – 2018.
119f.

Orientadores: Jorge Luís Machado do Amaral, Pedro Lopes de Melo.

Dissertação (Mestrado) – Universidade do Estado do Rio de Janeiro, Faculdade de Engenharia.

1. Engenharia eletrônica - Teses. 2. Aprendizado do computador - Teses. 3. Algoritmos - Teses. 4. Aparelho respiratório - Doenças - Teses. I. Amaral, Jorge Luís Machado do. II. Melo, Pedro Lopes de. III. Universidade do Estado do Rio de Janeiro, Faculdade de Engenharia. IV. Título.

CDU 004.891

Bibliotecária: Júlia Vieira – CRB7/6022

Autorizo, apenas para fins acadêmicos e científicos, a reprodução total ou parcial desta tese, desde que citada a fonte.

Assinatura

Data

Domingos Sávio Mattos de Andrade

Detecção de Alterações Respiratórias na Esclerose Sistêmica Através da Técnica de Oscilações Forçadas e Algoritmos de Aprendizado de Máquinas

Dissertação apresentada, como requisito parcial para obtenção do título de Mestre, ao programa de Pós-Graduação em Engenharia Eletrônica da Universidade do Estado do Rio de Janeiro. Área de concentração: Sistemas Inteligentes e Automação.

Aprovado em:

Banca Examinadora:

Prof. Dr. Jorge Luís Machado do Amaral (Orientador)

Faculdade de Engenharia - UERJ

Prof. Dr. Pedro Lopes de Melo (Orientador)

Laboratório de Instrumentação Biomédica - Instituto de Biologia - UERJ

Prof. Dr. Fabiano Saldanha Gomes de Oliveira

Instituto de Medicina Social - UERJ

Prof. PhD. Carlos Eduardo Pedreira

COPPE - UFRJ

Rio de Janeiro

2018

DEDICATÓRIA

Dedico este trabalho ao Laboratório de Redes Industriais e Sistemas de Automação (LARISA), ao Laboratório de Instrumentação Biomédica da UERJ (LIB - UERJ) e a todos os seus membros, que direta ou indiretamente contribuíram para que este trabalho tornasse realidade, mesmo diante do caos vivido pelo Estado do Rio de Janeiro.

AGRADECIMENTOS

Em primeiro lugar, só posso agradecer a Deus por ter dado forças para superar todos os obstáculos que surgem ao longo de nossa jornada. Agradeço aos meus pais e irmã, pela paciência, meus tios por palavras de incentivo e a minha avó, que por mais que a vida tenha lhe tirado grande parte de suas funções cognitivas e consciência, tenho certeza que torceu por mim. A meu avô (*In memoriam*) por tudo que representou em minha vida. À minha amiga Noemi, que sempre esteve ao meu lado, nos bons e maus momentos, com palavras de carinho. Agradeço, também, aos amigos que conheci ao longo do curso: Patrícia e Adriano, que sempre estiveram dispostos a ajudar. Aos amigos do LARISA, que sempre proporcionaram momentos de descontração e aprendizado em meio à crise. Aos professores Jorge Amaral e Pedro Lopes de Melo, minhas reverências por todos os ensinamentos recebidos. Sem o empenho deles não seria possível realizar esse trabalho.

Sonhe com o que você quiser. Vá para onde você queira ir.
Seja o que você quer ser, porque você possui apenas uma vida
e nela só temos uma chance de fazer aquilo que queremos.
Tenha felicidade bastante para fazê-la doce. Dificuldades
para fazê-la forte. Tristeza para fazê-la humana. E
esperança suficiente para fazê-la feliz.

Clarice Lispector

RESUMO

ANDRADE, Domingos Sávio Mattos de. *Detecção de alterações respiratórias na esclerose sistêmica através da técnica de oscilações forçadas e algoritmos de aprendizado de máquinas*. 119f. 2018. Dissertação (Mestrado em Engenharia Eletrônica) – Faculdade de Engenharia, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2018.

A Técnica de Oscilações Forçadas (TOF) tem mostrado um elevado potencial no estudo de alterações mecânicas no sistema respiratório de indivíduos portadores de doenças pulmonares. Devido a sua simplicidade de execução, requerendo pouca colaboração do paciente, ela fornece novos parâmetros, que complementam as avaliações pulmonares realizadas por métodos tradicionais, como a espirometria. A fim de aprimorar a eficácia dessa nova técnica, este trabalho propõe o uso de algoritmos de aprendizado de máquinas para auxiliar a investigação e diagnóstico precoce de alterações respiratórias em portadores de Esclerose Sistêmica. Os dados fornecidos pela TOF foram aplicados nos algoritmos: *K-Nearest Neighbor* (K-NN), *Radial Support Vector Machine* (RSVM), *Adaboost* (ADAB), *Random Forest* (RF) e Aprendizado em Múltiplas Instâncias (AMI), essa última usada como classificador e seletor de atributos. Dentre os experimentos realizados, a complacência dinâmica (C_{din}), foi o atributo que apresentou o melhor desempenho individual (AUC=0,77 para o grupo controle x exame espirométrico normal (grupo 1) e 0,94 para o grupo controle x restritivo (grupo 2)), indicando acurácia moderada e alta, respectivamente. No experimento com todos os atributos da TOF, o melhor desempenho foi alcançado pelo KNN (AUC=0,90) para o grupo 1, indicando alta acurácia e AUC=0,96, alcançada pelo ADAB para o grupo 2. O algoritmo AMI obteve AUC=0,72 para o grupo 1 e AUC=0,94 para o grupo 2. Com cinco atributos selecionados pelo algoritmo de múltiplas instâncias, o melhor resultado foi alcançado pelo algoritmo KNN (AUC = 0,88) para o grupo 1 e pelo ADAB (AUC=0,98) para o grupo 2. Como classificador, o AMI obteve AUC de 0,74 e 0,95, respectivamente. No experimento realizado com cinco atributos selecionados pelo especialista, o melhor desempenho foi alcançado pelo KNN (AUC=0,85) para o grupo 1 e pelo ADAB (AUC=0,98) para o grupo 2. Com a seleção direta de atributos o melhor desempenho também foi alcançado pelo KNN para o grupo 1 (AUC=0,89) e pelo ADAB para o grupo 2 (AUC=0,97). Os experimentos realizados mostraram que o uso de algoritmos de aprendizado de máquinas aumentou a acurácia do diagnóstico da Esclerose Sistêmica, podendo auxiliar no diagnóstico precoce da doença.

Palavras chave: Esclerose Sistêmica; Técnica de Oscilações Forçadas; TOF; Aprendizado de máquina; Múltiplas Instâncias; AUC.

ABSTRACT

ANDRADE, Domingos Sávio Mattos de. *Detection of respiratory changes in systemic sclerosis by forced oscillation technique and machine learning algorithms*. 119f. 2018. Dissertação (Mestrado em Engenharia Eletrônica) – Faculdade de Engenharia, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2018.

The Forced Oscillation Technique (TOF) has shown high potential in the study of mechanical alterations of respiratory diseases. Due to its simplicity of execution, requiring little patient's collaboration, it provides new parameters, which complement the pulmonary evaluations performed by traditional methods, such as spirometry. To prove the efficiency of this new technique, this work proposes the use of machine learning algorithms help the investigation and the early diagnosis of respiratory changes in patients with Systemic Sclerosis. The data provided by TOF were used on the following algorithms: K-Nearest Neighbor (KNN), Radial Support Vector Machine (RSVM), Adaboost (ADAB), Random Forest (RF) and Multiple Instance Learning (AMI) as classifier and attribute selector. From the experiments performed, dynamic compliance (C_{din}) was the attribute that presented the best individual performance (AUC = 0,77 for the control group x normal spirometric examination (group 1) and 0,94 for the control group x restrictive (group 2)), indicating moderate and high accuracy, respectively. In the experiment with all TOF attributes, the best performance was achieved by KNN (AUC = 0,90) for group 1, indicating high accuracy and AUC = 0,96, achieved by ADAB for group 2. The AMI algorithm obtained AUC = 0,72 for group 1 and AUC = 0,94 for group 2. With five attributes selected by the multi-instance algorithm, the best result was achieved by the KNN algorithm (AUC = 0,88) for group 1 and ADAB (AUC = 0,98) for group 2. As a classifier, the AMI obtained AUC of 0,74 and 0,95, respectively. In the experiment performed with five attributes selected by the expert, the best performance was achieved by KNN (AUC = 0,85) for group 1 and ADAB (AUC = 0,98) for group 2. With direct selection, the best performance was also achieved by KNN for group 1 (AUC = 0,89) and by ADAB for group 2 (AUC = 0,97). The experiments showed that the use of machine learning algorithms increased the accuracy of the diagnosis of Systemic Sclerosis, which may help in the early diagnosis of the disease.

Keywords: Systemic Sclerosis; Forced Oscillations Technique; TOF; Machine learning; Multiple Instances; AUC.

LISTA DE ILUSTRAÇÕES

Figura 1 - Esquema simplificado da Técnica de Oscilações Forçadas.....	21
Figura 2 - Voluntária realizando ensaios da TOF.....	22
Figura 3 - Descrição das variáveis P e Q no domínio do Tempo (A). Relações Geométricas entre as medidas empregando módulo e fase da impedância e as que se utilizam da resistência e da reatância (B).....	24
Figura 4 - Demonstração de molécula rotacionada	29
Figura 5 - (A) Aprendizado supervisionado (B) Aprendizado em Múltiplas Instâncias.	29
Figura 6 - Exemplo de bags e instâncias, sendo f1, f2 e f3 representam as três dimensões (atributos) de um exemplo.....	30
Figura 7 - Exemplo de <i>bag</i> e instância aplicado em uma imagem.....	31
Figura 8 - Exemplo de classificação do método 1-NN.....	35
Figura 9 - Exemplo com 3 e 5 vizinhos, respectivamente.....	36
Figura 10 - Estrutura de árvore de decisão ¹	38
Figura 11 - Exemplo de divisão recursiva de uma árvore de decisão	40
Figura 12 - Exemplo de Árvores Aleatórias.....	42
Figura 13 - Exemplo de hiperplano na classificação SVM.....	45
Figura 14 - Hiperplano ótimo para conjunto de dados linearmente separáveis.....	45
Figura 15 - Interpretação geométrica da distância X até o hiperplano ótimo para um padrão de entrada bidimensional	46
Figura 16 - Situações possíveis para o SVM.....	49
Figura 17 - Mapa não linear do espaço de entrada para o espaço de características.....	50
Figura 18 - Esquema simplificado do funcionamento do Adaboost	52
Figura 19 - Taxa de erro <i>versus</i> importância do classificador.....	54
Figura 20 - Fluxograma do modelo proposto	56
Figura 21 - Esquema ilustrativo da Validação Cruzada com 10 pastas.....	60
Figura 22 - Resultados de um teste em dois grupos (doente e saudável)	61
Figura 23 - Matriz de confusão ou tabela de contingência.....	61
Figura 24 - Exemplos de curva ROC	62
Figura 25 - Comparação dos valores da TOF para os grupos controle, esclerodermia (normal exame) e esclerodermia (restritivo)	66

Figura 26 - Curva ROC com os parâmetros da TOF individualmente para os grupos controle e normal exame.....	68
Figura 27 - Curva ROC com os parâmetros da TOF individualmente para os grupos controle e restritivo.....	70
Figura 28 - Curvas ROC obtidas no experimento com todos os atributos da TOF.....	72
Figura 29 - Análise comparativa da sensibilidade obtida nos cinco classificadores com a especificidade de 75% e 90% para os grupos controle e normal exame.....	74
Figura 30 - Curvas ROC obtidas no experimento com todos os atributos da TOF.....	76
Figura 31 - Análise comparativa da sensibilidade obtida nos cinco classificadores com a especificidade de 75% e 90% para os grupos controle e restritivo.....	77
Figura 32 - Curvas ROC obtidas no experimento com os parâmetros da TOF selecionados pelo especialista para os grupos controle e normal exame.....	79
Figura 33 - Curvas ROC obtidas no experimento com os parâmetros da TOF selecionados pelo especialista para os grupos controle e restritivo.....	80
Figura 34 - Curvas ROC obtidas no experimento com os parâmetros da TOF selecionados pelo melhor desempenho do Classificador AMI para os grupos Controle e Normal Exame.....	82
Figura 35 - Curvas ROC obtidas no experimento com os parâmetros da TOF selecionados pelo melhor desempenho do Classificador AMI para os grupos controle e restritivo.....	84
Figura 36 - Curvas ROC obtidas no experimento com os parâmetros da TOF escolhidos através da seleção direta para os grupos controle e normal exame.....	86
Figura 37 - Curvas ROC obtidas no experimento com os parâmetros da TOF escolhidos através da seleção direta de atributos para os grupos controle e restritivo.....	88
Figura 38 - Curvas ROC obtidas no experimento com os parâmetros da TOF selecionados pelo melhor desempenho do Classificador AMI com 3 atributos para os grupos controle e normal exame.....	90
Figura 39 - Curvas ROC obtidas no experimento com os parâmetros da TOF selecionados pelo melhor desempenho do Classificador AMI com 3 atributos para os grupos controle e restritivo.....	91
Figura 40 - Curvas ROC obtidas no experimento com 3 atributos da TOF escolhidos através da seleção direta para os grupos Controle e Normal Exame.....	93
Figura 41 - Curvas ROC obtidas no experimento com 3 atributos da TOF escolhidos através da seleção direta – Análise Controle <i>versus</i> Normal Restritivo.....	95

Figura 42 - Comparativo entre as acurácias obtidas nas seis metodologias em comparação com o BFP para a análise Controle <i>versus</i> Normal Exame.....	96
Figura 43 - Comparativo entre as acurácias obtidas nas seis metodologias testadas em comparação com o BFP para a análise Controle <i>versus</i> Restritivo	97
Figura 44 - Análise comparativa da sensibilidade obtida nos seis experimentos com a especificidade de 75% e 90% - Análise Controle <i>versus</i> Normal Exame.....	110
Figura 45 - Análise comparativa da sensibilidade obtida nos seis experimentos com a especificidade de 75% e 90% - Análise Controle <i>versus</i> Restritivo.....	111
Figura 46 - Análise comparativa da sensibilidade obtida nos seis experimentos com a especificidade de 75% e 90% para os atributos selecionados pelo AMI - Análise Controle <i>versus</i> Normal Exame.....	112
Figura 47 - Análise comparativa da sensibilidade obtida nos seis experimentos com a especificidade de 75% e 90% para os atributos selecionados pelo AMI - Análise Controle <i>versus</i> Restritivo	113
Figura 48 - Análise comparativa da sensibilidade obtida nos seis experimentos com a especificidade de 75% e 90% para os atributos escolhidos através da seleção direta - Análise Controle <i>versus</i> Normal Exame	114
Figura 49 - Análise comparativa da sensibilidade obtida nos seis experimentos com a especificidade de 75% e 90% para os atributos escolhidos através da seleção direta - Análise Controle <i>versus</i> Restritivo	115
Figura 50 - Análise comparativa da sensibilidade obtida nos seis experimentos com a especificidade de 75% e 90% para os atributos selecionados pelo AMI - Análise Controle <i>versus</i> Normal Exame.....	116
Figura 51 - Análise comparativa da sensibilidade obtida nos seis experimentos com a especificidade de 75% e 90% para os atributos selecionados pelo AMI - Análise Controle <i>versus</i> Restritivo	117
Figura 52 - Análise comparativa da sensibilidade obtida nos seis experimentos com a especificidade de 75% e 90% para os três atributos escolhidos através da seleção direta - Análise Controle <i>versus</i> Normal Exame	118
Figura 53 - Análise comparativa da sensibilidade obtida nos seis experimentos com a especificidade de 75% e 90% para os três atributos escolhidos através da seleção direta - Análise Controle <i>versus</i> Restritivo	119

LISTA DE TABELAS

Tabela 1 - Parâmetros fornecidos pela TOF	26
Tabela 2 - Conjunto de exemplos no formato atributo-valor	28
Tabela 3 - Pseudocódigo da configuração 1-NN.....	37
Tabela 4 - Pseudocódigo do <i>Random Forest</i>	43
Tabela 5 - Pseudocódigo do algoritmo Adaboost.....	54
Tabela 6 - Experimentos para determinação dos parâmetros dos algoritmos	64
Tabela 7 - Desempenho individual dos atributos da TOF na classificação de pacientes entre os grupos controle e normal exame	68
Tabela 8 - Desempenho individual dos atributos da TOF na classificação de pacientes entre os grupos controle e Restritivo	69
Tabela 9 - Resultados dos classificadores com todos os atributos da TOF.....	71
Tabela 10 - Comparação das AUCs entre os modelos com todos os atributos da TOF – Análise Controle <i>versus</i> Normal Exame	73
Tabela 11 - Resultados dos classificadores com todos os atributos da TOF.....	75
Tabela 12 - Comparação das AUCs entre os modelos com todos os atributos da TOF – Análise Controle <i>versus</i> Restritivo	76
Tabela 13 – Resultados obtidos na seleção de atributos pelo Especialista.....	78
Tabela 14 - Comparação das AUCs entre os modelos com a seleção de atributos pelo especialista – Análise Controle <i>versus</i> Normal Exame	80
Tabela 15 - Comparação das AUCs entre os modelos com a seleção pelo especialista – Análise Controle <i>versus</i> Restritivo	81
Tabela 16 - Resultados obtidos na seleção de atributos pelo AMI.....	82
Tabela 17 - Comparação das AUCs entre os modelos com cinco atributos selecionados pelo melhor desempenho AMI – Análise Controle <i>versus</i> Normal Exame	83
Tabela 18 - Comparação das AUCs entre os modelos com cinco atributos selecionados pelo melhor desempenho do algoritmo de AMI – Análise Controle <i>versus</i> Restritivo..	85
Tabela 19 – Resultados obtidos através da Seleção Direta de 5 Atributos.....	86
Tabela 20 - Comparação das AUCs entre os modelos com cinco atributos escolhidos através da seleção direta – Análise Controle <i>versus</i> Normal Exame	87
Tabela 21 - Comparação das AUCs entre os modelos com cinco atributos escolhidos através da seleção direta de atributos – Análise Controle <i>versus</i> Restritivo	88

Tabela 22 - Resultados obtidos na seleção de 3 atributos através do melhor desempenho do AMI	89
Tabela 23 - Comparação das AUC entre os modelos com três atributos selecionados pelo melhor desempenho AMI – Análise Controle <i>versus</i> Normal Exame	90
Tabela 24 - Comparação das AUCs entre os modelos com três atributos selecionados pelo melhor desempenho AMI - Análise Controle <i>versus</i> Restritivo.....	92
Tabela 25 – Resultados obtidos na Seleção Direta de 3 atributos	92
Tabela 26 - Comparação das AUC entre os modelos com três atributos escolhidos através da seleção direta – Análise Controle <i>versus</i> Normal Exame	94
Tabela 27 - Comparação das AUC entre os modelos com três atributos escolhidos através da seleção direta para os grupos Controle <i>versus</i> Restritivo.....	95
Tabela 28 - Combinações de 5 atributos do AMI para os grupos controle <i>versus</i> normal exame.....	106
Tabela 29 - Combinações de 5 atributos do AMI para os grupos controle <i>versus</i> restritivo.....	107
Tabela 30: Combinações de 3 atributos do AMI para os grupos controle <i>versus</i> normal exame.....	108
Tabela 31 - Combinações de 3 atributos do AMI para os grupos controle <i>versus</i> restritivo.....	109

LISTA DE ABREVIATURAS E SIGLAS

ADAB	<i>Adaboost</i>
AMI	Aprendizado em Múltiplas Instâncias
ANOVA	Análise de Variância
AUC	<i>Area Under the ROC curve</i>
BFP	<i>Best FOT Parameter</i>
COPD	<i>Chronic obstructive pulmonary disease</i>
DD	<i>Diverse Density</i>
DPOC	Doença Pulmonar Obstrutiva Crônica
ES	Esclerose Sistêmica
FFT	<i>Fast Fourier transform</i>
HUPE	Hospital Universitário Pedro Ernesto
IA	Inteligência Artificial
IC	Intervalo de confiança
KNN	<i>K Nearest Neighbor</i>
LIB	Laboratório de Instrumentação Biomédica da UERJ
MIL	<i>Multiple Instance Learning</i>
ML	<i>Machine Learning</i>
PNT	Pneumotacômetro
Prtools	<i>Pattern Recognition Toolbox</i>
RF	<i>Random Forest</i>
RN	Redes Neurais
ROC	<i>Receiver Operating Characteristic</i>
RSVM	<i>Radial Support Vector Machine</i>
TP	Transdutor de Pressão
UERJ	Universidade do Estado do Rio de Janeiro

SUMÁRIO

INTRODUÇÃO.....	16
1. ESCLEROSE SISTÊMICA.....	19
2. TÉCNICA DE OSCILAÇÕES FORÇADAS.....	21
3. APRENDIZADO DE MÁQUINAS	27
3.1. Aprendizado em Múltiplas Instâncias	28
3.1.1. Densidade Diversa	32
3.2. Algoritmo dos Vizinhos mais Próximos	33
3.3. Árvore de Decisão	37
3.3.1. Florestas Aleatórias	41
3.4. Máquina de Vetores de Suporte	44
3.5. Adaboost.....	51
4. MODELO PROPOSTO.....	56
4.1. Dados de Entrada.....	57
4.2. Extração de Atributos	57
4.3. Seleção de Atributos.....	58
4.4. Seleção e Treinamento do modelo	59
4.5. Medida de Desempenho do Modelo.....	60
4.6. Classificadores.....	63
5. ESTUDO DE CASO.....	65
5.1. Descrição do Conjunto de Dados	65
5.2. Experimento com apenas um atributo	67
5.2.1. Análise grupo controle <i>versus</i> grupo normal exame	67
5.2.2. Análise grupo controle <i>versus</i> grupo restritivo	69
5.3. Experimento com todos os atributos	70
5.3.1. Análise grupo controle <i>versus</i> grupo normal exame	70
5.3.2. Análise grupo controle <i>versus</i> grupo restritivo	74
5.4. Experimento com cinco atributos.....	78
5.4.1. Experimento com cinco atributos selecionados pelo especialista	78
5.4.2. Experimento com cinco atributos selecionados pelo melhor classificador de Aprendizado em Múltiplas Instâncias	81
5.4.3. Experimento com cinco atributos escolhidos através da seleção direta	85

5.5. Experimento com três atributos selecionados pelo melhor classificador de Aprendizado em Múltiplas Instâncias	89
5.6. Experimento com três atributos escolhidos através da seleção direta para os grupos controle e normal exame.....	92
5.7. Conclusões do capítulo.....	96
CONCLUSÃO.....	98
REFERÊNCIAS	100
APÊNDICE 1 - VALORES DOS ATRIBUTOS DA TOF.....	105
APÊNDICE 2 - COMBINAÇÕES DO AMI	106
APÊNDICE 3 – ANÁLISE COMPARATIVA DA SENSIBILIDADE E ESPECIFICIDADE ATRAVÉS DA SELEÇÃO PELO ESPECIALISTA	110
APÊNDICE 4 – ANÁLISE COMPARATIVA DA SENSIBILIDADE E ESPECIFICIDADE ATRAVÉS DA SELEÇÃO PELO AMI	112
APÊNDICE 5 – ANÁLISE COMPARATIVA DA SENSIBILIDADE E ESPECIFICIDADE ATRAVÉS DA SELEÇÃO DIRETA	114
APÊNDICE 6 – ANÁLISE COMPARATIVA DA SENSIBILIDADE E ESPECIFICIDADE ATRAVÉS DA SELEÇÃO PELO AMI.....	116
APÊNDICE 7 – ANÁLISE COMPARATIVA DA SENSIBILIDADE E ESPECIFICIDADE ATRAVÉS DA SELEÇÃO DIRETA	118

Introdução

A esclerose sistêmica (ES) é uma enfermidade autoimune, extremamente heterogênea na sua apresentação clínica e segue um curso variável e imprevisível. As manifestações pulmonares mais comuns da esclerodermia são a fibrose pulmonar, onde o pulmão perde a sua complacência e a hipertensão arterial. Os lobos pulmonares inferiores são primariamente afetados, porém pode haver uma expansão progressiva para as demais partes do pulmão. Nestes pacientes, tanto a fibrose pulmonar quanto a hipertensão arterial pulmonar estão associadas com a redução da sobrevida (MIRANDA, 2010).

O sintoma respiratório mais comum é a dispneia aos esforços, muitas vezes associada com tosse. Outras manifestações sintomáticas são: espessamento e derrame pleural, dilatação esofágica e sinais de esclerose miocárdica.

Por muitas vezes, a ES é subestimada em sua fase inicial, porque os pacientes apresentam sintomas pulmonares mínimos ou ainda permanecem assintomáticos no estágio inicial da doença, fatos que não permitem um diagnóstico conclusivo.

Trabalhos recentes mostraram que os resultados obtidos através da Técnica de Oscilações Forçadas (TOF) mostraram-se coerentes em descrever a fisiopatologia da Esclerose Sistêmica (MIRANDA, BRAGA, *et al.*, 2006). No entanto, segundo os autores, a TOF não é utilizada na avaliação de indivíduos portadores dessa doença por existirem poucos trabalhos na literatura que avaliem o potencial da técnica na avaliação da enfermidade. Além disso, trabalhos recentes mostram que os algoritmos de aprendizado de máquinas podem aumentar a acurácia diagnóstica e ser usados como apoio a decisão clínica no diagnóstico e estudo de doenças respiratórias.

A utilização da TOF em conjunto com algoritmos de *machine Learning* (ML) foi realizada em (AMARAL, LOPES, *et al.*, 2012), e (AMARAL, LOPES, *et al.*, 2013). Nesses artigos, os autores investigaram o potencial da combinação da TOF com ML para a identificação de alterações respiratórias causadas por diferentes fatores, dentre eles, a Doença Pulmonar Obstrutiva Crônica (DPOC).

Em outro trabalho científico mais recente, chamado “*High-Accuracy detection of airway obstruction in asthma using machine learning algorithms and forced oscillation measurements*”(AMARAL, LOPES, *et al.*, 2017) , os autores buscaram demonstrar que

a combinação de TOF e ML poderia ser aplicada no aprimoramento da detecção de obstrução respiratória na asma. Tal objetivo deveu-se ao fato de os testes tradicionais serem invasivos e muitas pessoas já acometidas pela doença não conseguiam realizar as manobras. Assim, a TOF aparece como um novo método, não invasivo, para a detecção da morbidade.

De toda sorte, os autores lançaram mão de algoritmos de aprendizado de máquinas para comparar seu desempenho com o melhor parâmetro da TOF. Nesse estudo, todos os classificadores utilizados obtiveram um aumento na acurácia diagnóstica, sendo dois deles quase chegando à faixa de alta acurácia. No artigo, os autores afirmam que os algoritmos de aprendizado de máquinas podem servir para ajudar no diagnóstico da doença, auxiliando no correto diagnóstico médico.

Em consonância com a linha de estudo mostrada acima, a presente dissertação pretende seguir a mesma vertente teórica, aplicando os métodos tradicionais e consolidados de algoritmos de aprendizado de máquinas para classificar os dados da Esclerose Sistêmica, fornecidos pela TOF. Além dos métodos tradicionais, será também avaliado o desempenho do algoritmo de aprendizado de múltiplas instâncias aplicado ao problema em estudo. O uso dos algoritmos justifica-se pelos resultados obtidos nos estudos acima apresentados, que se revelaram promissores na investigação de doenças respiratórias.

Nesse contexto, o objetivo do estudo foi aplicar algoritmos de aprendizado de máquinas visando aprimorar o diagnóstico de alterações respiratórias em portadores de Esclerose Sistêmica por meio dos parâmetros fornecidos pela TOF.

O método tradicional de diagnóstico de alterações respiratórias na esclerose sistêmica é através da espirometria, que realiza a análise de fluxos e volumes respiratórios. Esse é um exame simples que não caracteriza, em detalhes, o sistema respiratório humano, que é extremamente complexo. Essa simplicidade impede que alguns portadores da Esclerose Sistêmica sejam identificados por essa técnica, mas que são identificados pela TOF. Com o intuito de aprimorar a acurácia diagnóstica, propõe-se além da utilização de algoritmos de aprendizado de máquinas mais conhecidos, tais como *K-Nearest Neighbor*, *Adaboost*, *Random Forest*, *Radial Support Vector Machine*, a utilização do aprendizado em múltiplas instâncias.

Na literatura, há trabalhos que relacionam a utilização da técnica de aprendizado em múltiplas instâncias na área biomédica. Um deles é o artigo "*Classification of COPD with Multiple Instance Learning*" (CHEPLYGINA, Sørensen, et al., 2014),

que utilizou essa metodologia para classificar portadores de DPOC a partir de imagens. O resultado obtido foi superior aos relatados na literatura, com $AUC = 0,776$. Essa técnica se mostra adequada para dados fracamente rotulados, como ocorre na classificação da ES pela espirometria.

Os três primeiros capítulos são destinados à revisão teórica dos principais assuntos desta dissertação. O primeiro capítulo aborda a Esclerose Sistêmica, indicando suas causas e sintomas. O segundo capítulo apresenta os parâmetros fornecidos pela TOF e as vantagens do uso da técnica. O terceiro capítulo descreve os algoritmos de aprendizado de máquinas utilizados para classificar os dados da TOF. O quarto capítulo apresenta o modelo proposto para este trabalho. Por sua vez, o capítulo 5 mostra todos os resultados dos experimentos realizados e os principais pontos observados. Por fim, os principais pontos deste trabalho são apresentados na conclusão.

1. Esclerose Sistêmica

A ES é uma doença do tecido conectivo, crônica, multissistêmica, de etiologia não definida, caracterizada por alterações degenerativas e anormalidades vasculares na pele, articulações e órgãos viscerais como coração, rins e pulmões, culminando na fibrose cística (SANTOS, FARIA e TRAD, 2006), (OSTOJIC, CERINIC, *et al.*, 2007), (MIRANDA, 2010). A manifestação da doença é resultado de uma interação entre fatores genéticos, imunológicos e ambientais.

A ES é uma doença rara, um pouco mais comum em indivíduos negros, cuja incidência é de aproximadamente 10 casos por milhão de habitantes por ano (OMAR, RAED e ALEJANDRO, 1998), (MIRANDA, BRAGA, *et al.*, 2006), prevalente no sexo feminino, na proporção de 3:1, rara na infância e atinge seu pico de acometimento na 5ª década de vida, onde o acometimento pulmonar é uma significativa causa de morbidade e mortalidade. A detecção precoce do acometimento pulmonar é um fator importante na sobrevivência do indivíduo.

As manifestações pulmonares mais comuns da esclerodermia são a fibrose pulmonar, onde o pulmão perde a sua complacência, e que ocorre em aproximadamente 80% dos casos e a hipertensão arterial pulmonar, com incidência em 50% dos casos (OMAR, RAED e ALEJANDRO, 1998). Os lobos pulmonares inferiores são primariamente afetados, porém pode haver uma expansão progressiva para as demais partes do pulmão. Nestes pacientes, tanto a fibrose pulmonar quanto a hipertensão arterial pulmonar estão associadas com a redução da sobrevida (MIRANDA, 2010). Os sintomas respiratórios mais comuns são a dispnéia aos esforços, muitas vezes associada com tosse. Outras manifestações sintomáticas são: espessamento e derrame pleural, dilatação esofágica e sinais de esclerose miocárdica.

A doença pode ser classificada, clinicamente, segundo a extensão do acometimento cutâneo em: limitada, com espessamento distal da pele dos cotovelos e joelhos ou da face, e difusa, com grande comprometimento da pele estendendo-se, proximalmente, joelhos e cotovelos além de afetar, também, face, pescoço e tronco (DU BOIS, 2007). A classificação é baseada no exame físico ou no histórico clínico do paciente, necessitando o paciente apresentar um critério principal ou dois ou mais

critérios menores, descrito pela *American College of Rheumatology* (MASI, RODNAN e MEDSGER, 1980), (DU BOIS, 2007).

(1) Principal critério: espessamento e estiramento simétricos da pele dos dedos e da articulação metacarpofalangeana ou metatarsofalangeana, tórax, abdome, face/pescoço;

(2) Critérios menores: Esclerodactilia; úlceras digitais com perda da impressão digital como resultado de isquemia e ou fibrose pulmonar.

Em alguns casos é necessária a realização de exames complementares, laboratoriais, para detectar a presença de anticorpos específicos associados à doença, a fim de auxiliar o diagnóstico. Tais anticorpos estão presentes em mais de 90% dos pacientes com ES e estão associados com subtipos específicos da doença (ALLANORE, AVOUAC e KAHAN, 2008).

2. Técnica de Oscilações Forçadas

A TOF é um método não invasivo, descrito por (DU BOIS, BRODY, *et al.*), no ano de 1956, com o objetivo de mensurar as propriedades mecânicas do sistema respiratório (MELO, WERNECK e GIANELLA-NETO, 2000(a)). Esta técnica consiste na aplicação de sinais senoidais ao sistema respiratório por um dispositivo externo, onde as oscilações resultantes de fluxo, relacionadas com as oscilações de pressão correspondentes, sobrepõem às ondas de pressão da respiração normal, permitindo que seja estimada a impedância do sistema respiratório (Z_{rs}) (MELO, WERNECK e GIANELLA-NETO, 2000(a)), (FARIA, 2007), (AMARAL, LOPES, *et al.*, 2012), (MIRANDA, 2010), conforme demonstrado na Figura 1.

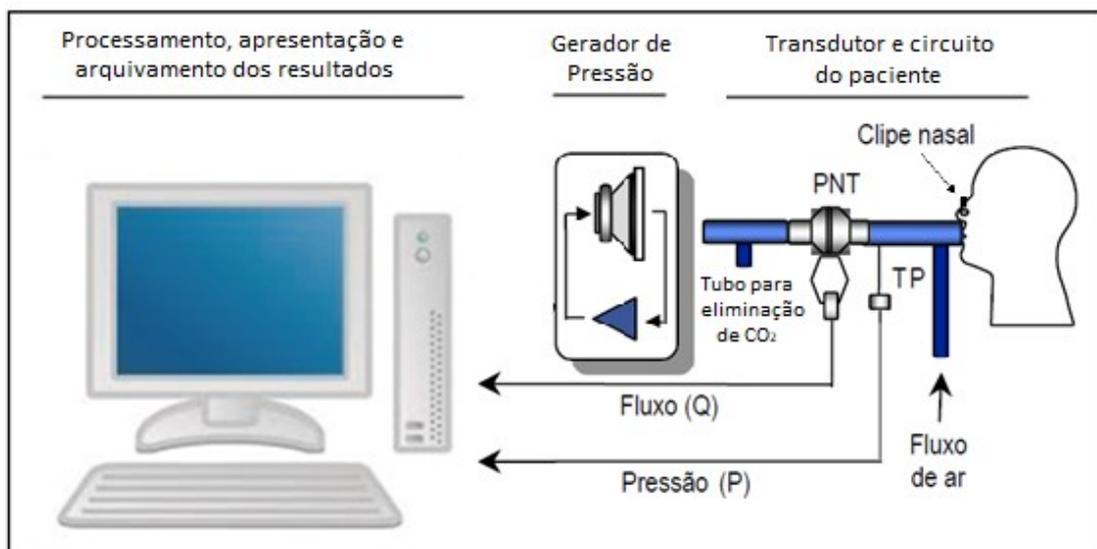


Figura 1 - Esquema simplificado da Técnica de Oscilações Forçadas

(Adaptado de (MELO, 2015))

As principais vantagens da TOF são: a simplicidade na realização do exame, que requer apenas a cooperação passiva do paciente, a obtenção de novos parâmetros complementares a avaliação pulmonar efetuada pelos métodos tradicionais e o tempo reduzido de realização dos exames e processamento dos dados (MELO, WERNECK e GIANELLA-NETO, 2000), (DU BOIS, BRODY, *et al.*, 1956(8)), (OOSTVEEN, MACLEOD, *et al.*, 2003), (DI MANGO, LOPES, *et al.*, 2006). Durante a manobra, o

indivíduo deve permanecer sentado, utilizando um clipe nasal, suportando as bochechas com as mãos e respirando calmamente em um bocal, conforme ilustrado na Figura 2.



Figura 2 - Voluntária realizando ensaios da TOF

(MIRANDA, 2010)

São aplicadas oscilações no sistema respiratório do indivíduo, na faixa compreendida entre 4 e 32Hz, com frequências múltiplas de 2Hz (FARIA, COSTA, *et al.*, 2004),(AMARAL, LOPES, *et al.*, 2012) contendo simultaneamente todas as frequências de interesse para análise, produzidas pelo alto-falante, com o objetivo de reduzir o tempo total do exame (MELO, WERNECK e GIANELLA-NETO, 2000(a)).

A utilização desta faixa de frequência permite que o efeito do sinal respiratório, cuja fundamental encontra-se em torno de 0,25Hz, seja minimizado. Os resultados obtidos nestas frequências refletem, principalmente, os efeitos das vias aéreas centrais e frequências abaixo de 2Hz apresentam resultados dominados pelas propriedades visco elásticas dos tecidos (MELO, WERNECK e GIANELLA-NETO, 2000(a)).

A resultante da respiração espontânea e das oscilações de pressão aplicadas resultam em um sinal de pressão (P) e fluxo (Q) que são captados por um transdutor de pressão (TP) e um pneumotacômetro (PNT), respectivamente. Tais sinais passam por um conjunto composto por amplificadores e filtros e, em seguida, por um analisador de

impedâncias onde, através da Transformada de Fourier (FFT) é efetuada a estimativa da impedância do sistema respiratório. Uma pequena bomba produz um fluxo de ar constante, renovando o ar inspirado pelo indivíduo (MELO, WERNECK e GIANELLA-NETO, 2000(a)).

Existe um índice de validade que garante que a medida obtida pela TOF é aceita, chamado de função de coerência. Tal índice é calculado através da correlação entre o sinal de pressão e de fluxo e, quando em seu valor unitário, garante que a medida está livre da influência de ruídos de origem fisiológicas ou associados à instrumentação (MELO, WERNECK e GIANELLA-NETO, 2000(a)), (FARIA, 2007), (MIRANDA, 2010), (AMARAL, LOPES, *et al.*, 2012). O valor mínimo dessa função que é considerado adequado e limítrofe para aceitação é de 0,9. Valores abaixo deste não são considerados válidos e a manobra é descartada (MELO, WERNECK e GIANELLA-NETO, 2000(a)), (FARIA, 2007), (MIRANDA, 2010), (AMARAL, LOPES, *et al.*, 2012).

A Transformada de Fourier permite decompor os sinais de pressão e fluxo em funções senoidais elementares e avaliar a alteração da impedância respiratória com a frequência. A equação (1) descreve matematicamente este processo:

$$Z_{rs}(f) = \frac{FFT(P)}{FFT(Q)} \quad (1)$$

Sendo:

$Z_{rs}(f)$: Impedância do sistema respiratório

$FFT(P)$: Transformada de Fourier da pressão P

$FFT(Q)$: Transformada de Fourier do fluxo Q

f : frequência desejada

Considerando cada uma das componentes senoidais que compõem o sinal, o comportamento associado à pressão (P) e ao fluxo (Q) pode ser descrito pelas equações (2) e (3), respectivamente.

$$P = P_m \text{sen}(\omega t) \quad (2)$$

$$Q = Q_m \text{sen}(\omega t + \varphi) \quad (3)$$

Sendo:

P_m : amplitude do sinal P

Q_m : amplitude do sinal Q

ω : frequência angular, igual a $2\pi f$

φ : diferença de fase entre os sinais P e Q

Na equação (3), assumindo o sinal de pressão como referência, φ representa a diferença de fase entre a pressão e o fluxo. A Figura 3(A) ilustra tal comportamento descrito acima.

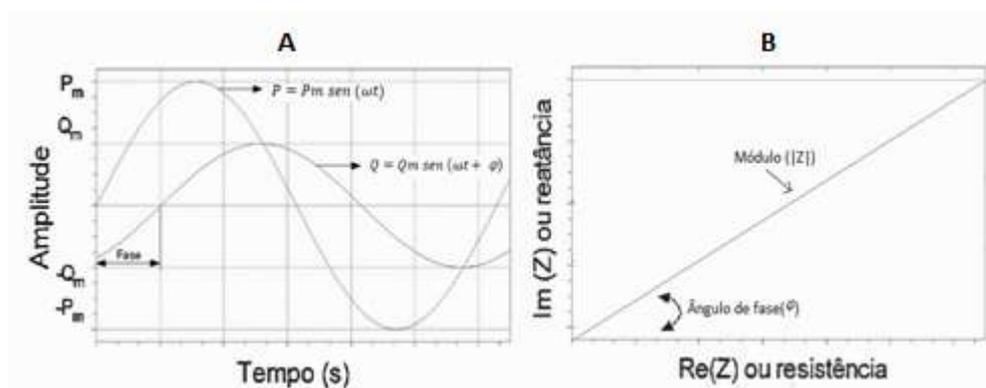


Figura 3 - Descrição das variáveis P e Q no domínio do Tempo (A). Relações Geométricas entre as medidas empregando módulo e fase da impedância e as que se utilizam da resistência e da reatância (B)

(Adaptado de (MELO, WERNECK e GIANELLA-NETO, 2000(a)))

A impedância mecânica descreve o efeito conjunto das propriedades resistivas, elásticas e inertivas do sistema respiratório, caracterizada pela razão das amplitudes dos sinais de pressão e fluxo (P_m / Q_m na Figura 3A), sendo conhecida como módulo da impedância ($|Z|$) e pela diferença de tempo entre os sinais, também demonstrado na Figura 3A, usualmente denominada diferença de fase (φ). A forma mais comum de representar o comportamento da impedância em ensaios utilizando a TOF emprega as componentes em fase, também chamados de reais da impedância ($\text{Re}(Z)$) ou resistência média do sistema respiratório (R_{rS}) e as componentes fora de fase, conhecidos como imaginários da impedância ($\text{Im}(Z)$) ou reatância (X_{rS}).

O primeiro componente denota a soma das resistências de via aérea extratorácica, intratorácica, tecido pulmonar e parede torácica. Por sua vez, o último descreve as propriedades elásticas do sistema, ou seja, compressibilidade de ar e

elasticidade do pulmão e parede torácica. As relações geométricas entre essas variáveis podem ser observadas na Figura 3 B, que mostra que o módulo da impedância representa a carga mecânica total oferecida pelo sistema respiratório, associada aos efeitos da resistência e da reatância através de relações geométricas simples, conforme a equação (4).

$$Z_{rs} = \sqrt{R_{rs}^2 + X_{rs}^2} \quad (4)$$

Cabe ressaltar que pode existir a condição em que a reatância, descrita por ($\text{Im}(Z)$), é insignificante. Assim, a impedância, descrita por ($\text{Re}(Z)$), representa apenas os efeitos da resistência do sistema respiratório. A fase da impedância pode ser descrita, matematicamente, pela equação (5).

$$\varphi = \tan^{-1} \left[\frac{\text{Im}(Z)}{\text{Re}(Z)} \right] \quad (5)$$

Também é possível listar as relações existentes entre a resistência e a reatância com o módulo ($|Z|$) e a fase (φ) da impedância.

$$R_{rs} = |Z_{rs}| \cos \varphi \quad (6)$$

$$X_{rs} = |Z_{rs}| \sin \varphi \quad (7)$$

A TOF é normalmente utilizada na faixa de frequências compreendidas entre 4 e 32 Hz. Assim, a resistência respiratória descreve a dissipação total de energia do sistema, que inclui a soma dos efeitos vindos de resistências relacionadas a quatro fatores: ao tecido pulmonar, a parede torácica, às vias aéreas e à redistribuição de gás nos pulmões. A reatância respiratória (X_{rs}) descreve o armazenamento de energia no sistema respiratório que, por sua vez, está associado à complacência dinâmica (C_{din}), enquanto a energia cinética é descrita pela inertância respiratória (I_{rs}). As propriedades elásticas estão associadas à complacência dinâmica (C_{din}). A relação entre X_{rs} , I_{rs} e C_{din} é descrita na equação (8).

$$X_{rs} = \omega I_{rs} - j \frac{1}{\omega C_{din}} \quad (8)$$

sendo:

$$\omega = 2\pi f$$

$$j = \sqrt{-1}.$$

A inertância respiratória (I_{rs}) descreve a energia cinética usada durante a aceleração do fluxo aéreo. Essa variável pode ser desprezada em análises realizadas em baixa frequência, sendo o sistema respiratório modelado apenas por um componente resistivo e um complacente. No caso da TOF, o efeito da aceleração do fluido torna-se mais evidente, devido à utilização de frequências mais elevadas, fazendo com que a inertância se torne relevante. Desta maneira, os exames da TOF permitem avaliar as propriedades elásticas por meio de resultados obtidos em baixas frequências, enquanto as propriedades inerciais são identificadas através da análise em frequências mais elevadas (MELO, 2015).

O parâmetro R_0 , avaliado na faixa de frequências entre 4-16 Hz, mostra como as resistências newtonianas associadas às vias aéreas e aos tecidos, bem como sua resistência tardia proveniente da distribuição do gás, reagem em frequências baixas. Por outro lado, o parâmetro S está associado à alteração na distribuição do fluxo de gás dentro do sistema respiratório de acordo com a frequência (MIRANDA, DIAS FARIA, *et al.*, 2013).

A Tabela 1 apresenta um resumo dos parâmetros fornecidos pela TOF.

Tabela 1 - Parâmetros fornecidos pela TOF

Parâmetro	Descrição
F_r	Frequência de ressonância
X_m	Reatância média
R_0	Resistência do intercepto
S	Coefficiente angular da curva de resistência
R_m	Resistência média do sistema respiratório
Z_{rs}	Impedância do sistema respiratório
C_{din}	Complacência dinâmica

3. Aprendizado de Máquinas

O Aprendizado de Máquina é uma das áreas da Inteligência Artificial (IA), cujo objetivo é desenvolver sistemas capazes de tomar decisões baseadas no conhecimento acumulado para melhorar o desempenho e a aquisição de experiência (LANGLEY e SIMON, 1995) e que consegue adquirir conhecimento a cada tomada de decisão. Isso significa que após a máquina ter gerado um resultado (decisão), este torna-se um novo conhecimento para a máquina (experiências), juntando-se a base de dados já existente para aprimorar a base de conhecimentos, adquirindo mais experiência a cada nova resposta gerada (tomada de decisão) (KEVIN e DA SILVA, 2016).

Segundo (MITCHELL, 1997): "Diz-se que um programa de computador aprende a partir de uma experiência (E) com respeito a algumas classes de tarefas (T) e uma dada medida de desempenho (P) se o seu desempenho nas tarefas T, medidas por P, melhoram com a experiência E".

Normalmente, os algoritmos de aprendizado de máquinas utilizam-se da experiência adquirida anteriormente para auxiliar o processo de tomada de novas decisões para melhorar o desempenho. Em consonância com as características da experiência adquirida, tem-se quatro formas diferentes de aprendizado: supervisionado, semi-supervisionado, não-supervisionado e aprendizado por reforço. A diferença entre os três primeiros modos de aprendizado é a presença ou não do atributo classe, que faz a rotulagem dos dados fornecidos ao algoritmo, chamado de conjunto de treinamento.

No aprendizado supervisionado, os rótulos são conhecidos, objetivando-se encontrar uma hipótese capaz de classificar novas observações entre as classes já existentes. Por sua vez, no aprendizado não supervisionado, os rótulos não são conhecidos e tem-se o objetivo de tentar estabelecer a existência de grupos ou similaridade entre os grupos. Já no aprendizado semi-supervisionado, apenas alguns poucos pontos do conjunto de treinamento são rotulados, enquanto grande parte deles são não rotulados. O objetivo desse tipo de aprendizado é encontrar uma hipótese capaz de classificar novas observações entre as classes já existentes. Esse tipo é um meio termo entre o aprendizado supervisionado e o não supervisionado.

Por outro lado, no aprendizado por reforço não existem pares "entrada/saída" para serem utilizados no treinamento. Após realizar uma ação, o agente imediatamente

recebe uma recompensa, mas não fica sabendo qual deveria ser a melhor ação para atingir o objetivo (maximizar o retorno em longo prazo). Ele precisa obter experiência dos possíveis estados, ações, transições e recompensas do sistema para atingir a o desempenho ótimo. Pode-se dizer, também, que uma mesma ação realizada em um mesmo estado pode levar a diferentes estados, com diferentes valores de retorno.

Em geral, o conjunto de treinamento é representado por meio de uma estrutura conhecida como tabela atributo-valor. A Tabela 2 mostra essa estrutura para o aprendizado supervisionado.

Tabela 2 - Conjunto de exemplos no formato atributo-valor

	A_1	A_2	...	A_M	Classe (Y)
E_1	x_{11}	x_{12}	⋮	x_{1m}	y_1
E_2	x_{21}	x_{22}	⋮	x_{2m}	y_2
⋮	⋮	⋮	⋮	⋮	⋮
E_M	x_{n1}	x_{n2}	⋮	x_{nm}	y_n

E_i = Exemplo ; A_i = Atributos ; x_{ij} = Possível valor do atributo A_i ; y_i = Classe

Nessa dissertação foram utilizados diferentes algoritmos de aprendizado. A seguir, todos eles serão detalhados e citadas as suas principais características.

3.1. Aprendizado em Múltiplas Instâncias

O aprendizado em múltiplas instâncias (AMI) foi proposto por (DIETTERICH, LATHROP e LOZANO-PEREZ), no ano de 1997, no contexto de previsão de reações químicas de drogas, cujo objetivo era prever se uma molécula de droga candidata iria se ligar a uma proteína alvo. Sabe-se que as moléculas são flexíveis e podem assumir diversas formas, onde cada forma representa uma substância diferente, conforme ilustrado na Figura 4. Cada uma das formas da molécula pode apresentar diferentes propriedades.

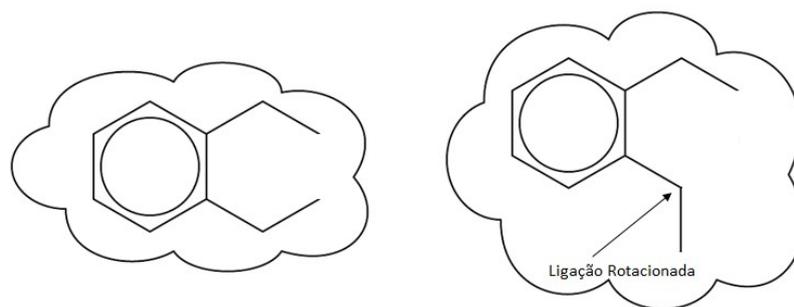


Figura 4 - Demonstração de molécula rotacionada
(adaptado de (DIETTERICH, LATHROP e LOZANO-PEREZ, 1997))

Desta forma, uma amostra positiva não representaria a forma específica que a molécula devesse possuir para se ligar à proteína, mas indicaria uma das possíveis formas que a molécula pudesse assumir. Em contrapartida, uma amostra negativa significaria que nenhuma das formas da molécula permitiria essa ligação. Esse cenário é considerado a primeira aplicação de AMI, onde o aprendizado dependia de múltiplas instâncias. Posteriormente, o AMI foi aplicado em diversas outras situações, tais como classificação de imagens (MARON e RATAN, 1998), classificação de texto, reconhecimento de voz, mineração de dados, entre outros (QUISPE, 2015).

O AMI pode ser considerado uma variação do aprendizado supervisionado Figura 5 (A), onde os rótulos são atribuídos a um conjunto de instâncias, conhecido por *bags*, e não a cada instância de forma individual.

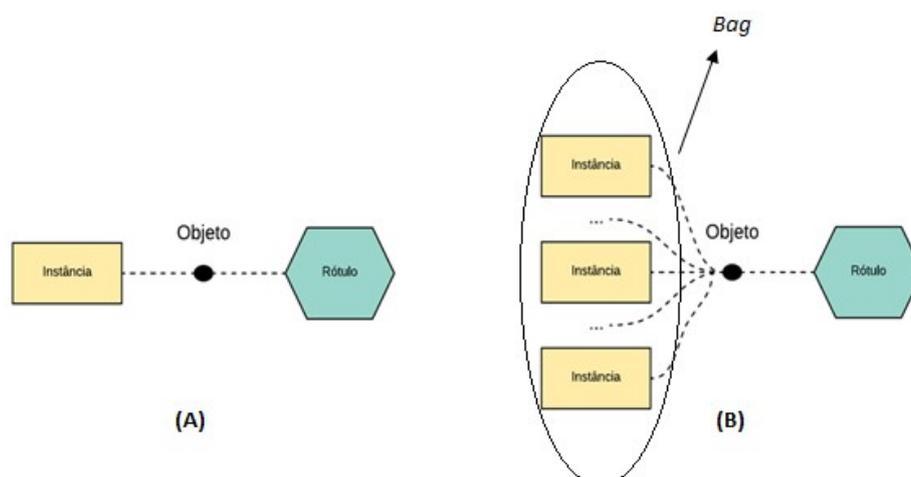


Figura 5 - (A) Aprendizado supervisionado (B) Aprendizado em Múltiplas Instâncias.

(Adaptado de (QUISPE, 2015))

Cada *bag* é rotulado como positivo ou negativo e cada um descrito como um conjunto de vetores. Essa variação de aprendizado foi chamada de aprendizado ambíguo, nos quais os exemplos não são completamente rotulados gerando, desse modo, ambiguidades nas classes das instâncias (MARON e RATAN, 1998).

A ideia central do AMI é a forma de rotulação dos *bags*: ele é rotulado como negativo (B_i^-) se todas as instâncias contidas nele forem negativas e rotulado como positivo (B_i^+) se, ao menos, uma das instâncias for positiva (DIETTERICH, LATHROP e LOZANO-PEREZ, 1997), (RAY e CRAVEN, 2005), (MARON, 1998), conforme ilustrado na Figura 6.

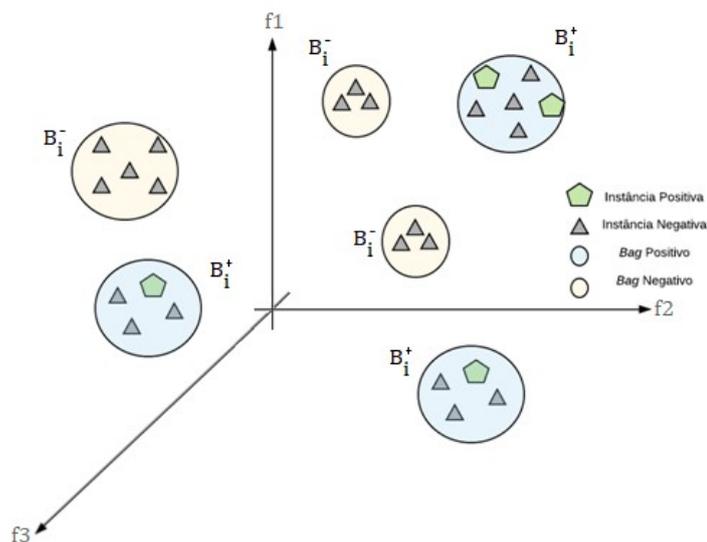


Figura 6 - Exemplo de bags e instâncias, sendo f1, f2 e f3 representam as três dimensões (atributos) de um exemplo.

(adaptado de (FARIA, 2016))

A Figura 7 (A) ilustra uma imagem que, ao ser analisada sobre a ótica do Aprendizado em Múltiplas Instâncias, pode ser considerada como sendo um *bag* positivo, devido ao fato de conter o objeto de interesse que, neste exemplo, é o cachorro. Desta maneira, um *bag* pode ser definido como uma coleção de instâncias ou regiões, cada uma delas representada por uma cor diferente.

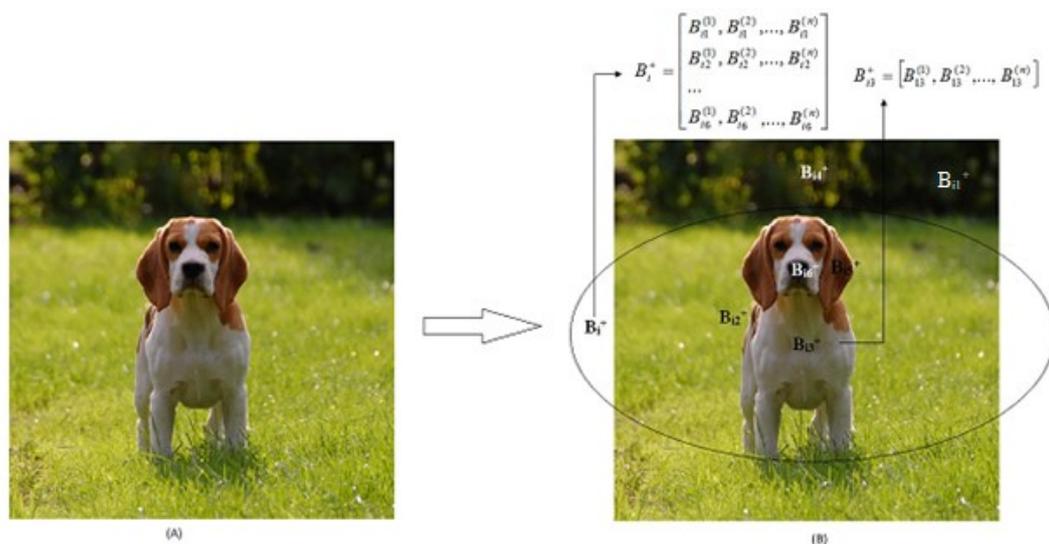


Figura 7 - Exemplo de *bag* e instância aplicado em uma imagem.

(adaptado de (FARIA, 2016))

A rotulação das regiões é distribuída da seguinte maneira: positiva - aquelas que contêm fragmentos (partes) do animal (3,5 e 6) e negativas - que são as que não contêm partes do animal (1 e 2).

Cabe ressaltar que cada uma das instâncias é formada por um vetor de características, sendo esses descritores da imagem. Ademais, especificamente neste exemplo, a rotulação das instâncias dentro do *bag* é conhecida, o que não ocorre, geralmente, nos casos reais.

Como forma de melhor compreender o problema do AMI, foi extraído de (DIETTERICH, LATHROP e LOZANO-PEREZ, 1997) o seguinte exemplo adaptado: “Suponha que um chaveiro profissional tenha sido contratado para abrir a porta da sala de suprimentos. Em um primeiro momento, ele não sabe qual a forma que a chave deve possuir para abrir aquela porta. Nesta empresa, existem diversos empregados e cada um deles possui um molho de chaves e estas chaves são capazes de abrir diversas portas. É de conhecimento dos funcionários que alguns deles possuem um privilégio especial, possuindo uma chave capaz de abrir esta porta específica, porém os funcionários não são cooperativos e ao invés de indicarem a chave certa para o chaveiro, eles apenas o entregaram o molho de chaves completa, devendo ele descobrir qual ou quais chaves destrancariam a porta da sala de suprimentos. Ainda, para piorar, o chaveiro não teria acesso direto a porta, o que o impediria de experimentar as chaves na fechadura de forma individual. Assim, ele apenas poderia inferir a chave certa pelo formato delas. Para resolver esse de problema é necessário descobrir qual ou quais chaves são comuns

em todos os molhos de chaves positivo (aquele que contem a chave que abre a porta desejada).”

Na literatura, existem diversos tipos de algoritmo que tratam do problema de Aprendizado em Múltiplas Instâncias. Na próxima sessão será detalhado o algoritmo utilizado nessa dissertação.

3.1.1. Densidade Diversa

O algoritmo de Densidade Diversa (DD), do inglês *Diversity Density*, foi proposto originalmente por (MARON e LOZANO-PEREZ, 1998). Nele e em (MARON, 1998) o algoritmo DD é descrito como uma medida de interseção dos *bags* positivos menos a união dos *bags* negativos.

A ideia central do algoritmo é encontrar um ponto conceito no espaço de características que seja próximo de, no mínimo, uma instância de cada *bag* positivo e distante das instâncias dos *bags* negativos.

Para detalhamento do algoritmo, adotam-se os *bags* positivos como B_i^+ e os *bags* negativos como B_i^- . Além disso, a instância j -ésima contida no *bag* é descrita como B_{ij}^+ , o valor da k -ésima característica desse ponto é descrito como B_{ijk}^+ e a instância negativa de um *bag* é representada por B_{ij}^- . Assumindo que um ponto conceito é um ponto único t , que é encontrado através da maximização da probabilidade $P(t|B_1^+, \dots, B_n^+, B_i^-, \dots, B_m^-)$, sobre todos os pontos $x_i \in \mathbb{R}^d$ no espaço de características.

Aplicando a regra de Bayes sobre a localização do ponto conceito, o que é semelhante a maximizar a probabilidade $P(t|B_1^+, \dots, B_n^+, B_1^-, \dots, B_m^-)$ e realizando suposições adicionais (FARIA, 2016),(MARON, 1998), obtém-se a seguinte equação:

$$\prod_{1 \leq i \leq n} P(B_i^+ | t) \prod_{1 \leq i \leq m} P(B_i^- | t) \quad (9)$$

A equação (9) define, de forma geral, a máxima Densidade Diversa. Também é necessário definir os termos em produtos para poder instanciá-los. De acordo com (FARIA, 2016), utilizando o modelo *noisy-or* (PARSONS e BIGHAM, 1996) para modelar a probabilidade de uma instância em um *bag* ser o ponto conceito t é dada por:

$P(x = t|B_i^+) = P(x = t|B_{i1}^+, B_{i2}^+, B_{i3}^+, \dots) = 1 - \prod_j(1 - P(x = t|B_{ij}^+))$. De forma análoga, $P(x = t|B_i^-) = P(x = t|B_{i1}^-, B_{i2}^-, B_{i3}^-, \dots) = \prod_j(1 - P(x = t|B_{ij}^-))$.

Assumindo, também, que cada instância em um *bag* possui um rótulo e que essas instâncias podem ser vistas como uma causa independente da rotulação do *bag*, então o modelo *noyse-or* pode ser aplicado. Dessa maneira, a probabilidade de um *bag* ser positivo é igual à probabilidade de que nem todas as instâncias dentro do *bag* sejam negativas. De forma análoga, a probabilidade de um *bag* ser negativa é igual à probabilidade de todas as instâncias dentro de um *bag* serem negativas.

Densidade Diversa utiliza a distância euclidiana como métrica para modelar a probabilidade associada à distância entre a instância em análise, no espaço de característica e o ponto conceito. Assim, se uma instância é próxima ao ponto conceito t , a probabilidade $P(x = t|B_{ij})$ é alta. A probabilidade de cada instância é definida pela equação (10).

$$P(x = t|B_{ij}) = \exp(-||B_{ij} - x||^2) \quad (10)$$

Onde:

$$||B_{ij} - x||^2 = \sum_k s_n^2 (B_{ijn} - x_n)^2 \quad (11)$$

n representa a n -ésima característica, B_{ijn} representa a n -ésima característica do ponto B_{ij} , x é o ponto conceito e S_n representa a contribuição (peso) de cada instância.

3.2. Algoritmo dos Vizinhos mais Próximos

Um método baseado em distância que é utilizado com frequência é o algoritmo dos vizinhos mais próximos (KNN), do inglês *k-nearest neighbors*. Este é um dos algoritmos de aprendizado de máquinas mais simples (FACELI, LORENA, *et al.*, 2011). A ideia central que está por trás desse algoritmo é que os objetos relacionados ao

mesmo conceito são semelhantes entre si, ou seja, um novo objeto é classificado com base nos exemplos do conjunto de treinamento que são próximos a ele.

Esse algoritmo é caracterizado como tendo um aprendizado preguiçoso (*lazy learner*), pois ele não tem a capacidade de aprender um modelo, mas sim memorizar os objetos do treinamento. Em contrapartida, uma vantagem é a de poder ser utilizado tanto para classificação quanto para regressão de maneira direta, além de requerer pouco esforço durante a etapa de treinamento.

O KNN tem variações de acordo com o número de vizinhos que serão considerados. Das variações possíveis, o mais simples é o 1-Vizinho mais próximo (ou 1-NN, do inglês *1-Nearest Neighbour*). Neste, cada objeto representa um ponto em um espaço definido pelos atributos. Assim, é possível calcular a distância entre cada dois pontos, através da métrica da distância euclidiana, dada pela equação (12).

$$d(X_i, X_j) = \sqrt{\sum_{i=1}^d (X'_i - X'_j)^2} \quad (12)$$

Na fase de treinamento, o algoritmo 1-NN memoriza os exemplos rotulados do conjunto de treinamento. Para classificar um exemplo cuja classe não é conhecida, é calculada, através da equação (12), a distância entre o vetor de valores do atributo e os exemplos rotulados. O rótulo de classe associado ao exemplo de treinamento mais próximo do exemplo de teste é utilizado para classificar o novo exemplo.

A Figura 8 apresenta um exemplo ilustrativo de aplicação do algoritmo 1-NN em duas classes. Nesse exemplo, considera-se um conjunto de dados onde objetos podem ser separados como "classe 1" ou como "classe 2". O ponto representado por "?" é o ponto de teste, ou seja, é o objeto a ser classificado. No espaço definido pelos atributos, e lançando mão da distância euclidiana, o objeto de treinamento mais próximo do objeto de teste pertence à classe 1, que é então atribuída ao objeto de teste.

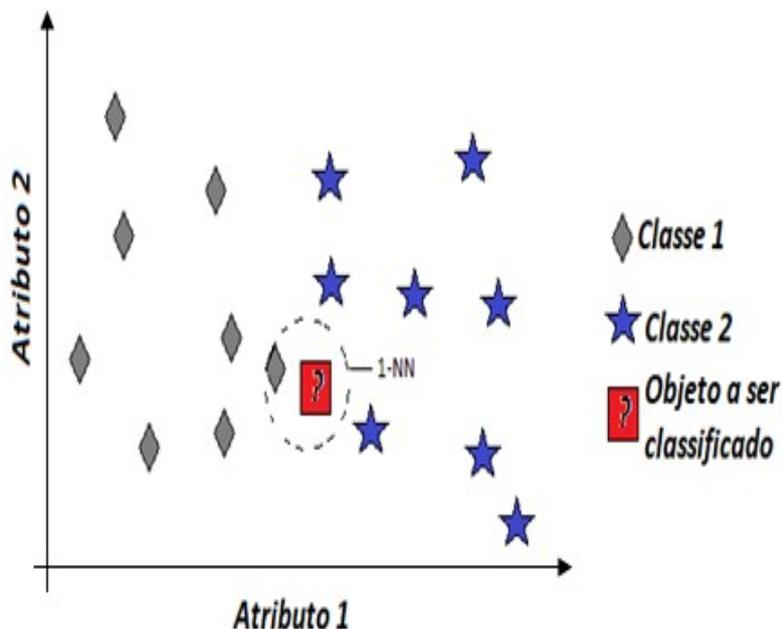


Figura 8 - Exemplo de classificação do método 1-NN
(Adaptado de (FACELI, LORENA, *et al.*, 2011))

Um dos problemas dos métodos baseados em distância é que eles têm seu desempenho afetado pela medida ou função de distância utilizada. Isso porque se pressupõe que os dados correspondem a pontos no espaço n -dimensional (R^n), ou seja, que seus atributos são numéricos, contínuos. Porém, há diversos problemas em que os dados possuem atributos qualitativos, necessitando fazer a codificação 1-de-N. Outro fator a ser observado no cálculo da distância entre objetos é a escala utilizada para os valores dos atributos, pois essas medidas são afetadas pela escala dos atributos. Por isso, a fim de minimizar esses efeitos, os atributos são usualmente normalizados.

Generalizando o algoritmo 1-NN tem-se o KNN que, diferentemente daquele, passa a considerar os K objetos do conjunto de treinamento mais próximos do ponto de teste x_t , onde K é um parâmetro do algoritmo, conforme ilustrado na Figura 9. Quando este valor é maior do que 1, para cada ponto de teste são considerados k vizinhos, em que cada um deles vota em uma classe. Essas previsões são agregadas de forma a classificar o ponto de teste. Essa agregação varia de forma diferente em problemas de classificação e regressão.

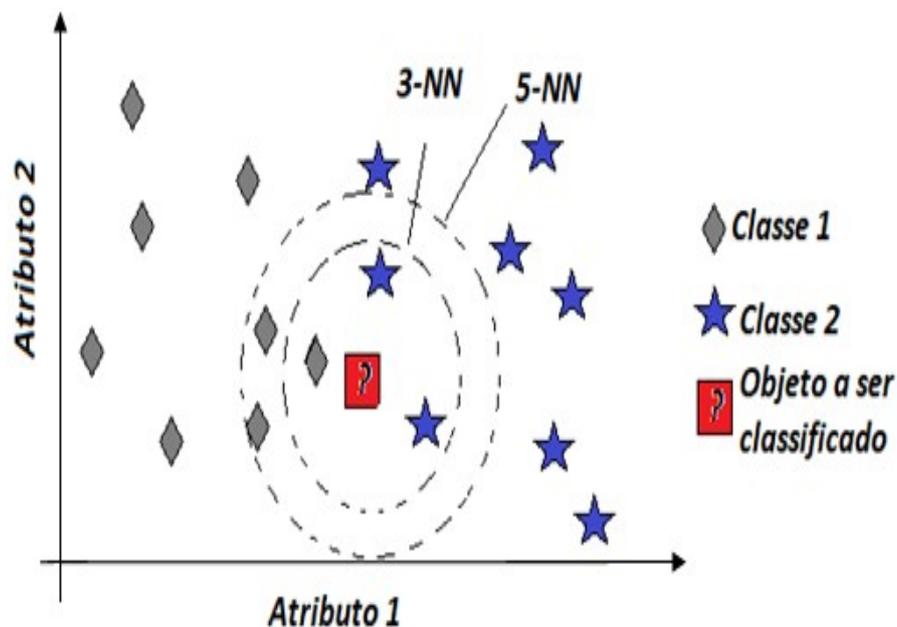


Figura 9 - Exemplo com 3 e 5 vizinhos, respectivamente

(Adaptado de (FACELI, LORENA, *et al.*, 2011))

De forma sucinta, em problemas de classificação em que a classe possui valores discretos, cada vizinho vota em uma classe e o objeto de teste é classificado na mais votada. Por sua vez, em problemas de regressão, podem ser usadas duas estratégias, dependendo do tipo da função de custo. Se a função de custo a ser minimizada for o desvio absoluto, deve ser utilizada a mediana, ao passo que se a função de custo a ser minimizada for o erro quadrático, deve ser utilizada a média dos valores obtidos para cada um dos k -vizinhos.

Utilizando o mesmo exemplo da Figura 8, o número de vizinhos mais próximos foi alterado para 3 e 5. Nota-se que para $k = 3$, o objeto de teste seria classificado como pertencente à classe 2, enquanto que para $k = 5$ o mesmo objeto seria classificado como classe 1. Percebe-se que a escolha do valor de k não é algo trivial, depende do usuário, varia de acordo com cada problema e influencia fortemente na classificação. Usualmente o valor de k é ímpar para evitar empates e deve ser avaliado para cada problema em particular.

A Tabela 3 indica, de forma sucinta, um pseudocódigo para o 1-NN.

Tabela 3 - Pseudocódigo da configuração 1-NN(Adaptado de (FACELI, LORENA, *et al.*, 2011))

Conjunto de dados para treinamento: $D = \{(x_i, y_i)\}, i = 1 \dots n$

Uma amostra que se deseja classificar: $a = \{x_a, y_a = ?\}$

Distância entre as instâncias: $d(x_i, x_j)$

Resultado da classificação da amostra a: y_a

$d_{min} \leftarrow +\infty$

Para $i = 1 : n$ faça

Se $d(x_i, x_j) < d_{min}$

$d_{min} \leftarrow d(x_i, x_j)$

$ind \leftarrow i$

Fim

Fim

$y_a = y_{ind}$

3.3. Árvore de Decisão

Árvores de decisão são modelos estatísticos que utilizam o aprendizado supervisionado para classificação e previsão de dados, lançando mão da estratégia de dividir para conquistar. Dessa forma, um problema complexo é dividido em problemas mais simples, aos quais recursivamente é aplicada a mesma estratégia a cada subproblema. (GAMA, 2004). As soluções dos subproblemas podem ser combinadas na forma de árvore, para resultar em uma solução do problema complexo. A força desse modelo vem da capacidade de dividir o espaço de instâncias em subespaços, onde cada subespaço é ajustado usando diferentes modelos (FACELI, LORENA, *et al.*, 2011).

Formalmente, os modelos em árvores são chamados de árvores de decisão, utilizado em problemas de classificação e árvores de regressão, em problemas de regressão. Em ambos, a interpretação dos modelos é semelhante, sendo adotada a nomenclatura de árvores de decisão, de uma forma genérica. Esta é um grafo acíclico direcionado em que cada nó ou é um nó de divisão, com dois ou mais sucessores, ou um nó folha.

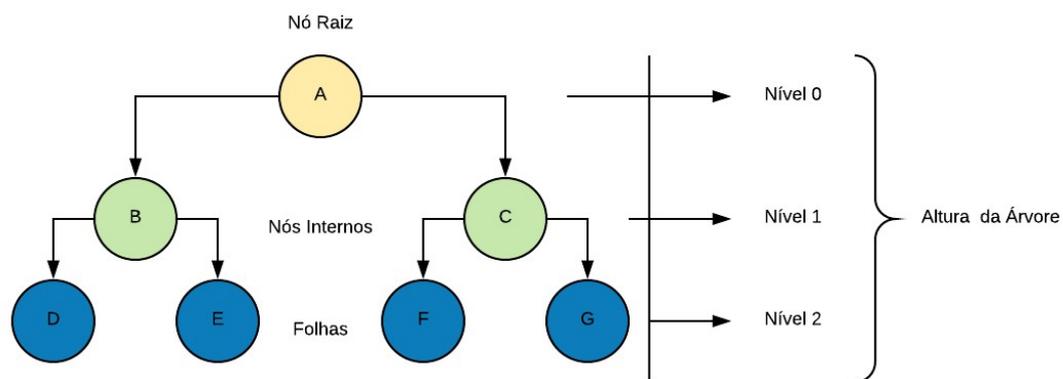


Figura 10 - Estrutura de árvore de decisão ¹

A árvore de decisão é o resultado de se fazer uma sequência ordenada de perguntas a cada passo, sendo que uma nova pergunta depende das respostas àquelas feitas anteriormente. O ponto de partida de uma árvore de decisão é chamado de nó raiz e consiste no conjunto de aprendizado e é localizado no topo da árvore. Um nó é um subconjunto do conjunto de atributos, que pode ser terminal, que se divide em nós filho, ou internos, conforme demonstrado na Figura 10. Por outro lado, o fim dessa sequência é uma previsão de classe. Tal divisão é determinada por uma condição sobre o valor de um único atributo e que divide os exemplos de acordo com as condições estabelecidas em outros nós. Cada exemplo do conjunto é finalizado em um nó terminal.

Uma árvore abrange todo o espaço de instâncias. Isso quer dizer que uma árvore pode fazer previsões para qualquer exemplo de entrada que esteja no espaço previsto de instâncias. O espaço de hipóteses das árvores de decisão enquadra-se no formalismo Forma Normal Disjuntiva (FND), onde os classificadores gerados por esse sistema codificam uma FND para cada classe. Para cada uma delas, as condições ao longo de um ramo (um percurso entre a raiz e uma folha) são conjunções de condições e os ramos são disjunções. Assim, cada ramo forma uma regra com uma parte condicional e uma conclusão. A parte condicional é formada por um conjunto de testes que envolvem um atributo particular, operador e um valor do domínio atributo. Esse tipo de teste corresponde a um hiperplano que é ortogonal ao eixo do atributo testado e paralelo aos outros eixos.

Uma vez definida a árvore de decisão, seu uso é imediato e rápido computacionalmente. Ademais, um ponto forte desse método é a sua interpretabilidade.

¹ Adaptado de: <http://slideplayer.com.br/slide/3322833/>

No entanto a sua construção, que é chamado de processo de indução, pode ser uma tarefa de alto custo computacional.

Para classificar um novo exemplo, basta começar pelo nó raiz, seguindo sempre cada nó de decisão, de acordo com o valor do atributo do novo exemplo, até que um nó folha seja encontrado. Dessa forma, a classificação é dada pela classe correspondente ao nó folha.

Na literatura, existem diversos métodos descritos para a divisão de uma árvore, mas a maior parte deles concorda que uma divisão que mantém a proporção de classes em todo o subconjunto não tem utilidade e uma divisão onde cada subconjunto tem somente exemplos de uma classe possui utilidade máxima (FACELI, LORENA, *et al.*, 2011). As medidas mais utilizadas para selecionar a melhor divisão são:

1. Medidas de diferença entre a distribuição do nó pai, ou seja, antes da divisão, e a distribuição nos subconjuntos obtidos por alguma função baseada na proporção de classe. Essas enfatizam a pureza dos subconjuntos. Quanto menor o grau de impureza, mais desbalanceada é a divisão das classes. Diz-se que o grau de impureza é nulo se todos os exemplos do nó pertencerem à mesma classe. De forma análoga, a impureza é máxima no nó se houver o mesmo número de exemplos para cada classe.
2. Medidas de diferença entre os subconjuntos divididos com base em alguma função sobre as proporções de classe. Essas enfatizam a disparidade dos subconjuntos.
3. Medidas estatísticas de independência entre as proporções de classe e os subconjuntos divididos. Essas medidas enfatizam o peso da evidência a confiança das predições da classe.

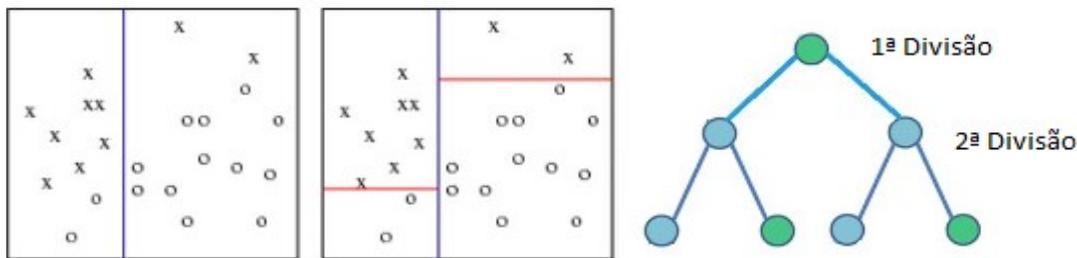


Figura 11 - Exemplo de divisão recursiva de uma árvore de decisão

(Adaptado de <https://www.quora.com/How-does-random-forest-work-for-regression-1>)

Além dessas, cita-se também o ganho de informação. O principal conceito do ganho de informação é a entropia. Ela mede a aleatoriedade (dificuldade de prever) do atributo alvo. A cada nó de decisão o atributo que permitir a maior redução da aleatoriedade de uma variável alvo será escolhido para dividir os dados, ou seja, os valores de um atributo definem as partições no conjunto de exemplo. Para cada atributo, o ganho de informação mede o quanto de redução houve na entropia das partições obtidas em consonância com o valor do atributo. Assim, supõe-se uma variável aleatória A , cujo domínio é $\{a_1, a_2, \dots, a_n\}$, onde a probabilidade de observar cada valor é p_1, p_2, \dots, p_n . A entropia A é dada pela equação (13):

$$A = \sum_i P_i \times \log_2 P_i \quad (13)$$

De forma mais genérica, o ganho de informação é calculado através da diferença entre o conjunto de exemplos e a soma ponderada da entropia das partições. A construção da árvore de decisão objetiva a redução da entropia, ou seja, reduzir a aleatoriedade da variável alvo.

Supondo a existência de duas classes, P e N , e um conjunto de exemplos S que contém p elementos da classe P e n elementos da classe n . A quantidade de informação necessária para decidir se um exemplo arbitrário pertence a P ou a N é definido pela equação (14):

$$H(p, n) = -\frac{p}{p+n} \log_2 \left(\frac{p}{p+n} \right) - \frac{n}{p+n} \log_2 \left(\frac{n}{p+n} \right) \quad (14)$$

Supondo que o atributo A foi selecionado, assumindo que o domínio de A tem v diferentes valores, a árvore resultante terá um conteúdo de informação esperado, descrito pela equação (15):

$$E(A, p, n) = \sum \frac{P_i + N_i}{P + N} H(p_i, n_i) \quad (15)$$

Onde P_i e N_i são o número de objetos de cada classe na subárvore associada com a partição i , baseada nos valores do atributo A . Desta maneira, o ganho de informação (G) obtido pode ser descrito pela equação (16):

$$G(A, p, n) = H(p, n) - E(A, p, n) \quad (16)$$

3.3.1. Florestas Aleatórias

A floresta aleatória, do inglês *Random Forest*, é uma técnica de aprendizado de máquina desenvolvida por (BREIMAN, 2001). Ela pode ser definida como sendo um classificador formado por um conjunto de árvores de decisões $\{h(x, \vartheta_k), k = 1, \dots\}$, onde ϑ_k são vetores aleatórios amostrados de forma independente e distribuídos igualmente em todas as árvores da floresta. O resultado do processo de classificação é a escolha da classe x com o maior número de votos dentre todas as árvores consideradas. A Figura 12 ilustra o funcionamento do algoritmo.

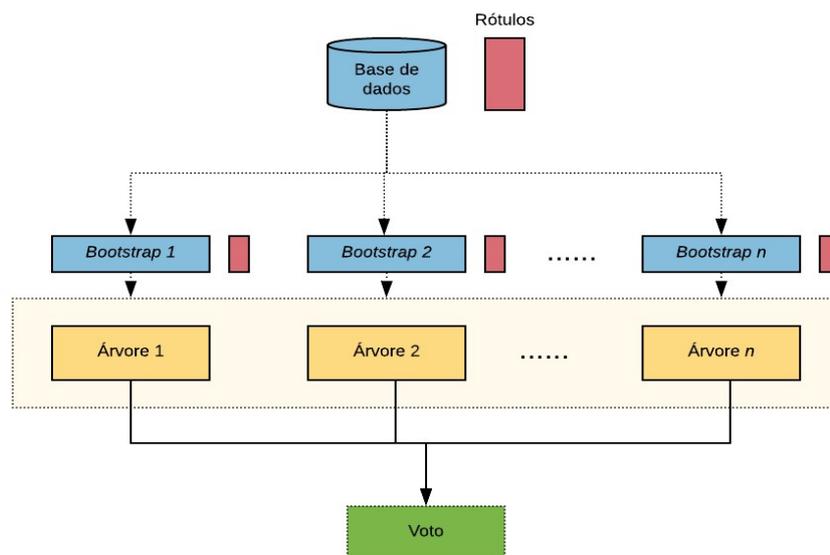


Figura 12 - Exemplo de Árvores Aleatórias

Os dados de entrada utilizados para a técnica de árvores aleatórias são selecionados, de forma aleatória, utilizando a técnica de *bootstrap*. Ela é utilizada, tipicamente, para estimar desvios-padrão, intervalos de confiança e valores críticos das amostras (SCHMIDHEINY, 2012). Nessa técnica, n subconjuntos são gerados a partir do conjunto de dados originais e os exemplos são amostrados aleatoriamente desse conjunto com repetição. Dessa maneira, é possível que um exemplo esteja presente em determinados subconjuntos de teste e o resultado é obtido pela média do desempenho de cada subconjunto de teste.

É comum adotar $n \geq 50$, pois a ideia básica é repetir o experimento diversas vezes para estimar o desempenho nos experimentos replicados. O *bootstrap* é um procedimento custoso e geralmente aplicado em pequenos conjuntos de dados (FACELI, LORENA, *et al.*, 2011).

Na sequência, para o desenvolvimento das árvores aleatórias, são realizados a seleção do conjunto de dados para o crescimento das árvores, a seleção das características dos preditores, o crescimento das árvores respeitando um critério de parada e, por fim, a estimativa do erro de crescimento, através do *out-of-bag* (IBAÑEZ, 2016)

O algoritmo *Random Forest* tem início com a seleção aleatória de um subconjunto Z , formado por amostras dos dados de treinamento com o total de p atributos. Em seguida, uma nova árvore T_b é construída em três etapas: seleção aleatória de m dos p atributos ($p \gg m$), escolha do melhor ponto de corte dentre os atributos

selecionados e divisão de um nó em dois nós filhos com base no ponto de corte escolhido. Esse procedimento é repetido para cada novo nó até alcançar o tamanho mínimo de nós (n_{min}). Com as árvores de decisão construídas, é possível configurar o algoritmo para regressão ou classificação (HASTIE, TIBSHIRANI e FRIEDMAN, 2008).

A vantagem desse algoritmo é que consegue lidar com muitos atributos de entrada mantendo sua rapidez na construção das redes. Como a escolha dos atributos de entrada é aleatória, as árvores construídas são descorrelacionadas. As desvantagens do método estão na sensibilidade a ruídos na base de dados e na dificuldade de interpretação do modelo (COSTA, 2012).

A Tabela 4 ilustra o pseudocódigo do *Random Forest*.

Tabela 4 - Pseudocódigo do *Random Forest*

(Adaptado de (HASTIE, TIBSHIRANI e FRIEDMAN, 2008))

Para $b = 1, \dots, B$ faça

 Seleciona um subconjunto Z com dados de treinamento

 Constrói uma árvore de decisão seguindo as três etapas:

1. Seleciona m atributos
2. Define o melhor atributo dentre m para ponto de corte
3. Divide o nó em dois nós filhos

Fim

Saída: Conjunto de árvores $\{T_b\}$

- Para classificação: Sendo $\hat{C}_b(x)$ a classe de um ponto x a ser classificado, tem-se:

$$\hat{C}_b(x) = \text{maioria dos votos } \{\hat{C}_b(x)\}_1^B$$

- Para regressão: $\hat{f}_{RF}^B(x) = \frac{1}{B} \sum_{b=1}^B T_b(x)$
-

3.4. Máquina de Vetores de Suporte

O algoritmo máquina de vetores de suporte, do inglês *support vector machine* (*SVM*) é, basicamente, uma máquina linear, onde objetiva-se construir um hiperplano como superfície de decisão, de forma que a margem de separação entre os objetos positivos e negativos seja máxima. Em conjuntos com padrão não linearmente separáveis, busca-se uma função de mapeamento φ apropriada, de modo a tornar o conjunto mapeado linearmente separável. Elas são embasadas pela teoria de aprendizado estatístico, desenvolvida por (VAPNIK, 1995), que estabelece uma série de princípios que devem ser observados na obtenção de classificadores com boa generalização.

A primeira formulação de SVM lida com problemas linearmente separáveis (GUYON, VAPNIK, *et al.*, 1992), de modo que fosse possível separar duas classes distintas sem perder generalidade, através de um hiperplano ótimo, resultante de um conjunto de treinamento linearmente separável (HAYKIN, 2001).

Considerando uma amostra de treinamento $\{(x_i, d_i)\}_{i=1}^N$, onde x_i é o padrão de entrada para o i -ésimo exemplo e d_i representa a resposta desejada, sendo $d_i = \{+1, -1\}$, representando as classes linearmente separáveis. A equação que separa as classes através de hiperplanos é definida pela equação (17), mostrada na Figura 13.

$$w^T \cdot x + b = 0 \quad (17)$$

Sendo x um vetor de entrada que representa os padrões do conjunto de treinamento, w um vetor de pesos ajustáveis e b um limiar, conhecido como bias. Assim:

$$\begin{cases} w^T \cdot x_i + b \geq 0 & \text{Para } d_i = +1 \\ w^T \cdot x_i + b < 0 & \text{Para } d_i = -1 \end{cases} \quad (18)$$

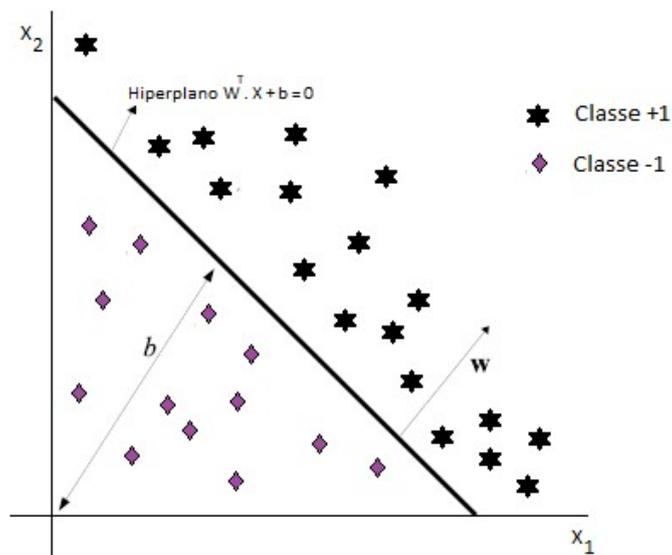


Figura 13 - Exemplo de hiperplano na classificação SVM
(Adaptado de (TAKAHASHI, 2012))

O objetivo da SVM é encontrar um hiperplano que seja capaz de separar o conjunto de treinamento sem erro e com uma margem de separação (ρ_0) máxima (HAYKIN, 2001). Sob esse ponto de vista, a superfície de decisão é conhecida como hiperplano ótimo. Para um espaço de entrada bidimensional, o hiperplano é ilustrado na Figura 14.

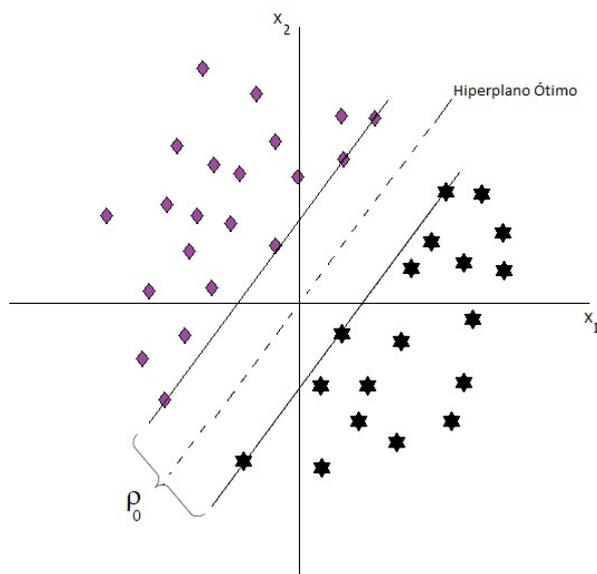


Figura 14 - Hiperplano ótimo para conjunto de dados linearmente separáveis
(Adaptado de (HAYKIN, 2001))

Sejam w_0 e b_0 valores ótimos para o peso e bias, respectivamente. Desta forma, o hiperplano ótimo representa uma superfície de decisão linear e é definido pela equação (19).

$$w_0^T \cdot x_0 + b_0 = 0 \quad (19)$$

E cuja função discriminante é definida pela equação (20) e fornece uma distância r de x até o hiperplano (w_0, b_0) , e pode ser representado pela equação (21).

$$g(x) = w_0^T \cdot x + b_0 \quad (20)$$

$$x = x_p + x_n \quad (21)$$

Na equação (21), x_p representa a projeção normal de x sobre o hiperplano ótimo e x_n é o vetor normal com a distância r , que pode ser reescrito de acordo com a equação (22).

$$x_n = r \cdot \frac{w_0}{\|w_0\|} \quad (22)$$

É possível perceber que r é positivo se x estiver no lado positivo do hiperplano ótimo e negativo caso esteja do lado negativo do hiperplano ótimo, conforme é possível verificar na Figura 15. Assim, por definição, considerando $g(x_p) = 0$ resulta na equação (23).

$$g(x) = w_0^T \cdot x + b_0 = r \|w_0\| \therefore r = \frac{|g(x)|}{\|w_0\|} \quad (23)$$

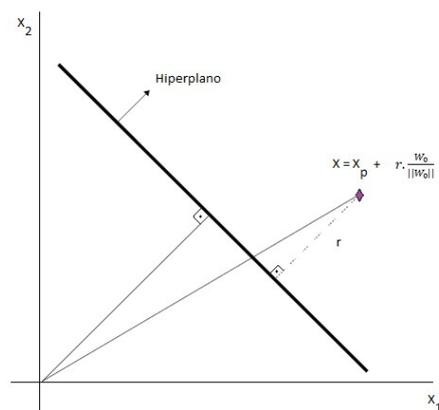


Figura 15 - Interpretação geométrica da distância X até o hiperplano ótimo para um padrão de entrada bidimensional

(Adaptado de (TAKAHASHI, 2012))

Para o conjunto de treinamento ser linearmente separável, ele deve satisfazer a seguinte restrição, onde w_0 e b_0 são obtidos somente no conjunto de treinamento:

$$\begin{cases} w_0^T \cdot x_i + b \geq 1 & \text{Para } d_i = +1 \\ w_0^T \cdot x_i + b \leq -1 & \text{Para } d_i = -1 \end{cases} \quad (24)$$

A equação (24) pode ser reescrita:

$$d_i(w_0^T \cdot x_i + b_0) \geq 1 \quad (25)$$

Dessa forma, se a equação (18) for válida, ou seja, caso os padrões sejam linearmente separáveis, pode-se sempre escalar w_0 e b_0 de modo que a equação (24) seja válida (HAYKIN, 2001). O escalamento não afeta a equação (19) e os pontos que satisfazem (x_i, d_i) para os quais uma das linhas da equação (24) é satisfeita com o sinal de igualdade são chamados de vetores de suporte, de onde se origina o nome “máquina de vetor de suporte”. São esses pontos que influenciam diretamente na localização do hiperplano ótimo com margem máxima, pois são esses os pontos mais próximos da superfície de decisão.

Considerando um ponto $x^{(j)}$ um vetor-suporte de uma classe positiva $d^{(w)} = +1$. Logo, por definição:

$$g(x^{(j)}) = w_0^T \cdot x^{(j)} + b_0 - 1, \quad \text{Para } d_i = +1 \quad (26)$$

Da equação (23) é possível perceber que a equação da distância do vetor de suporte $x^{(j)}$ até o hiperplano ótimo é dada por:

$$r = \frac{w_0^T \cdot x^{(j)} + b_0}{\|w_0\|} = \begin{cases} + \frac{1}{\|w_0\|}, & \text{se } d^{(w)} = +1 \\ - \frac{1}{\|w_0\|}, & \text{se } d^{(w)} = -1 \end{cases} \quad (27)$$

Da equação (27) infere-se que o sinal positivo indica que $x^{(j)}$ pertence ao lado positivo do hiperplano ótimo, ao passo que o negativo indica o contrário. Sendo ρ a máxima margem de separação entre duas classes de um conjunto de treinamento, então:

$$\rho = 2r = \frac{2}{\|w_0\|} \quad (28)$$

Observa-se que a equação (28) mensura a distância entre os hiperplanos da equação (25), de mesma forma que a distância entre os hiperplanos $w^T \cdot x + b = 0$ e $w^T \cdot x + b = -1$ é dada por $\frac{1}{\|w_0\|}$. Assim, percebe-se que a minimização de $\|w\|$ leva a uma maximização da margem.

A otimização do hiperplano é resolvida encontrando os valores ótimos de peso w e bias b , dada a amostra de treinamento $\{(x_i, d_i)\}_{i=1}^N$, que satisfaçam as restrições:

$$d_i(w^T \cdot x_i + b) \geq 1 \text{ para } i = 1, 2, \dots, N \quad (29)$$

Este pode ser resolvido minimizando a função custo φ em relação ao vetor de pesos w , satisfazendo a equação (29):

$$\varphi(w) = \frac{1}{2} w^T \cdot w \rightarrow \min \|\varphi(w)\| = \min \|w\|^2 \quad (30)$$

As SVMs lineares, também chamado de SVM com margens rígidas, são eficazes na classificação de conjuntos de dados linearmente separáveis. Porém, há diversos casos em que não é possível dividir os dados de treinamento por um hiperplano sem que haja erros de classificação. O objetivo da SVM é encontrar um hiperplano que minimiza a probabilidade desses erros de classificação.

Há casos onde não é necessário fazer um mapeamento do espaço de características no conjunto de treinamento. Assim, esses casos são tratados pela SVM com margens de separação entre classes suaves ou flexíveis, conhecidas também como *soft*, pois pode haver pontos (x_i, y_i) que violarão a condição que fora estabelecida em (29). As violações possíveis são descritas a seguir:

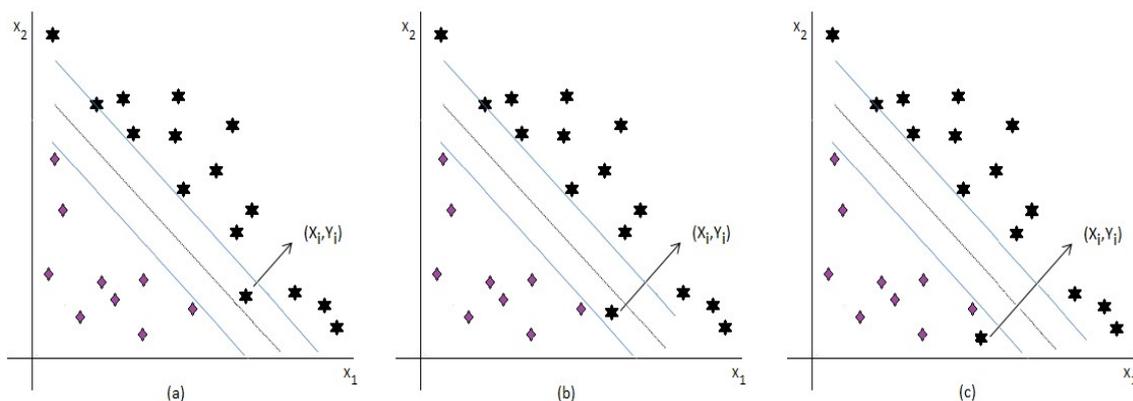


Figura 16 - Situações possíveis para o SVM

(Adaptado de (TAKAHASHI, 2012))

1. O ponto (x_i, y_i) , ilustrado na Figura 16 (a), está dentro da região de separação e no lado correto da superfície de decisão. Neste caso, houve uma escolha equivocada do hiperplano;
2. O ponto (x_i, y_i) , ilustrado na Figura 16 (b), está dentro da região de separação e no lado incorreto da superfície de decisão. Assim, houve uma escolha equivocada do hiperplano de margem maior;
3. O ponto (x_i, y_i) , ilustrado na Figura 16 (c), está fora da região de separação e no lado errado da superfície de decisão.

Para ser possível o tratamento do problema com casos contendo pontos de dados não separáveis, introduziu-se uma variável não negativa $\{\xi_i\}_{i=1}^N$ na definição do hiperplano de separação, como se segue:

$$d_i(w^T \cdot x_i + b) \geq 1 - \xi_i \quad (31)$$

Essas variáveis são conhecidas como *variáveis soltas* e têm como objetivo medir os desvios dos pontos (x_i, y_i) para a condição ideal de separação das classes. Assim, para $0 \leq \xi_i \leq 1$, o ponto de dados encontra-se dentro da região de separação, mas no lado correto da superfície de decisão, conforme ilustrado na Figura 16(a). Por outro lado, quando $\xi_i > 1$, o ponto encontra-se do lado incorreto do hiperplano de separação, conforme ilustrado na Figura 16 (b).

Os vetores de suporte são aqueles pontos de dados que satisfazem a equação (31), mesmo se $\xi_i > 0$. Dessa forma, caso um exemplo com $\xi_i > 0$ seja deixado de fora do conjunto de treinamento, a superfície de decisão permanecerá inalterada, de modo que os vetores de suporte são definidos da mesma forma para casos linearmente separáveis como para casos não separáveis.

Existem alguns casos onde não é possível dividir adequadamente os padrões dos conjuntos de treinamento por um hiperplano, mesmo após o ajuste das variáveis soltas. Assim, um mapeamento do domínio do espaço de entrada do conjunto de treinamento para um novo espaço, usando uma função *kernel* apropriada. Esta seria uma função que recebe dois pontos x_i e x_j do espaço de entrada e computa o produto escalar $\phi^T(x_i) \cdot \phi(x_j)$ no espaço de características.

A Figura 17 mostra o processo de transformação de um espaço de características não linearmente separável em um problema linearmente separável, através do aumento de dimensões, consequência do mapeamento feito por uma função *kernel*.

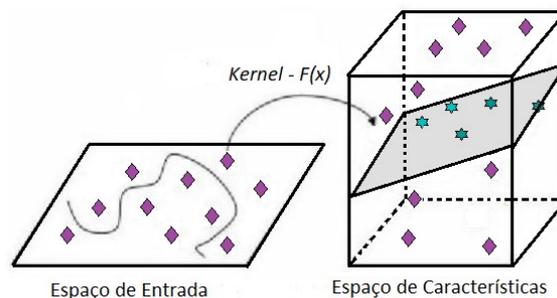


Figura 17 - Mapa não linear do espaço de entrada para o espaço de características
(Adaptado de (SANTOS, 2012))

O termo $\phi^T(x_i) \cdot \phi(x_j)$ é a representação do produto interno dos vetores x_i e x_j , com o *kernel* representado por:

$$k(x_i x_j) = \phi^T(x_i) \cdot \phi(x_j) \quad (32)$$

Cabe ressaltar que as funções ϕ devem pertencer a um domínio em que seja possível calcular os produtos internos, ou seja, tem que respeitar o *teorema de Mercer*. Este diz que os *kernels* devem ser matrizes positivas e definidas, ou seja, $k_{ij} =$

$k(x_i, x_j)$, para $i, j = 1, 2, \dots, N$, devem ser autovetores maiores que 0. Alguns *kernels* mais comuns são: RBF (*Radial Basis Function*), polinomiais e gaussiana.

A obtenção de classificadores por meio de SVM envolve a avaliação de um *kernel* apropriado, bem como a obtenção de parâmetros dessa função e do algoritmo de determinação do hiperplano ótimo. Essas escolhas afetam, significativamente, o desempenho do classificador através da superfície de decisão.

O algoritmo SVM possui a vantagem de apresentar boa capacidade de generalização, além de ser robusto diante de objetos de grande dimensão e poder ser usado em problemas de regressão. Também, por ter a sua função objetivo convexa, possui um único mínimo global. Por outro lado, o algoritmo é sensível à escolha de valores de parâmetros. Pode-se dizer, também, que o algoritmo é pouco interpretável e só consegue lidar com conjuntos de treinamento com dados numéricos.

3.5. Adaboost

O Adaboost, do inglês *Adaptive Boosting*, foi introduzido em 1995 por (FREUND e SCHAPIRE, 1995) como forma de resolver diversos problemas encontrados no algoritmo de *Boosting*. Ele apresenta propriedades que facilitam sua implementação e seu uso, dentre as quais é possível citar o baixo custo operacional pelo fato de envolver apenas programação linear, através do uso de funções lineares (FREUND e SCHAPIRE, 1999), diferentemente de outros métodos, que envolvem diferentes tipos de funções, além de ser capaz de fazer análises de dados de grandes dimensões. Também pode ser destacada a possibilidade de ser utilizado em diversas áreas e de ser combinado com outros classificadores.

Esse algoritmo tem por objetivo criar um classificador forte (alta acurácia) através da combinação de vários outros classificadores fracos (baixa acurácia). Assim, a cada iteração são feitos ajustes que aumentam a probabilidade dos pontos classificados erroneamente apareçam no próximo conjunto de treinamento. A Figura 18 ilustra o funcionamento do algoritmo.

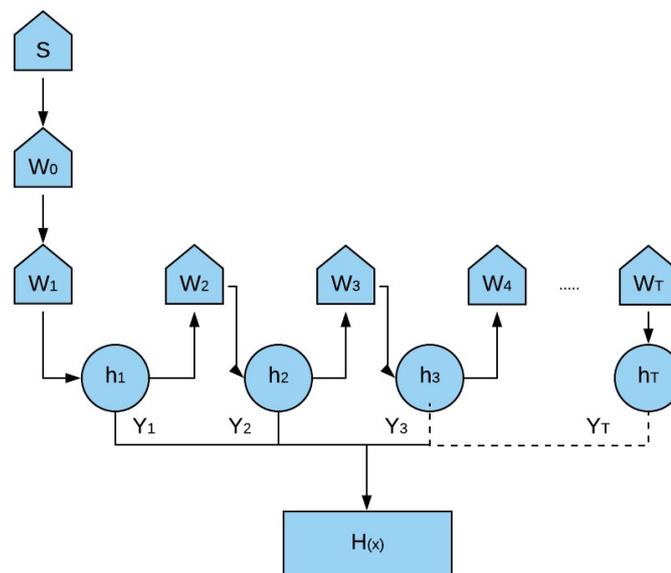


Figura 18 - Esquema simplificado do funcionamento do Adaboost

(Adaptado de (CHAVES, 2011))

Onde:

S representa o conjunto inicial;

w_0 Representa a normalização inicial;

$w_1 - w_T$ representam a distribuição de probabilidade aos exemplos de treinamento na iteração T ;

$h_1 - h_T$ são os classificadores gerados

$H_{(x)}$ é a hipótese final.

O algoritmo de aprendizado do *Adaboost* tem como entrada um conjunto de treinamento na forma $S = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, que são usadas para formular hipóteses e atualizar o vetor de pesos w_t . Cada x_i representa um vetor de atributos, cada y_i representa os possíveis grupos de classificação associados ao x_i e n representa o número total de amostras de treinamento.

O erro de treinamento é utilizado para ajustar as amostras de treinamento à hipótese h . Esse erro é dado pela razão entre o número de amostras classificadas de forma incorreta e o número total de amostras do conjunto, conforme a equação (33).

$$\varepsilon(h) = \frac{1}{n} \sum_{i=1}^n 1 \{h(x_i) \neq y_i\} \quad (33)$$

Sendo $1 \{.\}$ uma função que retorna 1 em caso de acerto e 0 caso contrário.

A cada ciclo de aprendizagem o *Adaboost* gera uma hipótese baseada na atualização dos pesos, tendo como objetivo priorizar a correta classificação dos dados que apresentam maiores pesos associados. Assim, diz-se que o objetivo desse algoritmo é gerar uma hipótese que minimize o erro de treinamento $\varepsilon(h)$. A atualização dos pesos é mostrada na equação (34):

$$D_{t+1}(i) = \frac{D_t(i) \exp(-\alpha_t y_i h_t(x_i))}{\sum_{i=1}^n D_t(i)} \quad (34)$$

Espera-se que as amostras de treinamento sejam geradas a partir da mesma distribuição D , em pares, no formato (x,y) . Assim, o erro de teste esperado de uma hipótese h é conhecido como *erro de generalização* (ε_t), sendo igual à probabilidade de classificar uma única amostra (x,y) de forma incorreta, de acordo com a equação (35). Para qualquer hipótese h a medida monitorada será uma soma de eventos de erro.

$$\varepsilon_t(h) = P_{(x,y)} \sim w_i [h(x) \neq y] \quad (35)$$

Por fim, os pesos anteriormente encontrados são revistos com o objetivo de aumentar os que são relacionados aos dados classificados de forma incorreta, iniciando-se uma nova rodada de testes. Assim, ao se realizar T rodadas, o *Adaboost* combina todas as hipóteses intermediárias h_t com o objetivo de gerar uma hipótese final $H(x)$, onde:

$$H(x) = \text{sign} \left(\sum_{t=1}^T \alpha_t h_t(x) \right) \quad (36)$$

Sendo:

$$\alpha_t = \frac{1}{2} \ln \left[\frac{1 - \varepsilon_t}{\varepsilon_t} \right] \quad (37)$$

Diz-se que a importância de um classificador fraco h_t é definida pela equação (37), conforme descrito na Figura 19:

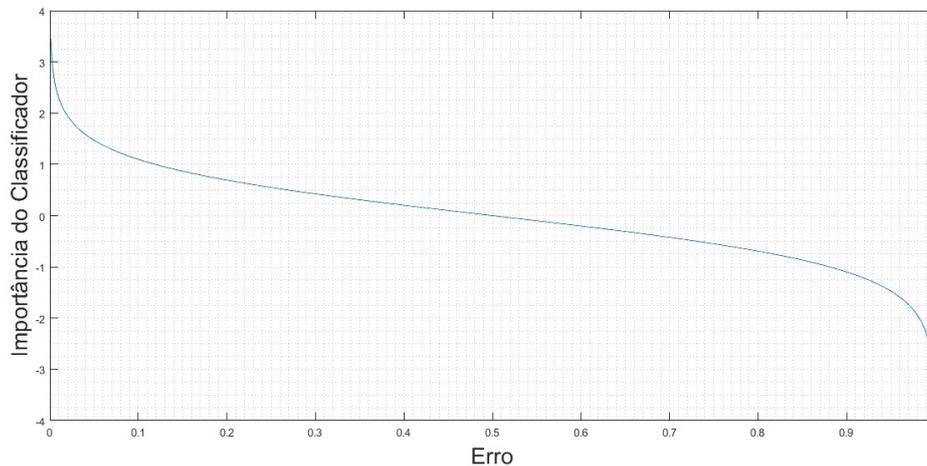


Figura 19 - Taxa de erro versus importância do classificador

(Adaptado de (CHAVES, 2011))

De acordo com a Figura 19, nota-se que quanto menor for a taxa de erro (ε_t), maior será a importância atribuída ao classificador. A

Tabela 5 ilustra um pseudocódigo para o algoritmo *Adaboost*.

Tabela 5 - Pseudocódigo do algoritmo Adaboost

(Adaptado de (SCHAPIRE, 2013))

Conjunto de dados para treinamento: $S = \{(x_1, y_1), \dots, (x_n, y_n)\}$, $y_i \in \{-1, +1\}$

Inicializa: $D_t(x_i) = \frac{1}{n}$ para $i=1, \dots, n$

Para $t = 1, \dots, T$ faça

Treina o classificador através de D_t

Calcula as hipóteses h_t

Seleciona h_t com menor erro ε_t

Escolha do α_t

Atualiza o valor de D_t para $i = 1, \dots, T$

Fim

- Saída $H(x)$
-

Como vantagem do algoritmo pode-se citar o baixo custo computacional, sua facilidade de implementação, flexibilidade de uso em diversas áreas de conhecimento, além de utilizar classificadores simples que, sucessivamente, vão se especializando em acertar a classificação que os classificadores anteriores fizeram de forma equivocada. Sua desvantagem é que o algoritmo pode conduzir ao *overfitting*, pois o algoritmo é sensível a dados ruidosos, também chamados de *outlier* (DIETTERICH, 2000). Isso significa que o algoritmo tende a enfatizar os pontos mais difíceis de serem classificados, podendo ocasionar um grande erro de generalização.

4. Modelo Proposto

Neste capítulo foram detalhados os algoritmos implementados cujo objetivo era realizar a correta classificação dos pacientes portadores da Esclerose Sistêmica utilizando os algoritmos de Aprendizado de Máquinas tradicionais 1-*Nearest Neighbor* (1-NN), *Adaboost* (ADAB), *Random Forest* (RF), *Radial Support Vector Machine* (RSVM) além do Aprendizado em Múltiplas Instâncias (AMI), por meio dos parâmetros fornecidos pela TOF.

Para tanto, foram utilizados o *software* Matlab R2012a e duas toolboxes desenvolvidas para esse ambiente: PRTOOLS (*Pattern Recognition Tools*)¹ e MIL (*Multiple Instance Learning*)², sendo ambas disponíveis na internet para ampla utilização.

O modelo proposto é descrito nos itens a seguir, de acordo com seu fluxograma resumido mostrado na Figura 20.

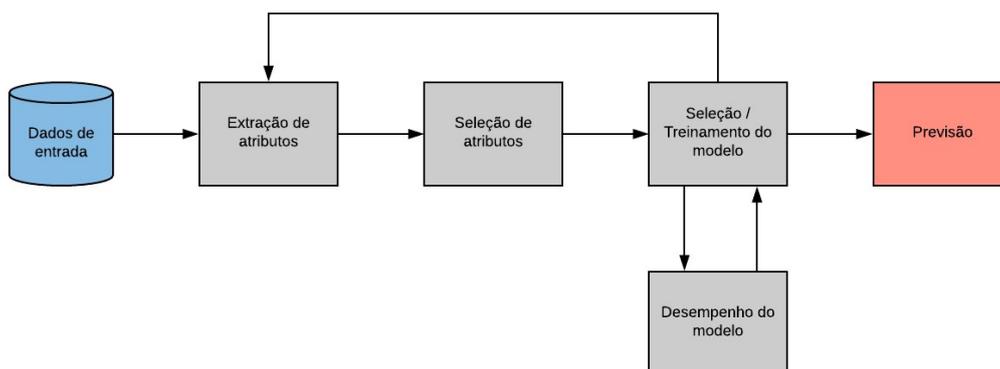


Figura 20 - Fluxograma do modelo proposto

¹ Prtools: *Toolbox for Pattern Recognition*. Disponível em: <<http://prtools.org/>>, Acessado em: 12/04/2018.

² Disponível em: <<http://www.37steps.com/prhtml/prtools/rbsvc.htm>>, Acessado em: 12/04/2018

4.1. Dados de Entrada

O conjunto de dados foi obtido através de ensaios realizados pelo sistema de aquisição de dados da TOF desenvolvido no Laboratório de Instrumentação Biomédica da UERJ (LIB - UERJ). As medições foram realizadas no Laboratório de Prova de Função Pulmonar, do Hospital Universitário Pedro Ernesto (HUPE) e também no LIB, aprovadas pelo comitê de ética do HUPE.

Para a obtenção desses dados, os indivíduos deviam permanecer sentados, respirando por um bocal acoplado ao sistema da TOF, onde eram aplicadas oscilações de pressão na faixa de 4-32 Hz, com incrementos de frequência de 2 Hz, com uma amplitude de aproximadamente 1 cmH₂O que eram produzidas por um autofalante. O sinal de pressão captado pelo transdutor foi transformado em sinal de tensão, de onde foi possível retirar todas as variáveis de interesse, descritas ao longo do capítulo 3.

Durante os exames para a obtenção dos dados, foram realizados 3 ensaios para cada indivíduo de aproximadamente 16 segundos cada, com intervalo de 1 minuto entre eles. Para tanto, eles permaneceram sentados, acoplados ao aparelho, ventilando espontaneamente, antes do início das manobras, para que fosse possível a adaptação ao equipamento. Posteriormente, permaneceram com a postura ereta com cabeça em posição neutra, utilizando um clipe nasal para não haver fuga de ar, sustentando as bochechas com as mãos para minimizar o efeito de dispersão dos sinais oscilatórios (MELO, WERNECK e GIANELLA-NETO, 2000(a)), (MIRANDA, 2010).

4.2. Extração de Atributos

Os dados da TOF foram, inicialmente, submetidos a um processo de normalização *Z-Score*. Essa foi necessária pelo fato das diferentes variáveis possuírem grandezas distintas. Assim, para amenizar os efeitos causados por elas, a normalização atuou realizando uma padronização nos atributos. Essa é uma técnica bastante utilizada na estatística e é representada pela Equação (38).

$$Z = \frac{x - \bar{x}}{s} \quad (38)$$

Na Equação (38), \bar{x} representa a média das amostras, e s é o desvio padrão dos dados. Nesta normalização, a nova variável z tem média zero e variância unitária.

Conforme será indicado no processo de treinamento, em primeiro lugar obtém-se a média e o desvio padrão do conjunto de treinamento. Uma vez que se dispõe desses valores, a normalização pode ser aplicada ao conjunto de treinamento e de teste durante a validação cruzada.

4.3. Seleção de Atributos

O objetivo da seleção de atributos de entrada é encontrar o menor número de características relevantes e que sejam capazes de descrever a doença e resultar em um desempenho satisfatório (GUYON e ELISSEEFF, 2003), (AMARAL, LOPES, *et al.*, 2012).

Outras motivações importantes para executar a seleção de atributos são: redução geral de dados, aumento da velocidade do algoritmo e ganhar conhecimento sobre o processo que gera os dados para permitir a visualização deles (2D ou 3D) (GUYON e ELISSEEFF, 2003), (AMARAL, LOPES, *et al.*, 2012). Ela também é importante porque contribui para a redução do número de parâmetros do modelo, reduzindo sua complexidade, conseqüentemente reduz a possibilidade de *overfit* em um conjuntos de dados de tamanho limitado.

A seleção de atributos pode ocorrer de duas maneiras distintas. A primeira forma seria através de um especialista capaz de indicar os principais descritores da doença, visando otimizar o desempenho do algoritmo e ganho de informação. A segunda maneira seria feita através de algoritmos, a saber: aplicação de filtros, busca pela abordagem *wrapper* e de métodos cuja seleção está embutida no processo de treinamento. O método de filtros realiza uma classificação ordenada dos atributos antes que estes sejam submetidos ao algoritmo (AMARAL, LOPES, *et al.*, 2012). Já a abordagem *Wrapper* utiliza algoritmos de aprendizado de máquinas para encontrar o

melhor subconjunto de atributos criados a partir do conjunto de dados original (HORTA, CARVALHO, *et al.*, 2010). O melhor subconjunto encontrado tem seus atributos selecionados para o classificador. Por outro lado, os métodos embutidos executam a seleção de atributos durante o treinamento e, em geral, são específicos para algum método de treinamento (AMARAL, LOPES, *et al.*, 2012).

Nessa etapa, foi utilizado o método de aprendizado em múltiplas instâncias como um seletor de atributos, onde os atributos que forneceram maiores valores de AUC atuaram como entradas para a avaliação nos demais algoritmos de aprendizado. É importante ressaltar que o especialista também forneceu um subconjunto de parâmetros. Além disso, também foram avaliados os resultados pela seleção de atributos através da função *featsel* disponível na *toolbox* do PRTools do software Matlab.

4.4. Seleção e Treinamento do modelo

A validação cruzada com k pastas foi usada durante o treinamento dos modelos. Essa técnica permite estimar o desempenho do modelo com dados não vistos anteriormente. Também permite comparar o desempenho de dois ou mais algoritmos diferentes.

Essa técnica foi utilizada devido ao tamanho do conjunto de dados, que é relativamente pequeno e, para esses casos, a validação cruzada fornece uma boa estimativa do erro de generalização do classificador (WITTEN e FRANK, 2005).

A ideia central deste método é dividir o conjunto de dados, aleatoriamente, em k partes iguais (ou aproximadamente iguais), segmentadas em pastas. Assim, o treinamento é realizado k vezes, onde $k-1$ pastas serão utilizadas para o treinamento do modelo e as demais para teste (REFAEILZADEH, TANG e LIU, 2009). Ao final dos experimentos é feita uma média dos resultados de cada teste, evitando que um resultado com uma medida de desempenho eventualmente muito alta ou muito baixa, devido a uma divisão específica dos dados em conjunto de treino e teste, represente o desempenho do classificador

Neste trabalho é feita a validação cruzada ($k = 10$) no conjunto de dados, de acordo com o esquema ilustrado na Figura 21.

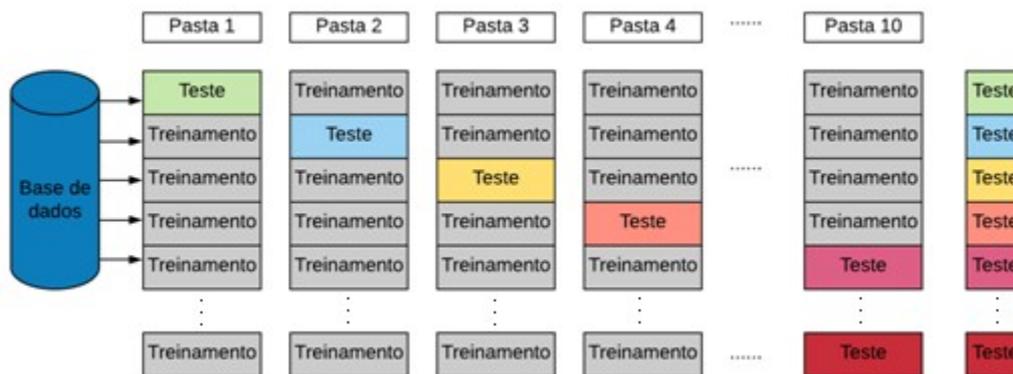


Figura 21 - Esquema ilustrativo da Validação Cruzada com 10 pastas

(Adaptado de: <http://blog-test.goldenhelix.com/wp-content/uploads/2015/04/B-fig-1.jpg>)

A métrica utilizada para a seleção dos classificadores e suas respectivas configurações estão descritas nos itens 4.5 e 4.6.

4.5. Medida de Desempenho do Modelo

A curva ROC (*Receiver Operating Characteristics*) é comumente utilizada para visualizar, organizar e selecionar classificadores baseados em sua performance. Ela teve origem na teoria de decisão da estatística e foi desenvolvida entre os anos de 1950 e 1960 para detectar sinais de radar.

Posteriormente, seu uso foi difundido na área de diagnóstico médico, onde há vasta literatura (ZOU, O'MALLEY e MAURI, 2007), pois, raramente, observa-se uma perfeita separação entre os grupos de duas populações, uma portadora da doença e outra não portadora da doença. Normalmente, há uma sobreposição entre as duas curvas que representam esses grupos (CASTANHO, YAMAKAMI e BARROS, 2004), conforme ilustrado na Figura 22.

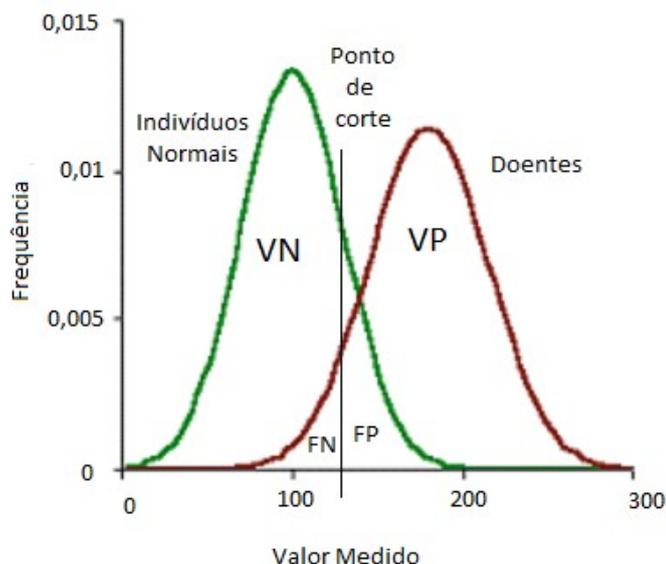


Figura 22 - Resultados de um teste em dois grupos (doente e saudável)

(Adaptado de: <http://www2.fm.usp.br/dim/bayes/index.php>)

Assim, a área sobre a curva ROC (AUC - *Area Under ROC Curve*) é empregada como uma medida de desempenho do teste como discriminador de doentes e saudáveis.

Dessa maneira, é possível dizer que qualquer que seja o ponto de corte, alguns indivíduos com a doença serão classificados corretamente (verdadeiros positivos - VP) e alguns serão classificados como negativos (falso negativos - FN). De forma análoga, indivíduos normais serão classificados corretamente como negativos (verdadeiros negativos - VN), enquanto alguns serão classificados como positivos (falso positivos - FP).

A matriz de confusão, ilustrada na Figura 23, ilustra as principais métricas de desempenho que podem ser utilizadas através da curva ROC, que são a acurácia, sensibilidade e especificidade. A classificação correta do modelo é indicada pela diagonal principal.

		Classe Verdadeira	
		P	N
Resultado da Classificação	P	Verdadeiro Positivo (VP)	Falso Positivo (FP)
	N	Falso Negativo (FN)	Verdadeiro Negativo (VN)

Figura 23 - Matriz de confusão ou tabela de contingência

A sensibilidade do modelo é definida pela a equação (39):

$$\text{Sensibilidade} = \frac{VP}{VP + FN} \quad (39)$$

Por outro lado, a especificidade do modelo é definida pela equação (40):

$$\text{Especificidade} = \frac{VN}{VN + FP} \quad (40)$$

Segundo (FAWCETT, 2006), para a construção da curva ROC devem ser usados os valores da sensibilidade no eixo y e 1 - especificidade no eixo x, representando a relação entre benefícios (verdadeiros positivos) e custos (falso positivos) de um modelo.

A Figura 24 ilustra três curvas hipotéticas representando a acurácia de diagnóstico. A curva A, com AUC = 1, representa o padrão ouro. Já a curva B, com AUC = 0,85, indica uma curva ROC típica e a curva C, com AUC = 0,5, resulta em uma chance aleatória. Quanto maior a AUC, ou seja, quanto mais próxima a 1, mais exato torna-se o diagnóstico (ZOU, O'MALLEY e MAURI, 2007).

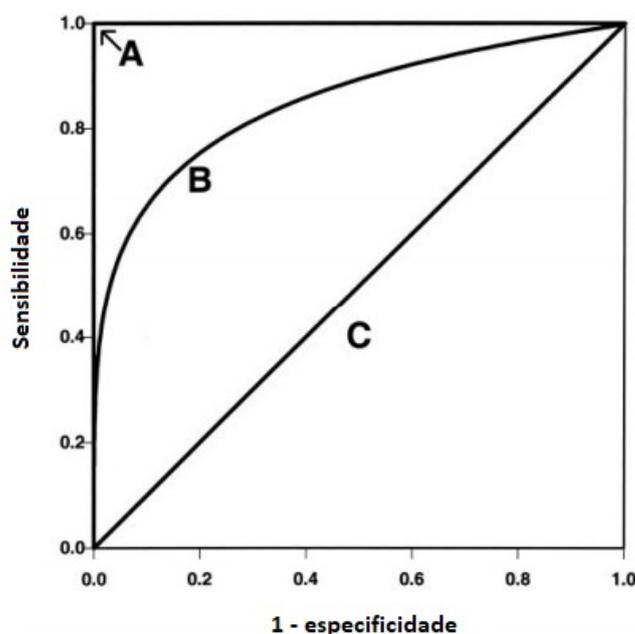


Figura 24 - Exemplos de curva ROC

(Adaptado de (ZOU, O'MALLEY e MAURI, 2007))

O ponto $(0,1)$, representado por A , representa classificações perfeitas, onde todos os exemplos positivos e negativos são classificados corretamente. sendo conhecido como *céu ROC*. Por outro lado, o ponto $(1,0)$ representa o ponto diametralmente oposto, chamado de *inferno ROC*. O ponto $(1,1)$ representa classificações sempre positivas e o ponto $(0,0)$ classificações sempre negativas. Desta maneira, classificadores que estiverem na região próxima ao ponto $(1,1)$ quase sempre rotulam os exemplos como positivos e classificadores na região $(0,0)$ rotulam a maioria dos exemplos como negativos.

Um classificador pode ser considerado melhor que o outro quando o seu ponto na curva ROC estiver acima e a esquerda do ponto correspondente do outro classificador (PRATI, 2006).

4.6. Classificadores

Os classificadores descritos no capítulo 3 foram implementados utilizando a *toolbox* PRTools. Os parâmetros foram definidos de acordo com seus desempenhos durante o treinamento, com base no erro encontrado durante este processo e utilizando a AUC como medida de desempenho, a saber: O algoritmo *Adaboost* foi implementado com 50, 100 e 150 árvores de decisão como classificadores simples, obtendo o melhor desempenho com 50 árvores; *Random Forest* foi construído com 10, 20 e 50 árvores geradas e tamanho do subconjunto de atributos igual a 1, 2 e 3, obtendo o melhor desempenho com 20 árvores e subconjunto de atributos igual a 2. Por sua vez, o algoritmo *K-Nearest Neighbor* foi implementado com $K=1, 3$ e 5 , obtendo o melhor desempenho com $K = 1$. No algoritmo *Radial Support Vector Machine* foram definidos dois parâmetros: o desvio padrão de base radial (r) e a constante de regularização (C). No AMI, apenas o número de atributos deveria ser definido, sendo usados 7, 5 e 3 respectivamente.

A determinação destes parâmetros foi realizada a partir de experimentos iniciais que se encontram na Tabela 6.

Tabela 6 - Experimentos para determinação dos parâmetros dos algoritmos

KNN			Adaboost			Random Forest											
Nº de vizinhos	C x NE	C x R	Nº de árvores	C x NE	C x R	Nº de árvores	C x NE	C x R	Nº de árvores	C x NE	C x R	Nº de árvores	C x NE	C x R			
	E_{AUC}	E_{AUC}		E_{AUC}	E_{AUC}		E_{AUC}	E_{AUC}		E_{AUC}	E_{AUC}		E_{AUC}	E_{AUC}			
K=1	0,1095	0,059	50	0,098	0,032	1*	10	0,156	0,053	2*	10	0,175	0,037	3*	10	0,122	0,046
K=2	0,1345	0,1	100	0,1	0,033		20	0,12	0,04		20	0,107	0,035		20	0,156	0,041
K=3	0,193	0,085	150	0,099	0,04		50	0,151	0,048		50	0,12	0,045		50	0,137	0,045

*Número de subconjuntos; E_{AUC} = Erro Padrão da AUC (1-AUC);

C x NE = Controle x Normal Exame; C x R = Controle x Restritivo

5. Estudo de Caso

Neste capítulo foram detalhados os experimentos implementados cujo objetivo era aplicar os parâmetros fornecidos pela TOF como entrada para os classificadores de aprendizado de máquinas. Foram propostas diferentes abordagens, visando à redução de atributos necessários para a classificação dos indivíduos e visando o aumento da acurácia de diagnóstico. Todas as análises foram realizadas em duas etapas: a primeira contempla o grupo controle e normal exame e a outra o grupo controle e restritivo.

5.1. Descrição do Conjunto de Dados

Foram selecionados 82 indivíduos, sendo 52 portadores de Esclerose Sistêmica e 30 sadios. Os portadores da doença foram divididos em dois grupos, sendo um denominado de normal ao exame, composto por 22 pacientes com exame espirométrico normal e outro grupo, denominado restritivo, formado por 30 pacientes diagnosticados com distúrbio respiratório restritivo pela espirometria. Para cada indivíduo, foram realizados 3 ensaios, totalizando 246 instâncias disponíveis para avaliação.

As características dos exames de TOF dos indivíduos pertencentes ao grupo controle, normal ao exame espirométrico e restritivo foram comparadas na Figura 25. Os gráficos de barras indicam os valores médios de cada uma das variáveis, com um intervalo de confiança de 95%, que também estão indicados nos gráficos.

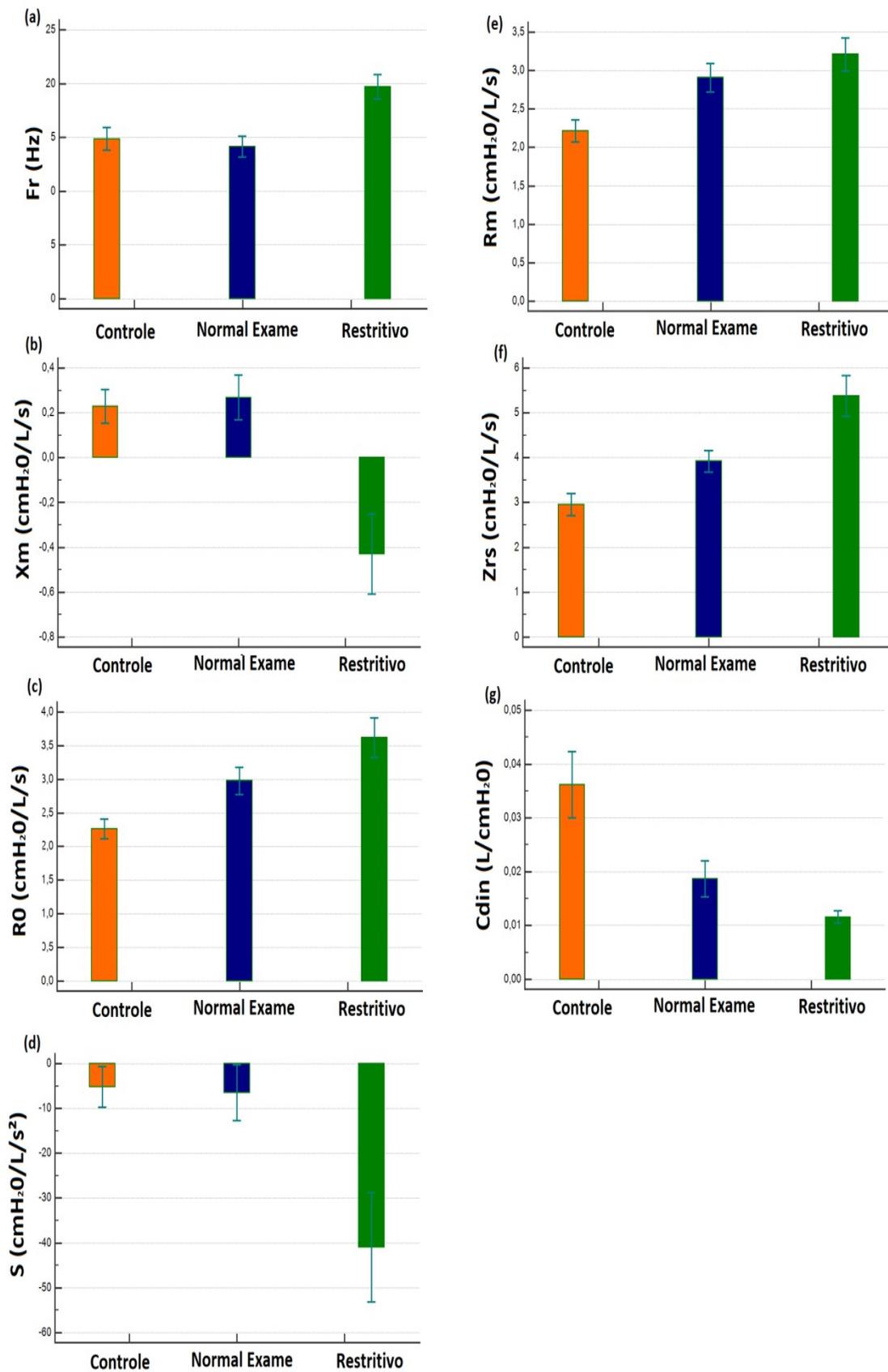


Figura 25 - Comparação dos valores da TOF para os grupos controle, esclerodermia (normal exame) e esclerodermia (restritivo)

Na Figura 25, é possível perceber visualmente que houve um aumento na média das variáveis R_0 , R_m , Z_{rs} , na comparação dos três grupos. Assim, é possível notar que os portadores da Esclerodermia possuem valores mais altos de resistências (R_0 e R_m) e de impedância (Z_{rs}). Por outro lado, a frequência de ressonância (F_r) e a inclinação da curva de resistência (S) possuem valores próximos para os grupos controle e normal exame, porém o primeiro possui valores mais elevados para o grupo restritivo, enquanto o segundo possui valores mais negativos. A complacência dinâmica (C_{din}) possui valores mais elevados para o grupo controle e valores semelhantes nos grupos normal exame e restritivo. Os valores de cada atributo estão detalhados no apêndice 1.

Todas as análises estatísticas foram realizadas por meio do MedCalc³, com a metodologia descrita em (DELONG, DELONG e CLARKE-PEARSON, 1988).

5.2. Experimento com apenas um atributo

A primeira análise realizada foi submeter cada atributo da TOF de forma a mensurar seu desempenho em classificar cada paciente. Todos os atributos tiveram as medidas de AUC, bem como seus intervalos de confiança de 95% (IC), sensibilidade e especificidade calculados (DELONG, DELONG e CLARKE-PEARSON, 1988).

5.2.1. Análise grupo controle versus grupo normal exame

A complacência dinâmica (C_{din}) foi o parâmetro que alcançou o melhor desempenho individual, com valor de $AUC = 0,77$, passando a ser denominado *BFP* (*Best FOT Parameter*). Os demais parâmetros apresentaram AUC variando entre 0,51 e 0,75. A Tabela 7 mostra os valores da AUC, IC, sensibilidade e especificidade dos atributos da TOF para o grupo em estudo.

³ MedCalc *Statistical Software version 14.8.1* (MedCalc Software bvba, Ostend, Belgium; <http://www.medcalc.org>; 2014)

Tabela 7 - Desempenho individual dos atributos da TOF na classificação de pacientes entre os grupos controle e normal exame

Característica	AUC	95% IC	Sensibilidade	Especificidade
F_r	0,51	0,43 - 0,59	3,0	83,3
X_m	0,55	0,47 - 0,633	63,6	62,2
R_0	0,74	0,66 - 0,81	80,3	61,1
S	0,51	0,43 - 0,59	36,4	82,1
R_m	0,75	0,68 - 0,82	59,1	80,0
Z_{rs}	0,75	0,67 - 0,81	84,4	56,7
C_{din}	0,77	0,72 - 0,85	60,6	85,6

Na Figura 26 é possível notar que os parâmetros inclinação da curva de resistência (S), frequência de ressonância (F_r) e reatância média (X_m), sozinhas, não são capazes de descrever a doença.

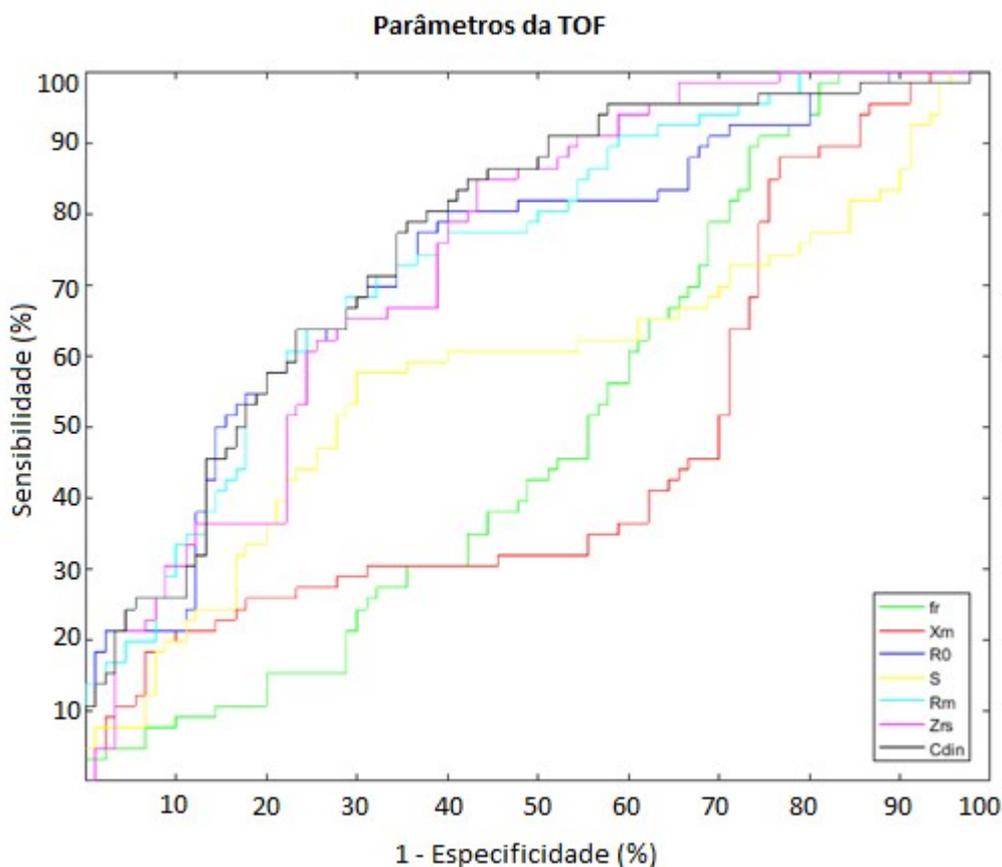


Figura 26 - Curva ROC com os parâmetros da TOF individualmente para os grupos controle e normal exame

5.2.2. Análise grupo controle *versus* grupo restritivo

Da mesma forma como ocorreu na análise anterior, novamente a complacência dinâmica (C_{din}) foi o parâmetro que alcançou o melhor desempenho individual, com valor de $AUC = 0,94$. Os demais parâmetros apresentaram AUC variando entre 0,70 e 0,87. A Tabela 8 mostra os valores da AUC , IC, sensibilidade e especificidade dos atributos da TOF para o grupo em estudo.

Tabela 8 - Desempenho individual dos atributos da TOF na classificação de pacientes entre os grupos controle e Restritivo

Característica	AUC	95% IC	Sensibilidade	Especificidade
F_r	0,75	0,679 - 0,811	94,4	44,4
X_m	0,77	0,696 - 0,825	60,0	92,2
R_0	0,82	0,752 - 0,870	66,7	85,6
S	0,70	0,616 - 0,756	48,9	91,1
R_m	0,79	0,726 - 0,849	82,2	62,2
Z_{rs}	0,87	0,814 - 0,917	87,8	70,0
C_{din}	0,94	0,893 - 0,969	88,9	85,6

Na Figura 27 é intuitivo notar o destaque do atributo C_{din} quando comparado com os demais atributos.

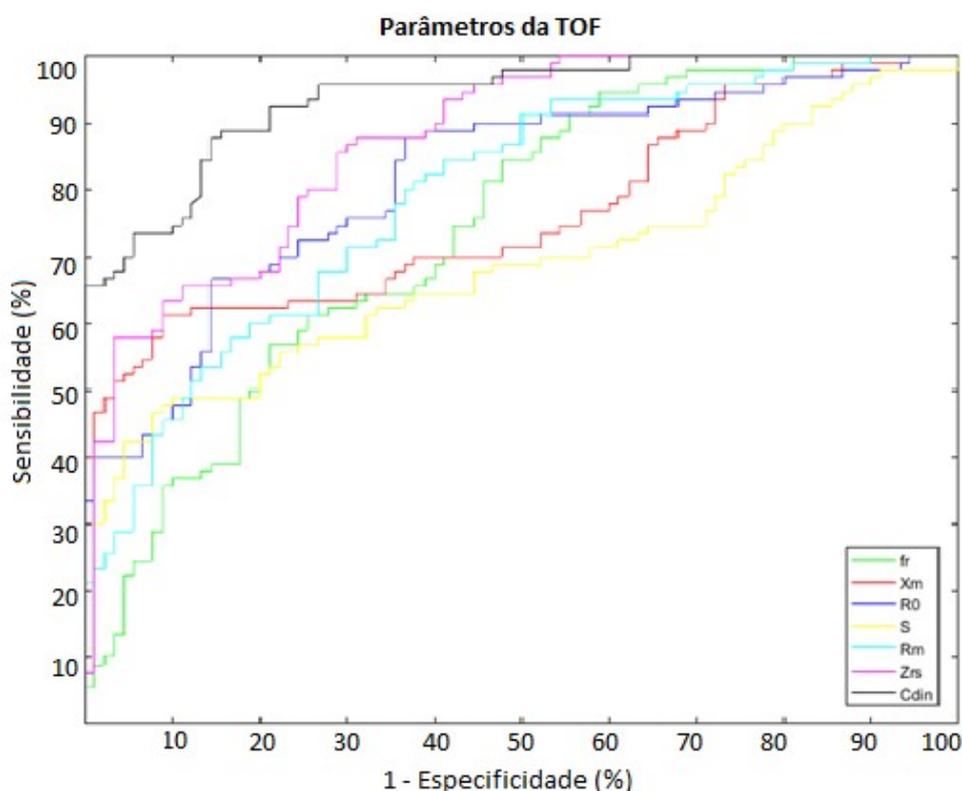


Figura 27 - Curva ROC com os parâmetros da TOF individualmente para os grupos controle e restritivo

5.3. Experimento com todos os atributos

O segundo experimento realizado utilizou todos os atributos da TOF nos seguintes algoritmos de aprendizado de máquinas: *K-Nearest Neighbor* (KNN), *Adaboost* (ADAB), *Random Forest* (RF), *Radial Support Vector Machine* (RSVM) e AMI (Aprendizado em Múltiplas Instâncias).

5.3.1. Análise grupo controle *versus* grupo normal exame

Os resultados dos classificadores obtidos foram comparados com o melhor parâmetro da TOF, descrito na seção 5.2. A Tabela 9 descreve os resultados para AUC, com intervalo de confiança de 95%, sensibilidade e especificidade. Cabe ressaltar que em todos os experimentos realizados utilizando os algoritmos de aprendizado de

máquinas foram utilizadas as mesmas partições para permitir que os resultados fossem confrontados.

Tabela 9 - Resultados dos classificadores com todos os atributos da TOF

	Sensibilidade(%)	Especificidade(%)	AUC
BFP	89,4 (79,4 - 95,6)	52,2 (41,4 - 62,9)	0,77 (0,69 - 0,83)
KNN	84,8 (73,9 - 92,5)	80,0 (70,2 - 87,7)	0,90 (0,84 - 0,94)
ADAB	89,4 (79,4 - 95,6)	76,6 (65,4 - 84,0)	0,87 (0,81 - 0,92)
RF	77,3 (65,3 - 86,7)	74,4 (64,2 - 83,1)	0,83 (0,76 - 0,89)
RSVM	89,4 (79,4 - 95,6)	76,7 (66,6 - 84,9)	0,87 (0,80 - 0,92)
AMI (7)	83,3 (72,1 - 91,4)	58,9 (48,0 - 69,2)	0,72 (0,65 - 0,80)

Através dos resultados obtidos, pode-se dizer que o melhor desempenho foi alcançado pelo algoritmo *K-Nearest Neighbor* (KNN), com uma $AUC = 0,90$. Esse resultado indica que o algoritmo alcançou alta acurácia diagnóstica ($0,9 \leq AUC \leq 1$). O segundo melhor desempenho, já considerado com acurácia moderada ($0,7 \leq AUC < 0,9$) foi alcançado pelos algoritmos Adaboost (ADAB) e *Radial Support Vector Machine* (RSVM), com uma $AUC = 0,87$.

Além da AUC, mostrada na Figura 28, foram calculados os valores de sensibilidade e especificidade, o que significa que foram indicadas as probabilidades de classificar um indivíduo como positivo quando ele for portador da doença e um resultado negativo quando ele for normal, respectivamente.

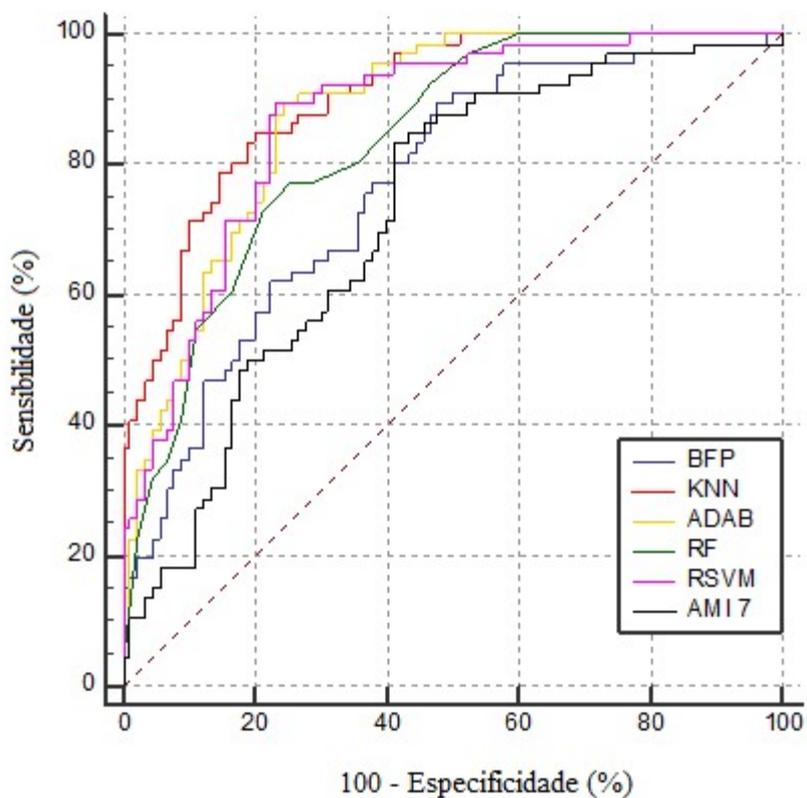


Figura 28 - Curvas ROC obtidas no experimento com todos os atributos da TOF

A Tabela 10 mostra a comparação das AUCs obtidas com o melhor parâmetro da TOF contra os demais classificadores e o erro padrão, em um intervalo de confiança de 95%. A interseção entre linha e coluna indica a diferença da AUC entre dois classificadores. Por exemplo, na interseção da primeira linha com a segunda coluna, há diferença significativa ($p < 0,005$) entre a AUC obtida pelo BFP e pelo ADAB, enquanto a interseção da segunda linha com a segunda coluna indica que não houve diferença significativa no par analisado ($p > 0,05$).

Tabela 10 - Comparação das AUCs entre os modelos com todos os atributos da TOF – Análise Controle *versus* Normal Exame

	KNN	ADAB	RF	RSVM	AMI(7)
BFP	0,135±0,0359 ⁺	0,108±0,0342 ⁺	0,0666±0,0280*	0,102±0,0369*	0,0436±0,0210*
KNN	-	0,0263±0,0236	0,0681±0,0262**	0,0328±0,0274	0,178±0,0396 ⁺⁺
ADAB	-	-	0,0418±0,0170	0,00657±0,0260	0,152±0,0388 ⁺⁺
RF	-	-	-	0,0353±0,0285	0,110±0,0347 ⁺
RSVM	-	-	-	-	0,145±0,0398 ⁺

BFP: *Best FOT Parameter* (Obtido sem o uso de classificadores; C_{din});

ADAB: *Adaboost*; 1-NN: *K-Nearest Neighbor* ($K = 1$); RF: *Random Forests*;

RSVM: *Radial Support Vector Machine*; AMI: Aprendizado em Múltiplas Instâncias

* $p < 0,05$

** $p < 0,01$

⁺ $p < 0,005$

⁺⁺ $p < 0,0001$

Além das análises anteriores, é possível comparar a sensibilidade observando a especificidade com os valores fixados em 75% e 90%. O primeiro valor representa uma especificidade moderada, enquanto o segundo indica uma alta especificidade. Essa análise limita os falsos positivos, permitindo uma análise onde o algoritmo dificilmente erra a classificação dos portadores da doença. Tal escolha limita a taxa de falsos positivos em 25% e 10% respectivamente.

A Figura 29 mostra que com o valor da especificidade fixado em 75%, houve um aumento no valor da sensibilidade de todos os classificadores, com exceção do AMI, quando comparados com o *BFP*. O mesmo ocorreu quando a especificidade foi fixada em 90%, havendo um aumento na sensibilidade de todos os classificadores, exceto no AMI. Esse apresentou um decréscimo substancial na sensibilidade.

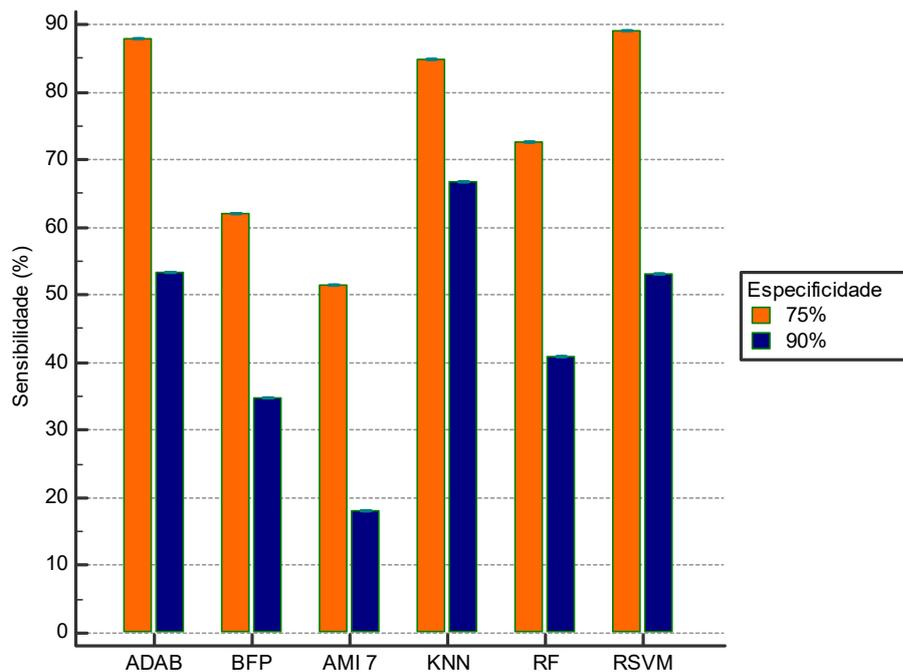


Figura 29 - Análise comparativa da sensibilidade obtida nos cinco classificadores com a especificidade de 75% e 90% para os grupos controle e normal exame

5.3.2. Análise grupo controle *versus* grupo restritivo

Seguindo o mesmo padrão realizado anteriormente, os resultados dos classificadores obtidos foram comparados com o melhor parâmetro da TOF para os grupos controle e restritivo. A Tabela 11 descreve os resultados para AUC, com intervalo de confiança de 95%, sensibilidade e especificidade.

Tabela 11 - Resultados dos classificadores com todos os atributos da TOF

	Sensibilidade(%)	Especificidade(%)	AUC
BFP	88,9 (80,5 - 94,5)	84,4 (75,3 - 91,2)	0,94 (0,89 - 0,97)
KNN	95,6 (89,0 - 98,8)	82,2 (72,7 - 89,0)	0,94 (0,90 - 0,97)
ADAB	95,6 (89,0 - 98,8)	85,6 (76,6 - 92,1)	0,96 (0,92 - 0,98)
RF	95,6 (89,0 - 98,8)	85,6 (76,6 - 92,1)	0,95 (0,92 - 0,98)
RSVM	90,0 (81,9 - 95,3)	86,7 (77,9 - 92,9)	0,91 (0,86 - 0,95)
AMI (7)	90,0 (81,9 - 95,3)	85,6 (76,6 - 92,1)	0,94 (0,89 - 0,97)

Através dos resultados obtidos, pode-se afirmar que o melhor desempenho foi alcançado pelo algoritmo *Adaboost* (ADAB), com uma AUC = 0,96. Esse resultado indica que o algoritmo alcançou alta acurácia diagnóstica. O segundo melhor desempenho foi alcançado pelo algoritmo *Random Forest* (RF), com AUC = 0,95.

Além da AUC, mostrada na Figura 30, foram calculados os valores de sensibilidade e especificidade, o que significa que foram indicadas as probabilidades de classificar um indivíduo como positivo quando ele for portador da doença e um resultado negativo quando ele for normal, respectivamente.

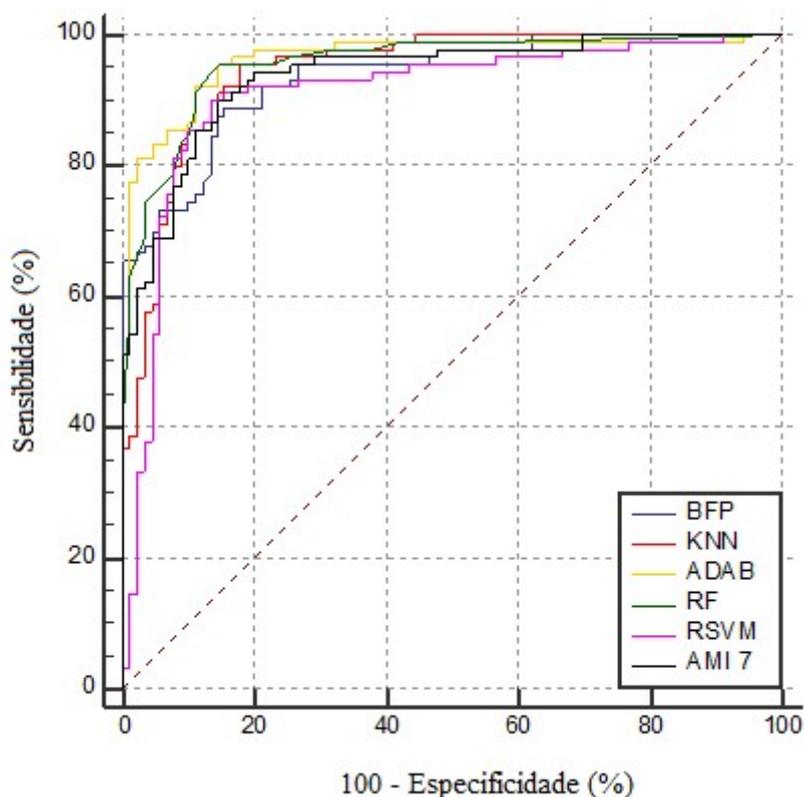


Figura 30 - Curvas ROC obtidas no experimento com todos os atributos da TOF

A Tabela 12 mostra a comparação das AUCs obtidas com o melhor parâmetro da TOF contra os demais classificadores e o erro padrão, em um intervalo de confiança de 95%. A interseção entre linha e coluna indica a diferença da AUC entre dois classificadores.

Tabela 12 - Comparação das AUCs entre os modelos com todos os atributos da TOF – Análise Controle *versus* Restritivo

	KNN	ADAB	RF	RSVM	AMI(Carac5)
BFP	0,00833±0,0191	0,0276±0,0149	0,0205±0,0141	0,0250±0,0190	0,00228±0,00821
KNN	-	0,0193±0,0162	0,0122±0,0147	0,0333±0,0228	0,00605±0,0173
ADAB	-	-	0,00710±0,00750	0,0526±0,0191 ⁺	0,0253±0,0150
RF	-	-	-	0,0455±0,0191	0,0182±0,0140
RSVM	-	-	-	-	0,0273±0,0155

BFP: *Best FOT Parameter* (Obtido sem o uso de classificadores; C_{din});

ADAB: *Adaboost*; 1-NN: *K-Nearest Neighbor* ($K = 1$); RF: *Random Forests*;

RSVM: *Radial Support Vector Machine*; AMI: *Aprendizado em Múltiplas Instâncias*

* $p < 0,05$

** $p < 0,01$

⁺ $p < 0,005$

⁺⁺ $p < 0,0001$

Além das análises anteriores, é possível comparar a sensibilidade observando a especificidade com os valores fixados em 75% e 90%. O primeiro valor representa uma especificidade moderada, enquanto o segundo indica uma alta especificidade. Essa análise limita os falsos positivos, permitindo uma análise onde o algoritmo dificilmente erra a classificação dos portadores da doença. Tal escolha limita a taxa de falsos positivos em 25% e 10% respectivamente.

A Figura 31 mostra que com o valor da especificidade fixado em 75%, houve um aumento no valor da sensibilidade de todos os classificadores quando comparados com o *BFP*. O mesmo ocorreu quando a especificidade foi fixada em 90%, havendo um aumento na sensibilidade de todos os classificadores

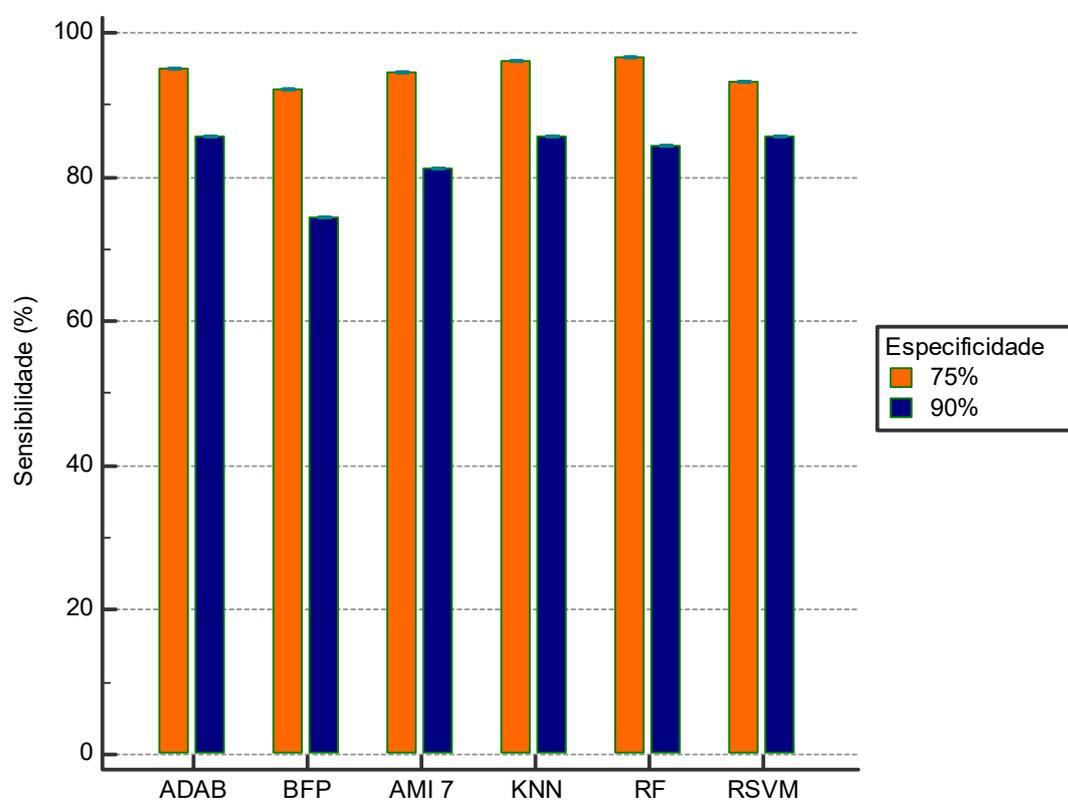


Figura 31 - Análise comparativa da sensibilidade obtida nos cinco classificadores com a especificidade de 75% e 90% para os grupos controle e restritivo

5.4. Experimento com cinco atributos

A exemplo do experimento anterior, os atributos da TOF foram utilizados nos seguintes algoritmos de aprendizado de máquinas: *K-Nearest Neighbor* (KNN), *Adaboost* (ADAB), *Random Forest* (RF), *Radial Support Vector Machine* (RSVM) e AMI (Aprendizado em Múltiplas Instâncias).

Diferentemente do modelo anterior, foram selecionados apenas cinco atributos. A seleção lançou mão da escolha pelo especialista, além de analisar todas as possíveis combinações de características classificadas pelo algoritmo de aprendizado em múltiplas instâncias (Apêndice 2), sendo a melhor delas alvo de análise posterior (seção 5.4.2). Também foi utilizada a seleção direta de atributos, através da função *featsel*, disponível no PRTools, cuja seleção de atributos é feita pelo 1-NN.

5.4.1. Experimento com cinco atributos selecionados pelo especialista

Os atributos selecionados pelo especialista foram R_0 , R_m , Z_{rs} e C_{din} , para ambas as análises. Os resultados estão detalhados na Tabela 13.

Tabela 13 – Resultados obtidos na seleção de atributos pelo Especialista

	Controle x Normal Exame			Controle x Restritivo			
	Sensibilidade (%)	Especificidade (%)	AUC	Sensibilidade (%)	Especificidade (%)	AUC	
Especialista	BFP	89,40	52,20	0,77	88,90	84,40	0,94
	KNN	86,40	68,90	0,85	92,20	86,70	0,93
	ADAB	77,30	76,70	0,83	91,10	93,30	0,98
	RF	74,20	78,90	0,82	93,30	91,10	0,97
	RSVM	87,90	64,60	0,82	94,40	81,10	0,93
	AMI	87,90	48,90	0,72	91,10	82,20	0,93

Através dos resultados obtidos, pode-se dizer que o melhor desempenho na análise controle *versus* normal exame foi alcançado pelo algoritmo *K-Nearest Neighbor* (KNN), com uma AUC = 0,85. Esse resultado indica que o algoritmo alcançou acurácia diagnóstica moderada. A curva ROC de cada classificador foi traçada na

Figura 32, ilustrando os resultados mostrados na Tabela 13.

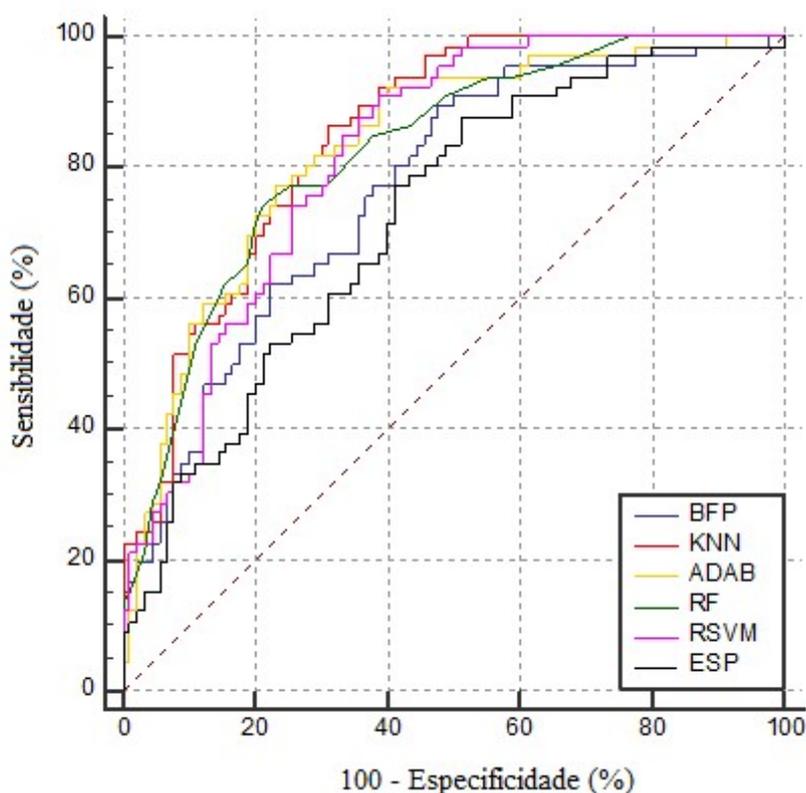


Figura 32 - Curvas ROC obtidas no experimento com os parâmetros da TOF selecionados pelo especialista para os grupos controle e normal exame

A Tabela 14 mostra a comparação das AUCs obtidas com o melhor parâmetro da TOF contra os demais classificadores e o erro padrão, desenvolvidos com base na escolha dos atributos pelo especialista. A interseção entre linha e coluna indica a diferença da AUC entre dois classificadores.

Tabela 14 - Comparação das AUCs entre os modelos com a seleção de atributos pelo especialista – Análise Controle *versus* Normal Exame

	KNN	ADAB	RF	RSVM	AMI(ESP)
BFP	0,0822±0,0358	0,0646±0,0362	0,0562±0,0315	0,0535±0,0365	0,0436±0,0206*
KNN	-	0,0175±0,0258	0,0259±0,0266	0,0286±0,0295	0,126±0,0410 ⁺
ADAB	-	-	0,00842±0,0173	0,0111±0,0310	0,108±0,0410**
RF	-	-	-	0,00269±0,0308	0,0998±0,0370**
RSVM	-	-	-	-	0,0971±0,0401*

BFP: *Best FOT Parameter* (Obtido sem o uso de classificadores; C_{din});

ADAB: *Adaboost*; 1-NN: *K-Nearest Neighbor* ($K = 1$); RF: *Random Forests*;

RSVM: *Radial Support Vector Machine*; AMI: Aprendizado em Múltiplas Instâncias

* $p < 0,05$

** $p < 0,01$

⁺ $p < 0,005$

⁺⁺ $p < 0,0001$

Por outro lado, analisando os resultados obtidos através da análise Controle *versus* Restritivo, pode-se dizer que o melhor desempenho foi alcançado pelo algoritmo *Adaboost* (ADAB), com uma AUC = 0,98. Esse resultado indica que o algoritmo alcançou alta acurácia diagnóstica. A Figura 33 mostra as Curvas ROC de cada atributo.

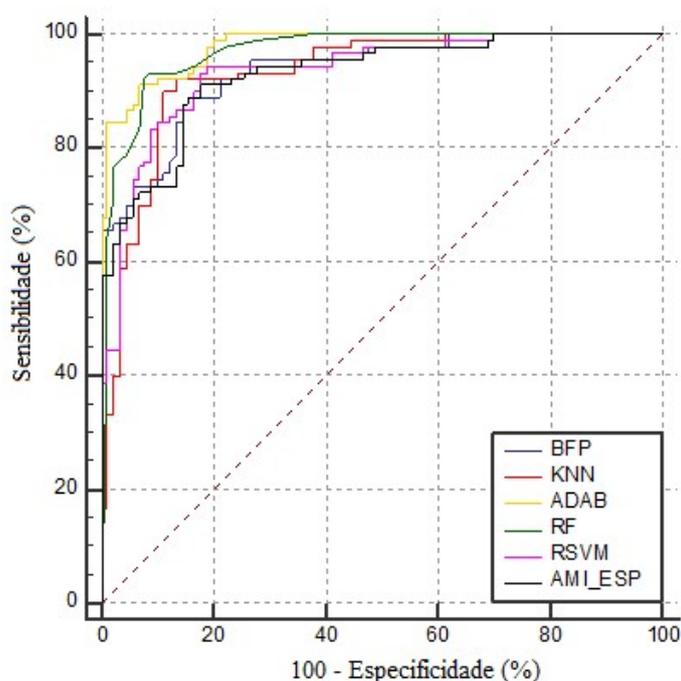


Figura 33 - Curvas ROC obtidas no experimento com os parâmetros da TOF selecionados pelo especialista para os grupos controle e restritivo

A Tabela 15 mostra a comparação das AUCs obtidas com o melhor parâmetro da TOF contra os demais classificadores e o erro padrão, desenvolvidos com base na escolha dos atributos pelo especialista.

Tabela 15 - Comparação das AUCs entre os modelos com a seleção pelo especialista – Análise Controle *versus* Restritivo

	KNN	ADAB	RF	RSVM	AMI(ESP)
BFP	0,00562±0,0204	0,0427±0,0145 ⁺	0,0322±0,0143*	0,00204±0,0139	0,0071±0,00521
KNN	-	0,0483±0,0146 ⁺	0,0378±0,0143**	0,00358±0,0185	0,00148±0,0212
ADAB	-	-	0,0105±0,00671	0,0447±0,0150**	0,0498±0,0159 ⁺
RF	-	-	-	0,0342±0,0156*	0,0393±0,0154*
RSVM	-	-	-	-	0,00506±0,0143

BFP: *Best FOT Parameter* (Obtido sem o uso de classificadores; C_{din});

ADAB: *Adaboost*; 1-NN: *K-Nearest Neighbor* ($K = 1$); RF: *Random Forests*;

RSVM: *Radial Support Vector Machine*; AMI: Aprendizado em Múltiplas Instâncias

* $p < 0,05$

** $p < 0,01$

⁺ $p < 0,005$

⁺⁺ $p < 0,0001$

No apêndice 3 estão disponíveis os gráficos com as análises com a sensibilidade fixada em 75% e 90%, para ambos os casos analisados acima.

5.4.2. Experimento com cinco atributos selecionados pelo melhor classificador de Aprendizado em Múltiplas Instâncias

A Tabela 16 mostra os resultados obtidos através da seleção pelo algoritmo de Múltiplas Instâncias, no qual todas as combinações com 5 atributos foram utilizadas no algoritmo AMI, sendo selecionadas as que obtiveram maior valor da AUC. Os atributos selecionados por essa técnica para a análise Controle *versus* Normal Exame foram: X_m , R_0 , S , R_m e C_{din} , de acordo com os resultados mostrados no apêndice 2.

Tabela 16 - Resultados obtidos na seleção de atributos pelo AMI

		Controle x Normal Exame			Controle x Restritivo		
		Sensibilidade (%)	Especificidade (%)	AUC	Sensibilidade (%)	Especificidade (%)	AUC
AMI	BFP	89,40	52,20	0,77	88,90	84,40	0,94
	KNN	80,30	81,10	0,88	90,00	91,10	0,95
	ADAB	86,40	70,00	0,84	91,10	95,60	0,98
	RF	80,30	81,10	0,85	91,10	91,10	0,97
	RSVM	87,90	74,40	0,83	88,90	87,90	0,92
	AMI	71,20	65,60	0,74	90,00	85,60	0,95

Através dos resultados obtidos, pode-se dizer que o melhor desempenho foi alcançado pelo algoritmo KNN, com $AUC = 0,88$, na análise Controle *versus* Normal Exame. A Figura 34 mostra as curvas ROC obtidas para todos os algoritmos analisados.

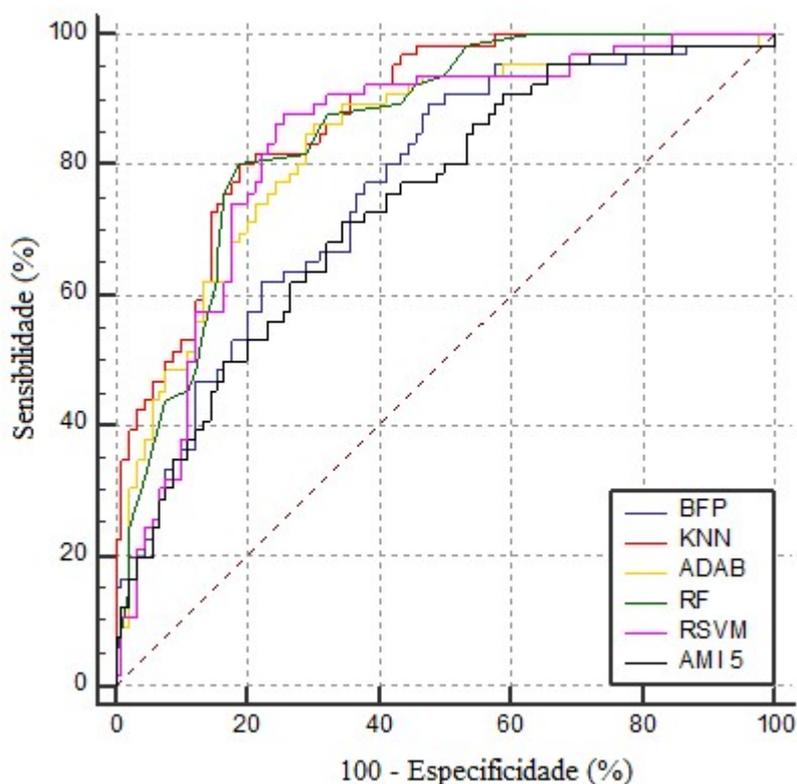


Figura 34 - Curvas ROC obtidas no experimento com os parâmetros da TOF selecionados pelo melhor desempenho do Classificador AMI para os grupos Controle e Normal Exame

A Tabela 17 mostra a comparação das AUCs obtidas com o melhor parâmetro da TOF e com os demais classificadores desenvolvidos com base na escolha dos atributos pelo melhor resultado obtido através do Aprendizado em Múltiplas Instâncias, com 5 atributos. A interseção entre linha e coluna indica a diferença da AUC entre dois classificadores.

Tabela 17 - Comparação das AUCs entre os modelos com cinco atributos selecionados pelo melhor desempenho AMI – Análise Controle *versus* Normal

Exame					
	KNN	ADAB	RF	RSVM	AMI(Carac5)
BFP	0,108±0,0371 ⁺	0,0689±0,0363	0,0856±0,0327**	0,0660±0,0364	0,0268±0,0214
KNN	-	0,0394±0,0285	0,0226±0,0267	0,0423±0,0307	0,135±0,0382 ⁺
ADAB	-	-	0,0168±0,0197	0,00286±0,0336	0,0956±0,0386*
RF	-	-	-	0,0196±0,0309	0,112±0,0356 ⁺⁺
RSVM	-	-	-	-	0,0928±0,0368*

BFP: *Best FOT Parameter* (Obtido sem o uso de classificadores; C_{din});

ADAB: *Adaboost*; 1-NN: *K-Nearest Neighbor* ($K = 1$); RF: *Random Forests*;

RSVM: *Radial Support Vector Machine*; AMI: Aprendizado em Múltiplas Instâncias

* $p < 0,05$

** $p < 0,01$

⁺ $p < 0,005$

⁺⁺ $p < 0,0001$

Por outro lado, quando a análise foi feita para os grupos Controle *versus* Restritivo, pode-se dizer que o melhor desempenho foi alcançado pelo algoritmo ADAB, com $AUC = 0,98$. Para tanto, foram selecionados para a análise os atributos F_R , R_0 , R_m , Z_{rs} e C_{din} . A Figura 35 ilustra as Curvas ROC obtidas:

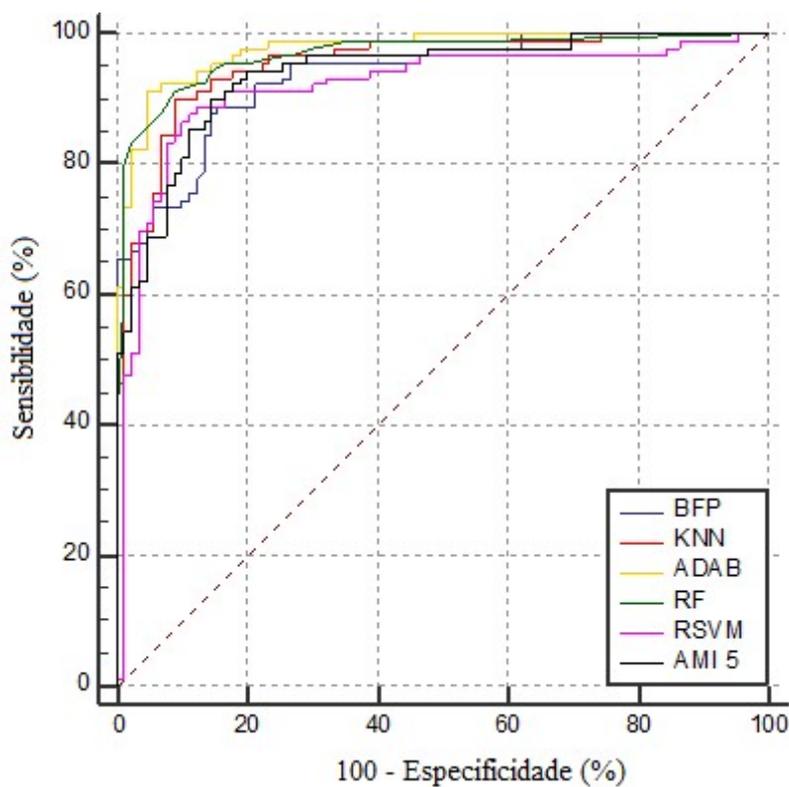


Figura 35 - Curvas ROC obtidas no experimento com os parâmetros da TOF selecionados pelo melhor desempenho do Classificador AMI para os grupos controle e restritivo

A Tabela 18 mostra a comparação das AUCs obtidas com o melhor parâmetro da TOF e com os demais classificadores desenvolvidos com base na escolha dos atributos pelo melhor resultado obtido através do Aprendizado em Múltiplas Instâncias com 5 atributos para os grupos controle e restritivo.

Tabela 18 - Comparação das AUCs entre os modelos com cinco atributos selecionados pelo melhor desempenho do algoritmo de AMI – Análise Controle versus Restritivo

	KNN	ADAB	RF	RSVM	AMI(5)
BFP	0,0182±0,0183	0,0399±0,0143**	0,0317±0,0145*	0,0178±0,0168	0,00228±0,00821
KNN	-	0,0217±0,0127	0,0135±0,0142	0,0360±0,0217	0,0159±0,0168
ADAB	-	-	0,00821±0,00611	0,0578±0,0186 ⁺	0,0377±0,0142**
RF	-	-	-	0,0496±0,0196*	0,0294±0,0146*
RSVM	-	-	-	-	0,0201±0,0135

BFP: *Best FOT Parameter* (Obtido sem o uso de classificadores; C_{din});

ADAB: *Adaboost*; 1-NN: *K-Nearest Neighbor* ($K = 1$); RF: *Random Forests*;

RSVM: *Radial Support Vector Machine*; AMI: Aprendizado em Múltiplas Instâncias

* $p < 0,05$

** $p < 0,01$

⁺ $p < 0,005$

⁺⁺ $p < 0,0001$

No apêndice 4 estão disponíveis os gráficos com as análises com a sensibilidade fixada em 75% e 90% para ambos os casos analisados acima e suas respectivas análises.

5.4.3. Experimento com cinco atributos escolhidos através da seleção direta

Nessa etapa, utilizou-se a *toolbox featsel*, disponível no PRTools, que realiza a seleção direta de atributos. Nela, o critério escolhido foi a taxa de acerto no algoritmo K-NN, com K igual a 1 (1-NN). O treinamento do classificador 1-NN foi feito através da validação cruzada *leave-one-out*, onde uma amostra n é testada com base nas $n-1$ amostras restantes. Dessa maneira, os atributos escolhidos pela *toolbox* foram: F_R, R_0, R_m, Z_{rs} e C_{din} para a análise Controle versus Normal Exame e X_m, R_0, S, R_m e Z_{rs} para a análise Controle versus Restritivo

A Tabela 19 mostra os resultados obtidos através da Seleção Direta para a análise Controle versus Normal Exame e Controle versus Restritivo.

Tabela 19 – Resultados obtidos através da Seleção Direta de 5 Atributos

	Controle x Normal Exame			Controle x Restritivo			
	Sensibilidade (%)	Especificidade (%)	AUC	Sensibilidade (%)	Especificidade (%)	AUC	
Seleção Direta	BFP	89,40	52,20	0,77	88,90	84,40	0,94
	KNN	87,90	78,90	0,89	90,00	91,10	0,95
	ADAB	89,40	70,00	0,84	91,10	93,30	0,97
	RF	83,30	58,90	0,81	93,30	90,00	0,96
	RSVM	81,80	75,60	0,83	88,90	87,80	0,92
	AMI	86,40	54,40	0,71	71,10	86,70	0,87

Através dos resultados obtidos, pode-se dizer que o melhor desempenho foi alcançado pelo algoritmo KNN, com AUC = 0,89, seguido pelo algoritmo ADAB, com AUC = 0,84. Esse resultado indica que o algoritmo alcançou acurácia diagnóstica moderada. O resultado obtido por essa técnica foi superior ao encontrado pela seleção através do AMI e do especialista.

A Curva ROC de cada classificador é mostrada na Figura 36, onde é possível notar que o melhor desempenho é obtido pelo algoritmo KNN.

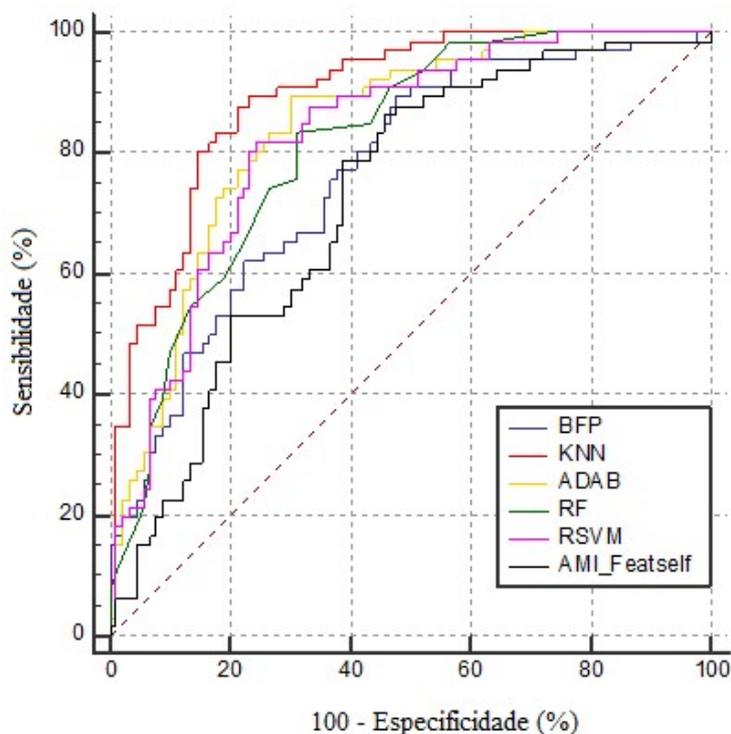


Figura 36 - Curvas ROC obtidas no experimento com os parâmetros da TOF escolhidos através da seleção direta para os grupos controle e normal exame

A Tabela 20 mostra a comparação das AUCs obtidas com o melhor parâmetro da TOF e com os demais classificadores desenvolvidos. A interseção entre linha e coluna indica a diferença da AUC entre dois classificadores.

Tabela 20 - Comparação das AUCs entre os modelos com cinco atributos escolhidos através da seleção direta – Análise Controle *versus* Normal Exame

	KNN	ADAB	RF	RSVM	AMI(ESP)
BFP	0,127±0,0395 ⁺	0,0742±0,0399	0,04830,0356	0,0594±0,0412	0,0461±0,0201*
KNN	-	0,0527±0,0272	0,0786±0,0296**	0,0675±0,0299*	0,173±0,0427 ⁺⁺
ADAB	-	-	0,0259±0,0168	0,0148±0,0323	0,120±0,0434**
RF	-	-	-	0,0111±0,0332	0,0944±0,0391*
RSVM	-	-	-	-	0,106±0,0437*

BFP: *Best FOT Parameter* (Obtido sem o uso de classificadores; C_{din});

ADAB: *Adaboost*; 1-NN: *K-Nearest Neighbor* ($K = 1$); RF: *Random Forests*;

RSVM: *Radial Support Vector Machine*; AMI: Aprendizado em Múltiplas Instâncias

* $p < 0,05$

** $p < 0,01$

⁺ $p < 0,005$

⁺⁺ $p < 0,0001$

Por outro lado, através dos resultados obtidos para a análise Controle *versus* Restritivo, pode-se dizer que o melhor desempenho foi alcançado pelo algoritmo ADAB, com AUC = 0,97, seguido pelo algoritmo RF, com AUC = 0,96. Esse resultado indica que o algoritmo alcançou alta acurácia diagnóstica. Cabe ressaltar que o desempenho dos algoritmos com os atributos selecionados por essa técnica foi inferior ao encontrado com os parâmetros escolhidos pelo especialista e pelo AMI. A Figura 37 mostra as Curvas ROC de cada classificador

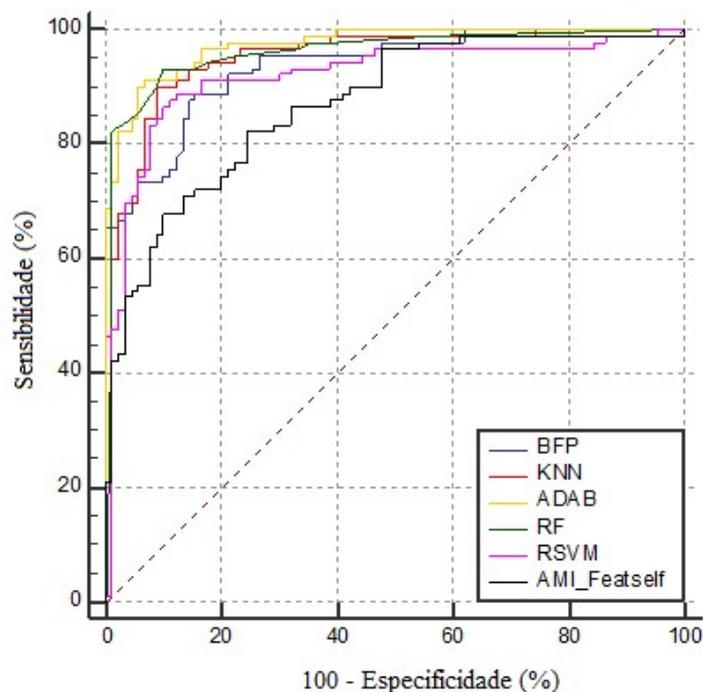


Figura 37 - Curvas ROC obtidas no experimento com os parâmetros da TOF escolhidos através da seleção direta de atributos para os grupos controle e restritivo

A Tabela 21 mostra a comparação das AUCs obtidas com o melhor parâmetro da TOF e com os demais classificadores desenvolvidos. A interseção entre linha e coluna indica a diferença da AUC entre dois classificadores

Tabela 21 - Comparação das AUCs entre os modelos com cinco atributos escolhidos através da seleção direta de atributos – Análise Controle *versus* Restritivo

	KNN	ADAB	RF	RSVM	AMI(ESP)
BFP	0,0182±0,0183	0,0380±0,0145**	0,0249±0,0157	0,0178±0,0168	0,0659±0,0196**
KNN	-	0,0198±0,0128	0,00673±0,0162	0,0360±0,0217	0,0841±0,0257 ⁺
ADAB	-	-	0,0130±0,00871	0,0558±0,0193 ⁺	0,104±0,0243 ⁺⁺
RF	-	-	-	0,0428±0,0204*	0,0908±0,0246 ⁺
RSVM	-	-	-	-	0,0480±0,0256

BFP: *Best FOT Parameter* (Obtido sem o uso de classificadores; C_{din});

ADAB: *Adaboost*; 1-NN: *K-Nearest Neighbor* ($K = 1$); RF: *Random Forests*;

RSVM: *Radial Support Vector Machine*; AMI: *Aprendizado em Múltiplas Instâncias*

* $p < 0,05$ ** $p < 0,01$ ⁺ $p < 0,005$ ⁺⁺ $p < 0,0001$

No apêndice 5 estão disponíveis os gráficos com as análises com a sensibilidade fixada em 75% e 90%, para ambos os casos analisados acima.

Na comparação com os métodos de seleção de atributos anteriores, o desempenho dos algoritmos com os atributos selecionados pela *toolbox* do PRTools para a análise entre os grupos controle e restritivo se mostrou inferior ao encontrado através da técnica de seleção pelo algoritmo de AMI e pelo especialista.

5.5. Experimento com três atributos selecionados pelo melhor classificador de Aprendizado em Múltiplas Instâncias

Nessa etapa, todas as combinações com 3 atributos foram utilizadas. Para isso, foram realizadas análises anteriores, que são mostradas no apêndice 2. Os atributos escolhidos para serem utilizados nos algoritmos de aprendizado de máquinas se deu da seguinte forma: S , R_m e C_{din} , para a análise Controle *versus* Normal Exame e R_0 , Z_{rs} e C_{din} , para a análise Controle *versus* Restritivo.

Tabela 22 - Resultados obtidos na seleção de 3 atributos através do melhor desempenho do AMI

		Controle x Normal Exame			Controle x Restritivo		
		Sensibilidade (%)	Especificidade (%)	AUC	Sensibilidade (%)	Especificidade (%)	AUC
AMI	BFP	89,40	52,20	0,77	88,90	84,40	0,94
	KNN	92,20	62,20	0,82	92,20	86,70	0,95
	ADAB	86,40	62,20	0,81	87,80	95,60	0,97
	RF	69,70	74,40	0,81	90,00	92,20	0,96
	RSVM	69,70	66,70	0,70	88,90	88,90	0,94
	AMI	75,80	66,70	0,75	88,90	85,60	0,94

Através dos resultados obtidos, pode-se dizer que o melhor desempenho, na análise Controle *versus* Normal Exame, foi alcançado pelo algoritmo KNN, com AUC = 0,82, seguida pelos algoritmos ADAB e RF, ambos com AUC = 0,81. Esses resultados indicam que os algoritmos alcançaram acurácia diagnóstica moderada. A curva ROC dos classificadores é mostrada na Figura 38

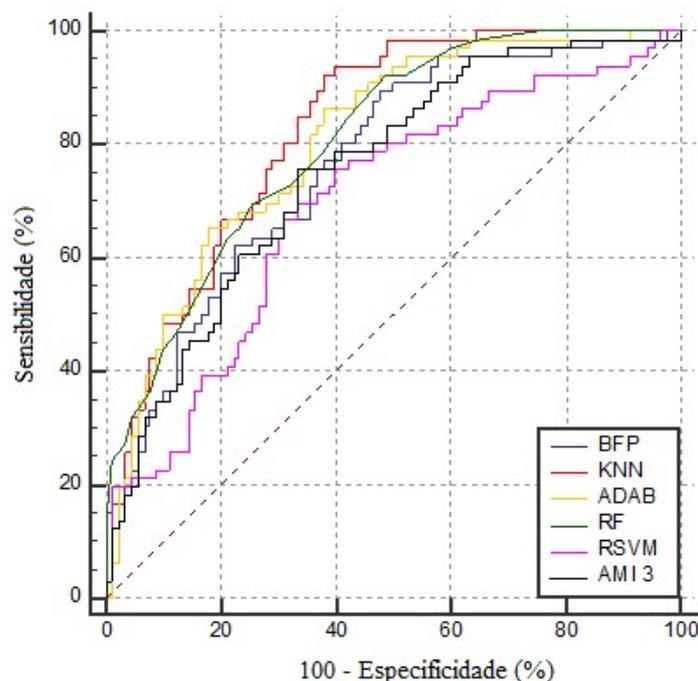


Figura 38 - Curvas ROC obtidas no experimento com os parâmetros da TOF selecionados pelo melhor desempenho do Classificador AMI com 3 atributos para os grupos controle e normal exame

A Tabela 23 mostra a comparação das AUC obtidas com o melhor parâmetro da TOF contra os demais classificadores e o erro padrão.

Tabela 23 - Comparação das AUC entre os modelos com três atributos selecionados pelo melhor desempenho AMI – Análise Controle *versus* Normal Exame

	KNN	ADAB	RF	RSVM	AMI(3)
BFP	0,0606±0,0388	0,0397±0,0354	0,0405±0,0302	0,0705±0,0304*	0,0152±0,0224
KNN	-	0,0209±0,0274	0,0201±0,0282	0,131±0,0432 ⁺	0,0758±0,0401
ADAB	-	-	0,0007±0,02	0,110±0,0414**	0,0549±0,0383
RF	-	-	-	0,111±0,0356 ⁺	0,0556±0,0341
RSVM	-	-	-	-	0,0554±0,0301

BFP: *Best FOT Parameter* (Obtido sem o uso de classificadores; C_{din});

ADAB: *Adaboost*; 1-NN: *K-Nearest Neighbor* (K = 1); RF: *Random Forests*;

RSVM: *Radial Support Vector Machine*; AMI: *Aprendizado em Múltiplas Instâncias*

* $p < 0,05$

** $p < 0,01$

⁺ $p < 0,005$

⁺⁺ $p < 0,0001$

Por outro lado, na análise Controle *versus* Restritivo, pode-se dizer que o melhor desempenho foi alcançado pelo algoritmo ADAB, com $AUC = 0,97$, seguida pelo algoritmo RF, com $AUC = 0,96$. Esses resultados indicam que os algoritmos alcançaram alta acurácia diagnóstica. A curva ROC dessa análise é mostrada na Figura 39.

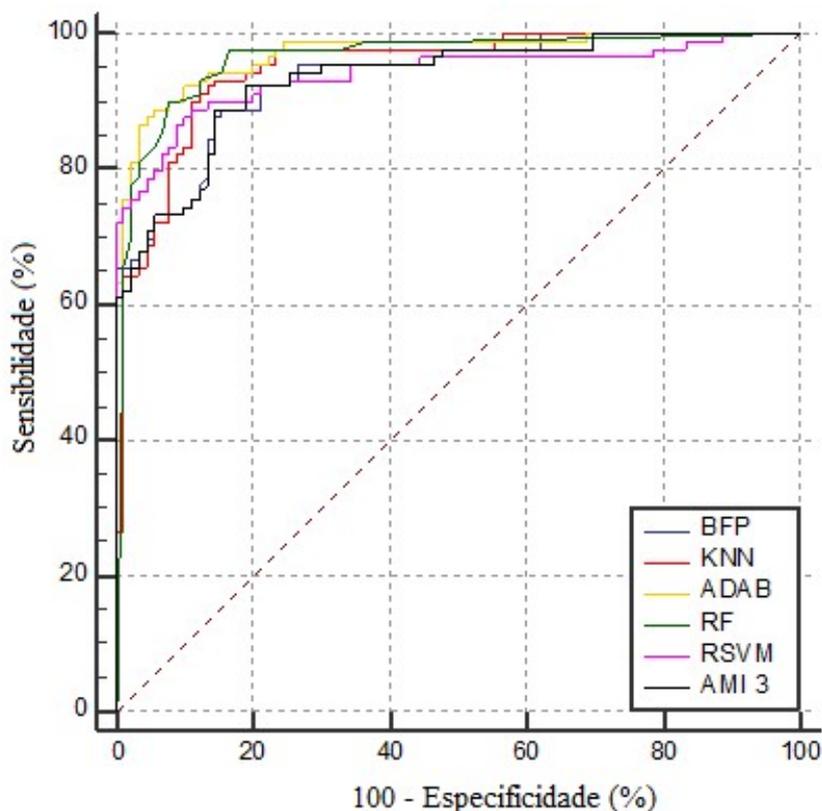


Figura 39 - Curvas ROC obtidas no experimento com os parâmetros da TOF selecionados pelo melhor desempenho do Classificador AMI com 3 atributos para os grupos controle e restritivo

A Tabela 24 mostra a comparação das AUCs obtidas com o melhor parâmetro da TOF e com os demais classificadores desenvolvidos com base na escolha dos atributos pelo melhor resultado obtido através do Aprendizado em Múltiplas Instâncias, com 3 atributos. A interseção entre linha e coluna indica a diferença da AUC entre dois classificadores.

Tabela 24 - Comparação das AUCs entre os modelos com três atributos selecionados pelo melhor desempenho AMI - Análise Controle versus Restritivo

	KNN	ADAB	RF	RSVM	AMI(3)
BFP	0,0131±0,0188	0,0349±0,0156*	0,0280±0,0154	0,00451±0,0102	0,00235±0,00192
KNN	-	0,0217±0,0120	0,0149±0,0164	0,00864±0,0218	0,0155±0,0194
ADAB	-	-	0,00685±0,0102	0,0304±0,0180	0,0372±0,0164*
RF	-	-	-	0,0235±0,0187	0,0304±0,0160
RSVM	-	-	-	-	0,00685±0,00974

BFP: *Best FOT Parameter* (Obtido sem o uso de classificadores; C_{din});

ADAB: *Adaboost*; 1-NN: *K-Nearest Neighbor* ($K = 1$); RF: *Random Forests*;

RSVM: *Radial Support Vector Machine*; AMI: Aprendizado em Múltiplas Instâncias

* $p < 0,05$

** $p < 0,01$

⁺ $p < 0,005$

⁺⁺ $p < 0,0001$

No apêndice 6 estão disponíveis os gráficos com as análises com a sensibilidade fixada em 75% e 90%, para ambos os casos analisados acima.

5.6. Experimento com três atributos escolhidos através da seleção direta para os grupos controle e normal exame

Através da seleção direta de 3 atributos, foram escolhidos R_0 , Z_{rs} e C_{din} para a análise Controle versus Normal exame e X_m , R_0 e C_{din} para a análise Controle versus Restritivo. A Tabela 25 mostra os resultados obtidos para ambas as análises.

Tabela 25 – Resultados obtidos na Seleção Direta de 3 atributos

	Controle x Normal Exame			Controle x Restritivo			
	Sensibilidade (%)	Especificidade (%)	AUC	Sensibilidade (%)	Especificidade (%)	AUC	
Seleção Direta	BFP	89,40	52,20	0,77	88,90	84,40	0,94
	KNN	98,50	56,70	0,82	92,20	86,70	0,95
	ADAB	71,20	77,80	0,79	93,30	90,00	0,98
	RF	75,80	71,10	0,82	92,20	91,10	0,97
	RSVM	86,40	67,80	0,82	88,90	88,90	0,94
	AMI	83,30	57,80	0,72	73,30	84,40	0,82

Através dos resultados obtidos, pode-se dizer que os melhores desempenhos foram alcançados pelos algoritmos KNN, RF e RSVM, para a análise controle *versus* Normal Exame, ambos com uma $AUC = 0,82$. Esses resultados indicam que os algoritmos alcançaram acurácia diagnóstica moderada. A Figura 40 ilustra as curvas ROC de cada classificado.

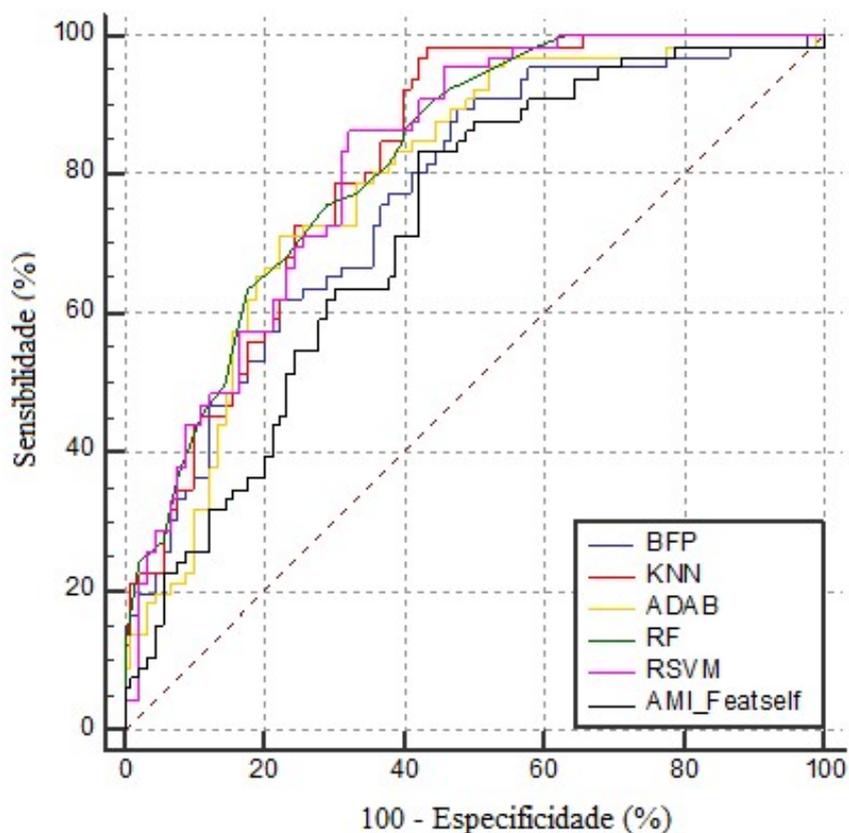


Figura 40 - Curvas ROC obtidas no experimento com 3 atributos da TOF escolhidos através da seleção direta para os grupos Controle e Normal Exame

A Tabela 26 mostra a comparação das AUCs obtidas com o melhor parâmetro da TOF e com os demais classificadores desenvolvidos com base na escolha de atributos pelo melhor resultado obtido através da função da seleção direta, com 3 atributos. Na análise Controle versus Normal Exame

Tabela 26 - Comparação das AUC entre os modelos com três atributos escolhidos através da seleção direta – Análise Controle *versus* Normal Exame

	KNN	ADAB	RF	RSVM	AMI(Carac3)
BFP	0,0515±0,0382	0,0222±0,0391	0,0530±0,0360	0,0502±0,0366	0,0455±0,0198*
KNN	-	0,0293±0,0244	0,00152±0,0224	0,00135±0,0281	0,097±0,0431*
ADAB	-	-	0,0308±0,0153*	0,0279±0,0296	0,0677±0,0432
RF	-	-	-	0,00286±0,0268	0,0985±0,0398*
RSVM	-	-	-	-	0,0956±0,0408*

BFP: *Best FOT Parameter* (Obtido sem o uso de classificadores; C_{din});

ADAB: *Adaboost*; 1-NN: *K-Nearest Neighbor* ($K = 1$); RF: *Random Forests*;

RSVM: *Radial Support Vector Machine*; AMI: Aprendizado em Múltiplas Instâncias

* $p < 0,05$

** $p < 0,01$

⁺ $p < 0,005$

⁺⁺ $p < 0,0001$

Por outro lado, através dos resultados obtidos para a análise Controle *versus* Restritivo, pode-se dizer que o melhor desempenho foi alcançado pelo algoritmo ADAB, com $AUC = 0,98$. Esses resultados indicam que o algoritmo alcançou alta acurácia diagnóstica. Esse resultado foi seguido pelo algoritmo RF, com $AUC = 0,97$ e também com alta acurácia diagnóstica. A curva ROC para essa análise é mostrada na Figura 41.

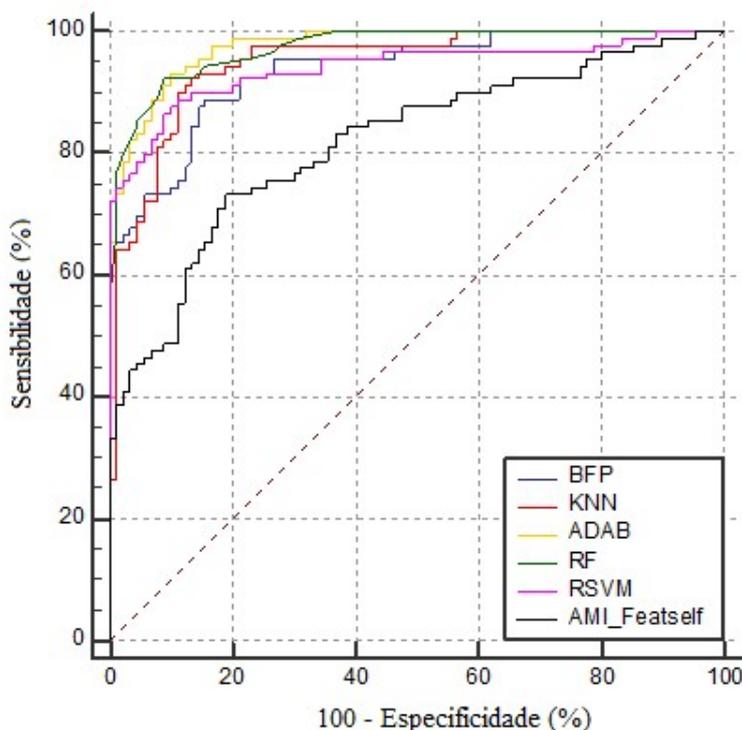


Figura 41 - Curvas ROC obtidas no experimento com 3 atributos da TOF escolhidos através da seleção direta – Análise Controle *versus* Normal Restritivo

A Tabela 27 mostra a comparação das AUCs obtidas com o melhor parâmetro da TOF e com os demais classificadores desenvolvidos com base na escolha dos atributos pelo melhor resultado obtido através da seleção direta de atributos, para a análise Controle *versus* Restritivo.

Tabela 27 - Comparação das AUC entre os modelos com três atributos escolhidos através da seleção direta para os grupos Controle *versus* Restritivo

	KNN	ADAB	RF	RSVM	AMI(Featself)
BFP	0,0515±0,0382	0,0222±0,0391	0,0530±0,0360	0,0502±0,0366	0,0455±0,0198*
KNN	-	0,0293±0,0244	0,00152±0,0224	0,00135±0,0281	0,097±0,0431*
ADAB	-	-	0,0308±0,0153*	0,0279±0,0296	0,0677±0,0432
RF	-	-	-	0,00286±0,0268	0,0985±0,0398*
RSVM	-	-	-	-	0,0956±0,0408*

BFP: *Best FOT Parameter* (Obtido sem o uso de classificadores; C_{din});

ADAB: *Adaboost*; 1-NN: *K-Nearest Neighbor* ($K = 1$); RF: *Random Forests*;

RSVM: *Radial Support Vector Machine*; AMI: *Aprendizado em Múltiplas Instâncias*

* $p < 0,05$ ** $p < 0,01$ + $p < 0,005$ ++ $p < 0,0001$

No apêndice 7 estão ilustrados os gráficos com as análises com a sensibilidade fixada em 75% e 90%, para ambos os casos.

5.7. Conclusões do capítulo

A fim de consolidar os resultados, a Figura 42 e a Figura 43 resumem todos os resultados obtidos com as diferentes metodologias abordadas nessa dissertação. A linha tracejada representa o resultado da BFP e os gráficos de barras os resultados obtidos com os experimentos.

É possível notar que o melhor resultado para a análise entre os grupos controle e normal exame foi alcançado pelo algoritmo *K-Nearest Neighbor* (KNN), com a utilização de todos os parâmetros da TOF. Para esses grupos, devido à similaridade entre os dados, a redução de atributos não se mostrou eficiente em aumentar a acurácia diagnóstica, porém, a utilização dos algoritmos de aprendizado de máquinas se mostrou eficaz.

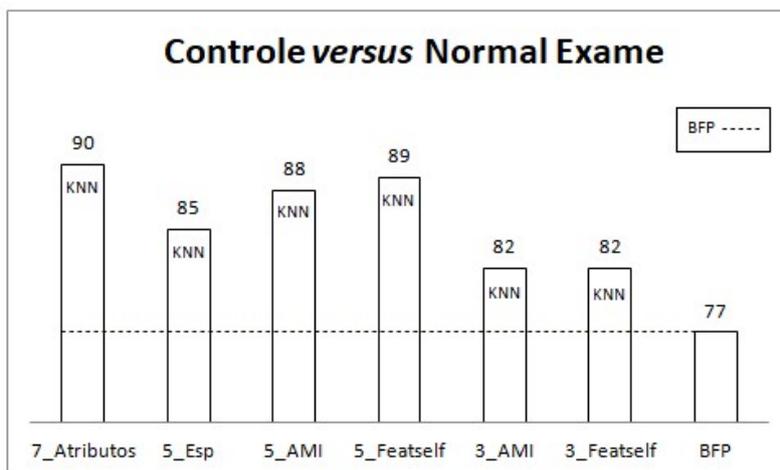


Figura 42 - Comparativo entre as acurácias obtidas nas seis metodologias em comparação com o BFP para a análise *Controle versus Normal Exame*

Por outro lado, na análise entre os grupos controle e restritivo, mostradas na Figura 43, os melhores resultados foram alcançados pelo algoritmo *Adaboost* (ADAB). Em três dos testes, os resultados foram similares em termos da AUC, porém, quando

comparada a sensibilidade entre os modelos, o melhor resultado foi obtido através da seleção de atributos pelo algoritmo AMI. Dessa forma, o ADAB se mostrou com maior sensibilidade e maior especificidade que nas demais análises.

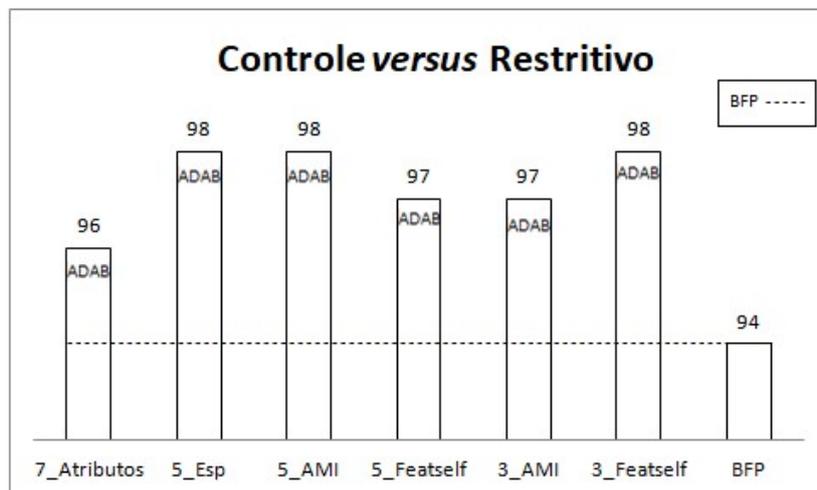


Figura 43 - Comparativo entre as acurácias obtidas nas seis metodologias testadas em comparação com o BFP para a análise *Controle versus Restritivo*

Conclusão

A utilização de algoritmos de aprendizado de máquinas, para detecção de alterações respiratórias em portadores de esclerose sistêmica, através da técnica de oscilações forçadas, mostrou-se eficiente, aumentando a acurácia diagnóstica.

Durante todo o processo, os classificadores *K-Nearest Neighbor*, *Adaboost*, *Radial Support Vector Machine*, *Random Forest* e Aprendizado em Múltiplas Instâncias apresentaram valores de AUC maiores do que os valores obtidos pelo melhor parâmetro da TOF, confirmando a hipótese proposta inicialmente, associada ao possível aprimoramento no diagnóstico.

Durante os testes realizados, a complacência dinâmica (C_{din}) foi o atributo que obteve o melhor desempenho individual em ambas as análises realizadas - controle *versus* normal exame e controle *versus* restritivo. Porém, na primeira análise, se verificava uma dificuldade em separar os grupos, pois os pacientes pertencentes a ambos estão no limiar entre ter ou não a doença, fazendo com que os atributos da TOF tivessem valores muito próximo entre os grupos. Isso não ocorreu na segunda análise, onde os grupos controle e restritivo possuíam uma separação bem delineada.

De toda forma, quando os atributos da Esclerose Sistêmica foram aplicados aos algoritmos de aprendizado de máquinas, foi possível obter valores considerados com alta acurácia para a análise controle *versus* normal exame, comprovando que a utilização de algoritmos de aprendizado de máquinas pode servir como apoio às decisões clínicas, na investigação de detecções respiratórias decorrentes da Esclerose Sistêmica. A utilização dos classificadores KNN e *Adaboost* permitiu melhorar, significativamente, a acurácia atingindo alta acurácia no diagnóstico clínico.

A decisão de reduzir atributos foi pensada levando em consideração os custos inerentes a obtenção dos dados. A Esclerose Sistêmica é uma doença de baixa frequência e com diferentes tipos de manifestações clínicas, dificultando a condução de pesquisas com grande número de pacientes. Ela deve analisada com muito cuidado, pois o estudo mostrou que a acurácia diagnóstica pode melhorar com o uso de menos atributos da TOF, mas em alguns casos a sensibilidade e especificidade do algoritmo tornam-se baixas, podendo levar a falsos positivos e a falsos negativos.

A utilização do algoritmo de AMI se mostrou eficiente na seleção de atributos, onde conseguiu atingir melhor resultado para a análise controle *versus* normal exame que o conseguido pela seleção do especialista, porém o algoritmo de AMI não se mostrou tão eficiente quanto os classificadores tradicionais.

Referências

- ALLANORE, Y.; AVOUAC, J.; KAHAN, A. Systemic sclerosis: an update in 2008. *Joint Bone Spine*, 75, p. 650-655, 2008.
- AMARAL, J. L. M. et al. Machine learning algorithms and forced oscillation measurements applied to the automatic identification of chronic obstructive pulmonary disease. *Computer Methods and Programs in Biomedicine*, p. Volume 105(3), 183-194, 2012.
- AMARAL, J. L. M. et al. An improved method of early diagnosis of smoking-induced respiratory changes using machine learning algorithms. *Computer methods and programs in biomedicine* 112, p. 441-454, 2013.
- AMARAL, J. L. M. et al. High-accuracy detection of airway obstruction in asthma using machine learning algorithms and forced oscillation measurements. *Computer Methods and Programs in Biomedicine*, p. 113-125, 2017.
- BREIMAN, L. Random forests. *Machine learning*, p. V.45 (1), 5-32, 2001.
- CASTANHO, M. J. P.; YAMAKAMI, A.; BARROS, L. C. Avaliação de um teste em medicina usando uma curva ROC fuzzy. *Biomatemática*, v. 14, p. 19-28, 2004.
- CHAVES, B. B. Estudo do algoritmo Adaboost de aprendizagem de máquina aplicado a sensores e sistemas embarcados. *Dissertação de mestrado, Escola Politécnica, Universidade de São Paulo, São Paulo: [s.n.], 2011.*
- CHEPLYGINA, V. et al. Classification of COPD with Multiple Instance Learning. *International Conference on Pattern Recognition (ICPR)*, 2014.
- COSTA, H. S. R. M. Estudo comparativo de abordagens ao problema de débito de transações bancárias em contas com saldo insuficiente no segmento de negócios (Dissertação). *Departamento de Matemática Aplicada - Faculdade de Ciências da Universidade do Porto. [S.l.]. 2012.*
- DELONG, E. R.; DELONG, D. M.; CLARKE-PEARSON, D. L. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, p. 44, 837-845, 1988.
- DI MANGO, A. M. et al. Changes in respiratory mechanics with increasing degrees of airway obstruction in COPD: detection by forced oscillation technique. *Respir Med*, p. 100(3), 399-410, 2006.
- DIETTERICH, T. G. Ensemble methods in machine learning. *Proceedings of the First International Workshop on Multiple Classifier Systems*, 2000.

- DIETTERICH, T. G.; LATHROP, R. H.; LOZANO-PEREZ, T. Solving the multiple instance problem with axis-parallel rectangles. *Artificial Intelligence*, p. Vol.89, 31-71, 1997.
- DU BOIS, A. B. et al. Oscillation mechanics of lungs and chest in man. *J. Appl Physiol*, p. 587-594, 1956(8).
- DU BOIS, R. M. Mechanisms of scleroderma-induced lung disease. *Proc. Am. Thorac. Soc.*, v. 4, p. 434-438, 2007.
- FACELI, K. et al. *Inteligência Artificial: Uma abordagem de Aprendizado de Máquina*. Rio de Janeiro: LTC, 2011.
- FARIA, A. C. D. Efeito da carga tabágica sobre as propriedades resistivas e reativas do sistema respiratório, 2007.
- FARIA, A. C. D. et al. Alterações precoces nas propriedades resistivas e elásticas do sistema respiratório decorrentes do tabagismo. *J. bras. pneumol.*, p. Volume 30(3), S69-S70, 2004.
- FARIA, A. W. C. Uma nova abordagem para aprendizado de múltiplas instâncias, baseada em seleção de instâncias via estimador de densidade por kernel, p. Tese (Doutorado) - Universidade Federal de Minas Gerais, Escola de Engenharia, 2016.
- FAWCETT, T. An introduction to ROC analysis. *Pattern Recognition Letters*, v. 27, n. 8, p. 861-874, 2006.
- FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *EuroCOLT*, London, p. 23-37, 1995.
- FREUND, Y.; SCHAPIRE, R. E. A short introduction to Boosting. *Journal of Japanese Society for Artificial Intelligence*, v. 14, n. 5, p. 771-780, 1999.
- GAMA, J. Functional Trees. *Machine Learning*, p. (55), 219-250, 2004.
- GUYON, I. et al. Structural risk minimization for character recognition. *Neural Information Processing Systems*, v. 4, 1992.
- GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. *Journal of Machine Learning Research* 3 , p. 1157-1182, 2003.
- HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, 2008.
- HAYKIN, S. *Redes Neurais: Princípios e prática*. [S.l.]: Bookman, 2001.
- HORTA, R. A. M. et al. Comparação de Técnicas de Seleção de Atributos para Previsão de Insolvência de Empresas Brasileiras no Período 2005-2007. XXXIV Encontro da ANPAD, Rio de Janeiro, 2010.

- IBAÑEZ, M. M. Uso de redes neurais nebulosas e florestas aleatórias na classificação de imagens em um projeto de ciência cidadã. Dissertação (Mestrado em Computação Aplicada) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos: INPE, 2016.
- KEVIN, M. A.; DA SILVA, E. M. Utilização do Algoritmo de Máquina de Vetores de Suporte (SVM) para Predição de Dados Climáticos. *Encoinfo*, p. 149-158, 2016.
- LANGLEY, P.; SIMON, H. A. Applications of machine learning and rule induction. *Communications of the ACM*, p. 54-64, 1995.
- MARON, O. Learning from Ambiguity, p. Doctoral Thesis, Dept. of Electrical Engineering and Computer Science, M.I.T., 1998.
- MARON, O.; LOZANO-PEREZ, T. A framework for multiple-instance learning. *Advances In Neural Information Processing Systems (NIPS)*, p. 570-576, 1998.
- MARON, O.; RATAN, A. L. Multiple-instance learning for natural scene classification. *International Conference on Machine Learning*, p. 341-349, 1998.
- MASI, A. T.; RODNAN, G. P.; MEDSGER, T. A. J. Preliminary criteria for the classification of systemic sclerosis (scleroderma). *Arthritis Rheumatology*, p. 581-590, 1980.
- MELO, P. L. Técnica de oscilações forçadas na prática pneumológica: Princípios e exemplos de potenciais aplicações. *Pulmão*, RJ, v. 24(1), p. 42-48, 2015.
- MELO, P. L.; WERNECK, M. M.; GIANELLA-NETO, A. Avaliação da mecânica ventilatória por oscilações forçadas: fundamentos e aplicações clínicas. *J. Pneumologia*, p. 194-206, 2000(a).
- MELO, P. L.; WERNECK, M. M.; GIANNELLA-NETO, A. New impedance spectrometer for scientific and clinical studies of the respiratory system. *Rev Sci Instr*, 71 (7), p. 2867-2872, 2000.
- MIRANDA, I. A. Análise da mecânica ventilatória de indivíduos portadores de esclerose sistêmica através da Técnica de Oscilações Forçadas. [S.l.]. 2010.
- MIRANDA, I. A. et al. Oscilações forçadas na avaliação funcional de indivíduos portadores de esclerodermia. *Pulmão RJ*, V.15(1), p. 2-6, 2006.
- MIRANDA, I. A. et al. On the Respiratory Mechanics Measured by Forced Oscillation Technique in Patients with Systemic Sclerosis. *PLoS ONE*, v. 8(4), abril 2013.
- MITCHELL, T. M. *Machine Learning*. [S.l.]: McGraw-Hill Science, 1997.
- NAGELS, J. et al. Mechanical properties of lungs and chest wall during spontaneous breathing. *J Appl Physiol*, p. 49(3): 408-416, 1980.

- OMAR, A. M.; RAED, A. D.; ALEJANDRO, C. A. Manifestations of scleroderma pulmonary disease. *Clinics in Chest Medicine*, p. 713-731, 1998.
- OOSTVEEN, E. et al. The forced oscillation technique in clinical practice: methodology, recommendations and future developments. *Eur Respir J.*, p. Volume 22(6), 1026-1041, 2003.
- OSTOJIC, P. et al. Interstitial lung disease in systemic sclerosis. *Lung*, p. 211-220, 2007.
- PARSONS, S.; BIGHAM, J. Possibility theory and the generalised noisy or model. *Proceedings of the Sixth International Conference on Information Processing and the Management of Uncertainty*, 1996. 853-858.
- PRATI, R. Novas abordagens em aprendizado de máquina para a geração de regras, classes desbalanceadas e ordenação de casos. Universidade de São Paulo. São Paulo - SP. 2006.
- QUISPE, S. C. Uma abordagem visual para apoio ao aprendizado multi-instâncias. Dissertação de mestrado, p. USP - São Paulo - SP, 2015.
- RAY, S.; CRAVEN, M. Supervised versus Multiple Instance Learning: an empirical comparison. *Proceedings of the 22nd international conference on machine learning*, p. 697-704, 2005.
- REFAEILZADEH, L.; TANG, L.; LIU, H. Cross Validation, *Encyclopedia of Database Systems*. [S.l.]: Springer, 2009.
- SANTOS, C. A. M. Análise de Características para Detecção de Nudez em Imagens, Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas, Manaus, 2012.
- SANTOS, M. K.; FARIA, F. B.; TRAD, C. S. Comprometimento pulmonar na esclerose sistêmica: revisão de casos. *Radiol Bras.*, p. 181-184, 2006.
- SCHAPIRE, R. E. Explaining adaboost. *Empirical Inference*, p. 37-52, 2013.
- SCHMIDHEINY, K. The bootstrap. *Short Guides to Microeconometrics* - Universität Basel, 2012.
- TAKAHASHI, A. Máquina de vetores-suporte intervalar , p. Tese (Doutorado) - Universidade Federal do Rio Grande do Norte. Centro de Tecnologia. Programa de Pós-Graduação em Engenharia Elétrica, 2012.
- VAPNIK, V. N. *The nature of statistical learning theory*. New York: Springer-Verlag, 1995.

WITTEN, I. H.; FRANK, E. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann, n. Segunda Edição, 2005.

ZOU, K. H. Comparison of Correlated Receiver Operating Characteristic Curves Derived from Repeated Diagnostic Test Data. Academic Radiology, p. 8, 225-233, 2001.

ZOU, K. H.; O'MALLEY, A. J.; MAURI, L. Receiver-Operating Characteristic Analysis for Evaluating Diagnostic Tests and Predictive Models. Statistical Primer for Cardiovascular Research, n. 115, p. 654-657, 2007.

Apêndice 1 - Valores dos atributos da TOF

O gráfico (a) da Figura 25 explicita a comparação da frequência de ressonância (F_r) entre os três grupos analisados. Os valores médios apresentados foram: $14,86 \pm 5,04$, $14,16 \pm 3,96$ e $19,71 \pm 5,38$ para o grupo controle, normal exame e restritivo, respectivamente.

O gráfico (b) da Figura 25 apresenta a comparação da reatância média (X_m) entre os três grupos analisados. Os valores médios apresentados foram: $0,23 \pm 0,36$, $0,27 \pm 0,41$ e $0,43 \pm 0,85$ para o grupo controle, normal exame e restritivo, respectivamente.

O gráfico (c) da Figura 25 apresenta a comparação da reatância média (R_0) entre os três grupos analisados. Os valores médios apresentados foram: $2,27 \pm 0,70$, $2,98 \pm 0,84$ e $3,62 \pm 1,41$ para o grupo controle, normal exame e restritivo, respectivamente.

O gráfico (d) da Figura 25 apresenta a comparação da inclinação da curva de resistência (S) entre os três grupos analisados. Os valores médios apresentados foram: $-5,19 \pm 21,77$, $-6,49 \pm 25,21$ e $-40,95 \pm 57,93$ para o grupo controle, normal exame e restritivo, respectivamente.

O gráfico (e) da Figura 25 apresenta a comparação da inclinação resistência média (R_m) entre os três grupos analisados. Os valores médios apresentados foram: $2,21 \pm 0,6949$, $2,91 \pm 0,74$ e $3,21 \pm 1,03$ para o grupo controle, normal exame e restritivo, respectivamente.

O gráfico (f) da Figura 25 apresenta a comparação da impedância média (Z_{4Hz}) entre os três grupos analisados. Os valores médios apresentados foram: $2,95 \pm 1,18$, $3,92 \pm 0,97$ e $5,38 \pm 2,17$ para o grupo controle, normal exame e restritivo, respectivamente.

O gráfico (g) da Figura 25 apresenta a comparação da complacência dinâmica (C_{din}) entre os três grupos analisados. Os valores médios apresentados foram: $0,036 \pm 0,0295$, $0,0187 \pm 0,0137$ e $0,0116 \pm 0,0055$ para o grupo controle, normal exame e restritivo, respectivamente.

Apêndice 2 - Combinações do AMI

Tabela 28 - Combinações de 5 atributos do AMI para os grupos controle *versus* normal exame

Atributos	AUC
1, 2, 3, 4, 5	0,7012
1, 2, 3, 4, 6	0,7304
1, 2, 3, 4, 7	0,7248
1, 2, 3, 5, 6	0,7271
1, 2, 3, 5, 7	0,7324
1, 2, 3, 6, 7	0,7205
1, 2, 4, 5, 6	0,7285
1, 2, 4, 5, 7	0,7301
1, 2, 4, 6, 7	0,7234
1, 2, 5, 6, 7	0,7177
1, 2, 4, 5, 6	0,7205
1, 3, 4, 5, 7	0,7002
1, 3, 4, 6, 7	0,7247
1, 3, 5, 6, 7	0,7221
1, 4, 5, 6, 7	0,7316
2, 3, 4, 5, 6	0,7324
2, 3, 4, 5, 7	0,7372
2, 3, 4, 6, 7	0,7127
2, 3, 5, 6, 7	0,7295
2, 4, 5, 6, 7	0,7163
3, 4, 5, 6, 7	0,7135

onde:

1 = F_r , 2 = X_m , 3 = R_0 , 4 = S , 5 = R_m , 6 = Z_{rs} , 7 = C_{din}

Tabela 29 - Combinações de 5 atributos do AMI para os grupos controle *versus* restritivo

Atributos	AUC
1, 2, 3, 4, 5	0,8283
1, 2, 3, 4, 6	0,8786
1, 2, 3, 4, 7	0,901
1, 2, 3, 5, 6	0,872
1, 2, 3, 5, 7	0,884
1, 2, 3, 6, 7	0,8761
1, 2, 4, 5, 6	0,8786
1, 2, 4, 5, 7	0,9002
1, 2, 4, 6, 7	0,8992
1, 2, 5, 6, 7	0,8899
1, 2, 4, 5, 6	0,8786
1, 3, 4, 5, 7	0,9011
1, 3, 4, 6, 7	0,8901
1, 3, 5, 6, 7	0,9383
1, 4, 5, 6, 7	0,9301
2, 3, 4, 5, 6	0,8701
2, 3, 4, 5, 7	0,9294
2, 3, 4, 6, 7	0,9289
2, 3, 5, 6, 7	0,9294
2, 4, 5, 6, 7	0,92
3, 4, 5, 6, 7	0,9002

onde:

$1 = F_r, 2 = X_m, 3 = R_0, 4 = S, 5 = R_m, 6 = Z_{rs}, 7 = C_{din}$

Tabela 30: Combinações de 3 atributos do AMI para os grupos controle *versus* normal exame

Atributos	AUC	Atributos	AUC
1, 2, 3	0,7402	2, 3, 7	0,7281
1, 2, 4	0,6152	2, 4, 5	0,7315
1, 2, 5	0,7301	2, 4, 6	0,7388
1, 2, 6	0,7261	2, 4, 7	0,7375
1, 2, 7	0,7006	2, 5, 6	0,7327
1, 3, 4	0,7254	2, 5, 7	0,7394
1, 3, 5	0,7215	2, 6, 7	0,7316
1, 3, 6	0,7227	3, 4, 5	0,7364
1, 3, 7	0,7006	3, 4, 6	0,7175
1, 4, 5	0,7021	3, 4, 7	0,7419
1, 4, 6	0,7132	3, 5, 6	0,7254
1, 4, 7	0,7352	3, 5, 7	0,7449
1, 5, 6	0,7369	3, 6, 7	0,7118
1, 5, 7	0,7012	4, 5, 6	0,7247
1, 6, 7	0,736	4, 5, 7	0,7473
2, 3, 4	0,7251	4, 6, 7	0,7187
2, 3, 5	0,7338	5, 6, 7	0,7210
2, 3, 6	0,7382		

onde:

$1 = F_r, 2 = X_m, 3 = R_0, 4 = S, 5 = R_m, 6 = Z_{rs}, 7 = C_{din}$

Tabela 31 - Combinações de 3 atributos do AMI para os grupos controle *versus* restritivo

Atributos	AUC	Atributos	AUC
1, 2, 3	0,8226	2, 3, 7	0,9002
1, 2, 4	0,7602	2, 4, 5	0,82
1, 2, 5	0,8283	2, 4, 6	0,8701
1, 2, 6	0,8786	2, 4, 7	0,9289
1, 2, 7	0,9201	2, 5, 6	0,8905
1, 3, 4	0,8226	2, 5, 7	0,93
1, 3, 5	0,8283	2, 6, 7	0,90
1, 3, 6	0,8786	3, 4, 5	0,82
1, 3, 7	0,8871	3, 4, 6	0,8473
1, 4, 5	0,8283	3, 4, 7	0,9008
1, 4, 6	0,8786	3, 5, 6	0,848
1, 4, 7	0,8983	3, 5, 7	0,9041
1, 5, 6	0,879	3, 6, 7	0,9337
1, 5, 7	0,9003	4, 5, 6	0,8443
1, 6, 7	0,8912	4, 5, 7	0,923
2, 3, 4	0,8176	4, 6, 7	0,9032
2, 3, 5	0,8169	5, 6, 7	0,8905
2, 3, 6	0,8701		

onde:

$$1 = F_r, 2 = X_m, 3 = R_0, 4 = S, 5 = R_m, 6 = Z_{rs}, 7 = C_{din}$$

Apêndice 3 - Análise comparativa da Sensibilidade e Especificidade através da seleção pelo Especialista

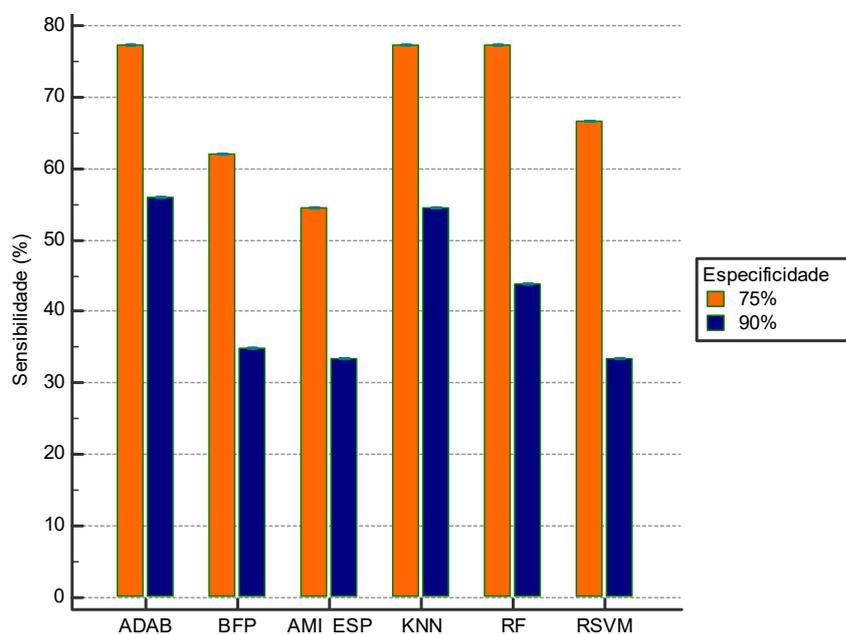


Figura 44 - Análise comparativa da sensibilidade obtida nos seis experimentos com a especificidade de 75% e 90% - Análise Controle *versus* Normal Exame

A Figura 44 mostra que com o valor da especificidade fixado em 75%, houve um aumento no valor da sensibilidade de todos os classificadores, quando comparados com o *BFP*, exceto no *AMI*. Dentre eles, o *ADAB*, *KNN* e *RF* ficaram na faixa de sensibilidade moderada.

De forma análoga, quando a especificidade foi fixada em 90%, também houve um aumento na sensibilidade do *ADAB*, *KNN* e *RF*, em comparação com o *BFP*, porém nenhum deles atingiu a faixa de sensibilidade moderada.

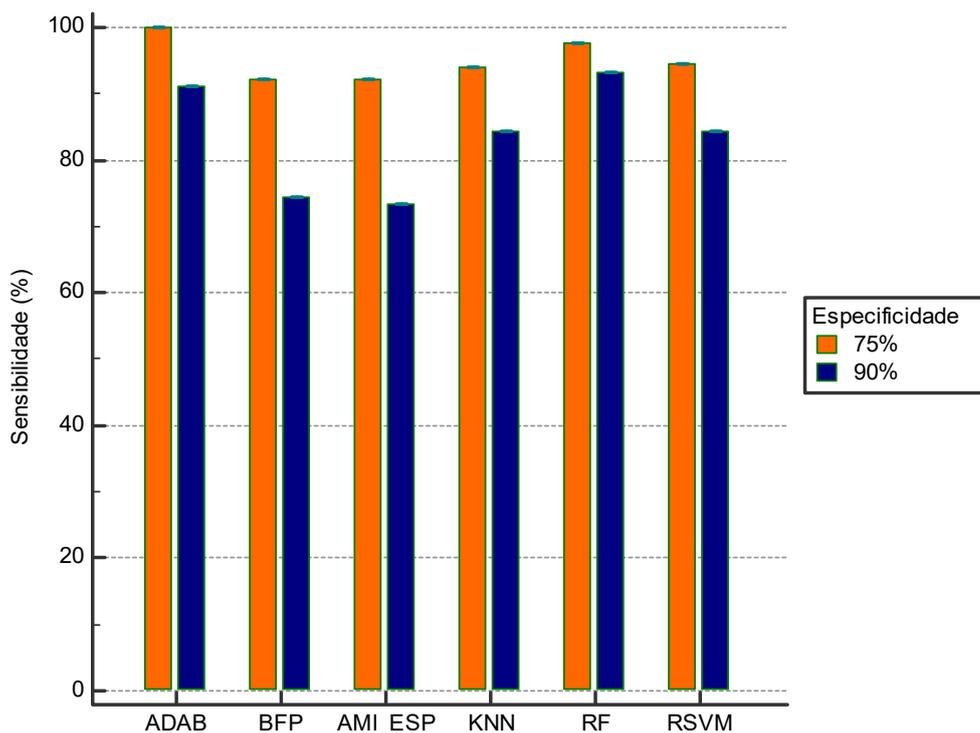


Figura 45 - Análise comparativa da sensibilidade obtida nos seis experimentos com a especificidade de 75% e 90% - Análise Controle *versus* Restritivo

A Figura 45 mostra que com o valor da especificidade fixado em 75%, houve um aumento no valor da sensibilidade dos classificadores ADAB, KNN, RF e RSVM, quando comparados com o *BFP*. O AMI teve desempenho similar ao *BFP*. Todos os classificadores ficaram na faixa de alta sensibilidade.

De forma análoga, quando a especificidade foi fixada em 90%, foi observado o mesmo comportamento da análise anterior, onde houve aumento na sensibilidade dos classificadores ADAB, KNN, RF e RSVM.

Apêndice 4 - Análise comparativa da Sensibilidade e Especificidade através da seleção pelo AMI

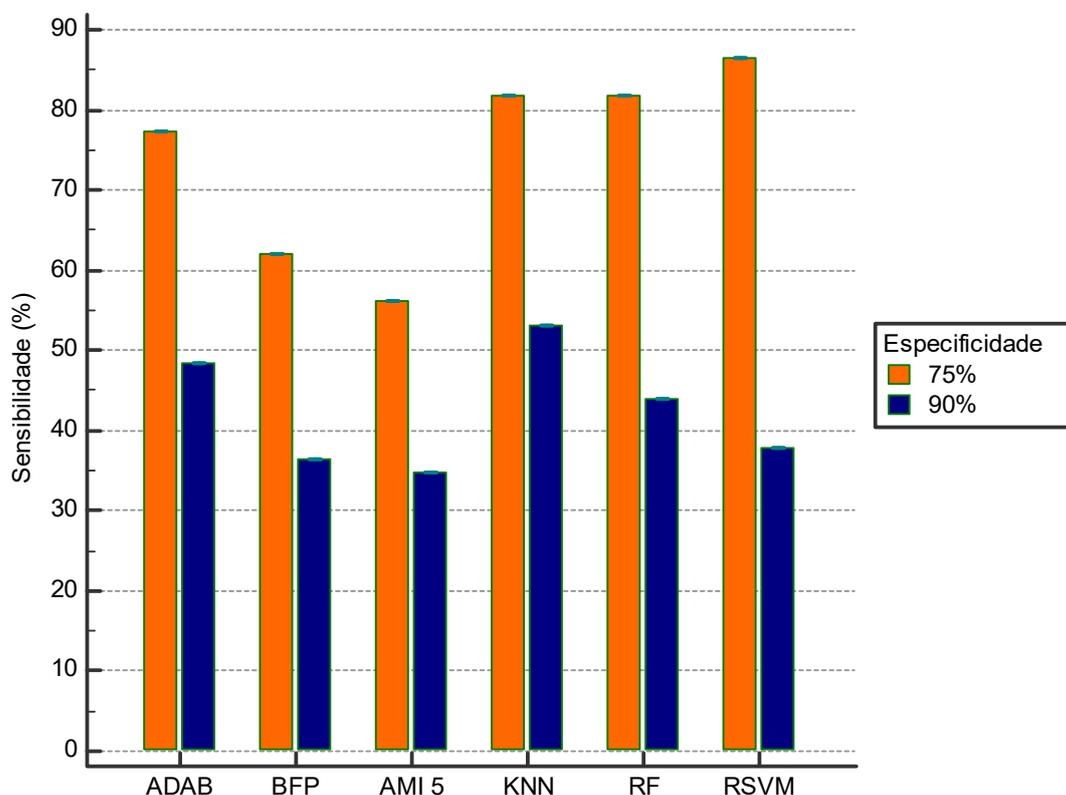


Figura 46 - Análise comparativa da sensibilidade obtida nos seis experimentos com a especificidade de 75% e 90% para os atributos selecionados pelo AMI - Análise Controle *versus* Normal Exame

A Figura 46 mostra que com o valor da especificidade fixado em 75% os algoritmos ADAB, KNN, RF e RSVM tiveram um aumento na sensibilidade, quando comparados com o BFP. Todos esses ficaram na faixa de sensibilidade moderada.

Analogamente ao resultado anterior, quando a especificidade foi fixada em 90% os algoritmos citados acima tiveram comportamento semelhante, porém todos ficaram abaixo da faixa de sensibilidade moderada.

Pode-se perceber que apenas um dos parâmetros selecionados pelo algoritmo de aprendizado em múltiplas instâncias divergiu da escolha do especialista como melhores

preditores da doença. Mesmo assim, houve um aumento no valor das AUCs dos algoritmos, quando comparadas com a seleção feita pelo especialista.

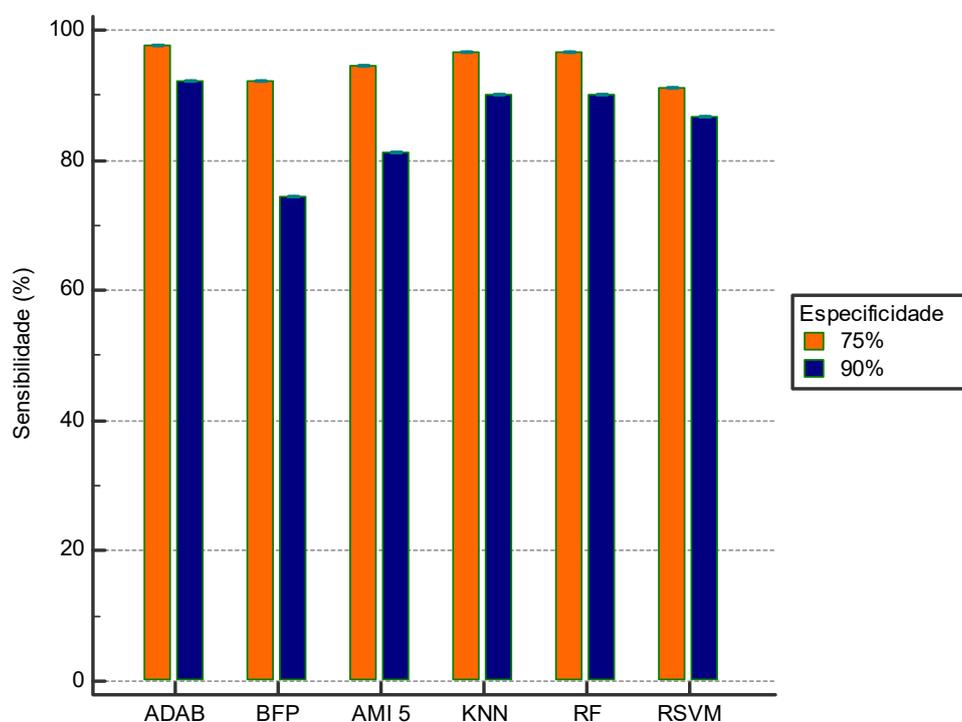


Figura 47 - Análise comparativa da sensibilidade obtida nos seis experimentos com a especificidade de 75% e 90% para os atributos selecionados pelo AMI - Análise Controle *versus* Restritivo

A Figura 47 mostra que com o valor da especificidade fixado em 75% todos os algoritmos tiveram um aumento na sensibilidade, quando comparados com o BFP. Todos eles ficaram na faixa de alta sensibilidade.

Na sequência, a especificidade foi fixada em 90% e os algoritmos citados acima tiveram comportamento semelhante, sempre melhores que o BFP, porém apenas o ADAB atingiu a faixa de alta sensibilidade.

Em comparação com a seleção pelo especialista, os resultados obtidos por essa técnica foram levemente inferiores para os grupos em análise. Mesmo assim, é possível dizer que os resultados são satisfatórios, pois atingiram alta acurácia diagnóstica.

Apêndice 5 - Análise comparativa da Sensibilidade e Especificidade através da seleção Direta

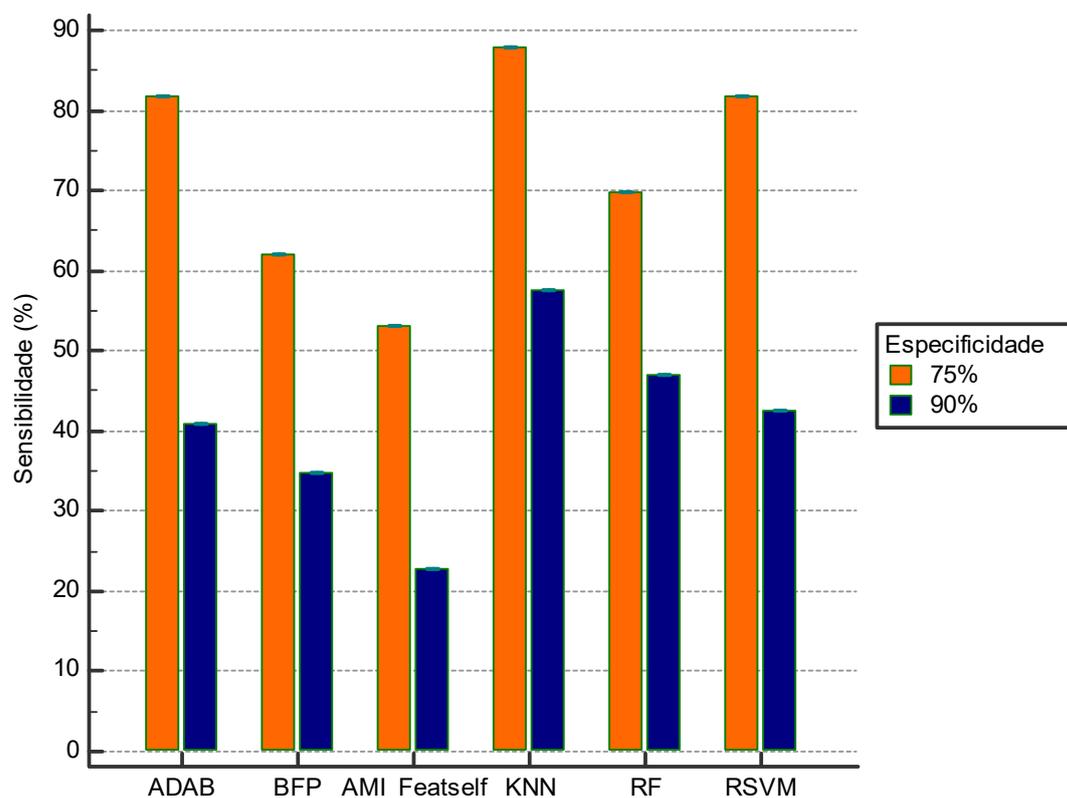


Figura 48 - Análise comparativa da sensibilidade obtida nos seis experimentos com a especificidade de 75% e 90% para os atributos escolhidos através da seleção direta - Análise Controle *versus* Normal Exame

A Figura 48 mostra que com o valor da especificidade fixado em 75% os algoritmos ADAB, KNN, RF e RSVM tiveram um aumento na sensibilidade, quando comparados com o BFP. Todos eles ficaram na faixa de alta sensibilidade.

Na sequência, quando a especificidade foi fixada em 90%, os algoritmos citados acima tiveram comportamento semelhante, sempre melhores que o BFP, porém abaixo da faixa de sensibilidade moderada.

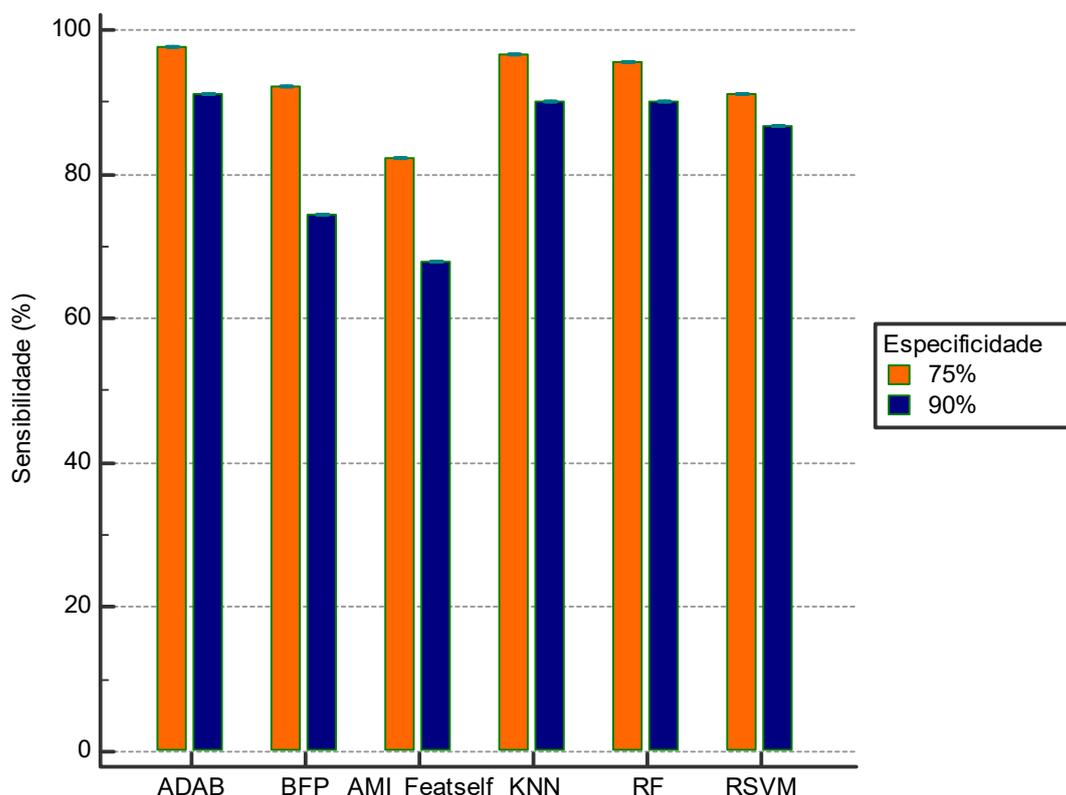


Figura 49 - Análise comparativa da sensibilidade obtida nos seis experimentos com a especificidade de 75% e 90% para os atributos escolhidos através da seleção direta - Análise Controle *versus* Restritivo

A Figura 49 mostra que com o valor da especificidade fixado em 75% os algoritmos ADAB, KNN, RF e RSVM tiveram um aumento na sensibilidade, quando comparados com o BFP. Todos eles ficaram na faixa de alta sensibilidade.

Na sequência, quando a especificidade foi fixada em 90%, os algoritmos citados acima tiveram comportamento semelhante, sempre melhores que o BFP, e todos na faixa de sensibilidade moderada, exceto o algoritmo de AMI, que ficou abaixo da faixa de sensibilidade moderada.

Apêndice 6 - Análise comparativa da Sensibilidade e Especificidade através da seleção pelo AMI

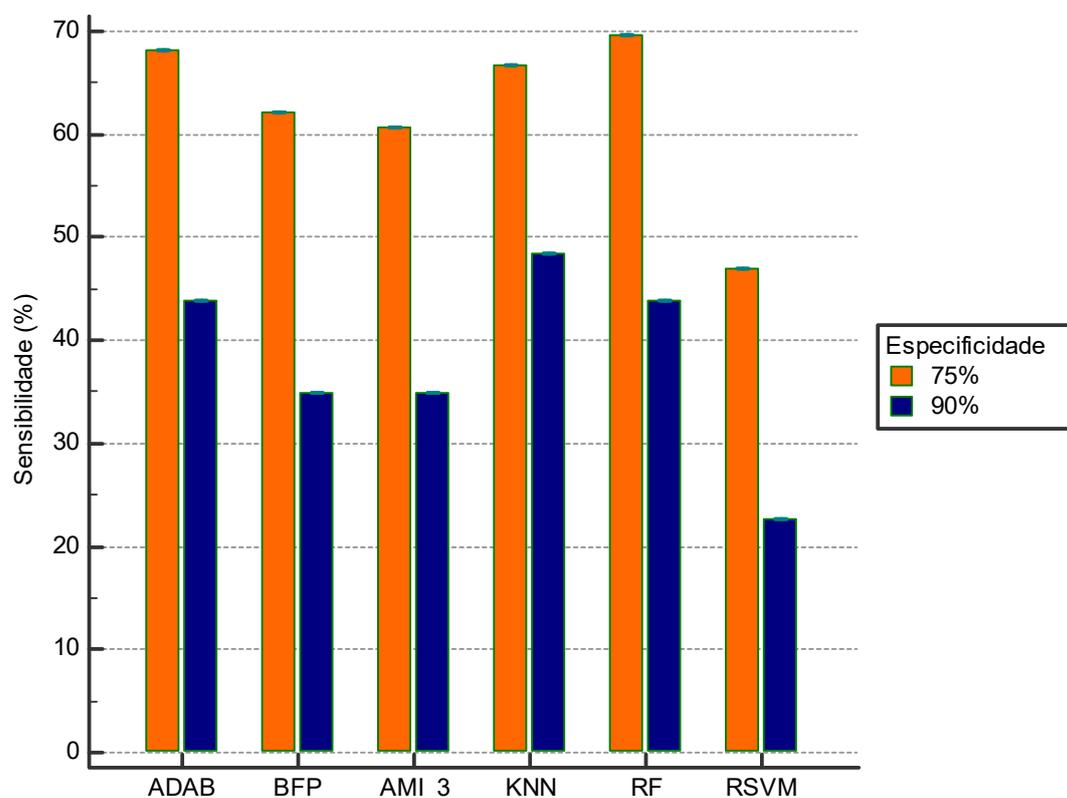


Figura 50 - Análise comparativa da sensibilidade obtida nos seis experimentos com a especificidade de 75% e 90% para os atributos selecionados pelo AMI - Análise Controle *versus* Normal Exame

A Figura 50 mostra que com o valor da especificidade fixado em 75% os algoritmos ADAB, KNN e RF tiveram um aumento na sensibilidade, quando comparados com o BFP. Todos ficaram abaixo da faixa de sensibilidade moderada. Isso indica que o algoritmo permite mais falsos positivos.

Na sequência, quando a especificidade foi fixada em 90%, os algoritmos citados tiveram um comportamento semelhante ao descrito anteriormente, também com a sensibilidade abaixo da faixa moderada.

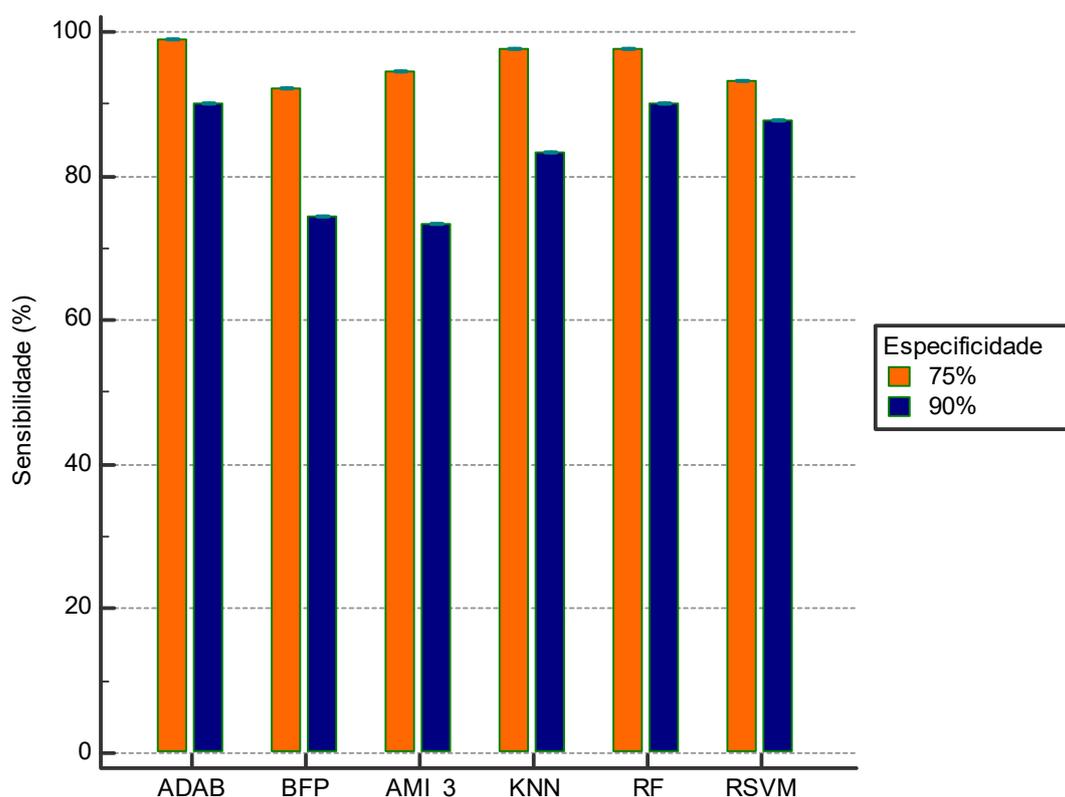


Figura 51 - Análise comparativa da sensibilidade obtida nos seis experimentos com a especificidade de 75% e 90% para os atributos selecionados pelo AMI - Análise Controle *versus* Restritivo

Na Figura 51 é possível observar que com o valor da especificidade fixado em 75% todos os algoritmos tiveram um aumento na sensibilidade, quando comparados com o BFP. Todos eles ficaram na faixa de alta sensibilidade.

Na sequência, quando a especificidade foi fixada em 90%, os algoritmos ADAB, KNN, RF e RSVM também conseguiram obter uma sensibilidade maior que a conseguida pelo BFP. O algoritmo de AMI obteve comportamento semelhante ao BFP.

Apêndice 7 - Análise comparativa da Sensibilidade e Especificidade através da seleção Direta

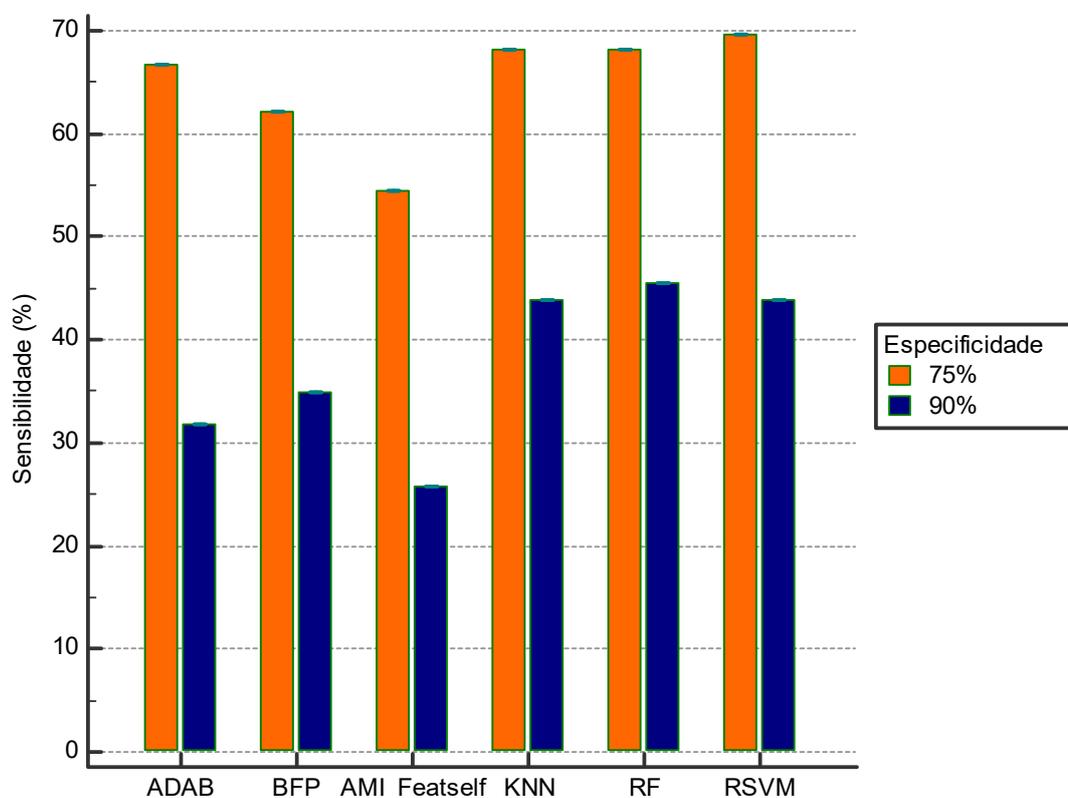


Figura 52 - Análise comparativa da sensibilidade obtida nos seis experimentos com a especificidade de 75% e 90% para os três atributos escolhidos através da seleção direta - Análise Controle *versus* Normal Exame

A Figura 52 mostra que com o valor da especificidade fixado em 75% os algoritmos ADAB, KNN, RF e RSVM tiveram um aumento na sensibilidade, quando comparados com o BFP. Todos eles ficaram abaixo da faixa de sensibilidade moderada.

Na sequência, quando a especificidade foi fixada em 90%, os algoritmos KNN, RF e RSVM tiveram um aumento na sensibilidade, porém continuaram abaixo da faixa de sensibilidade moderada.

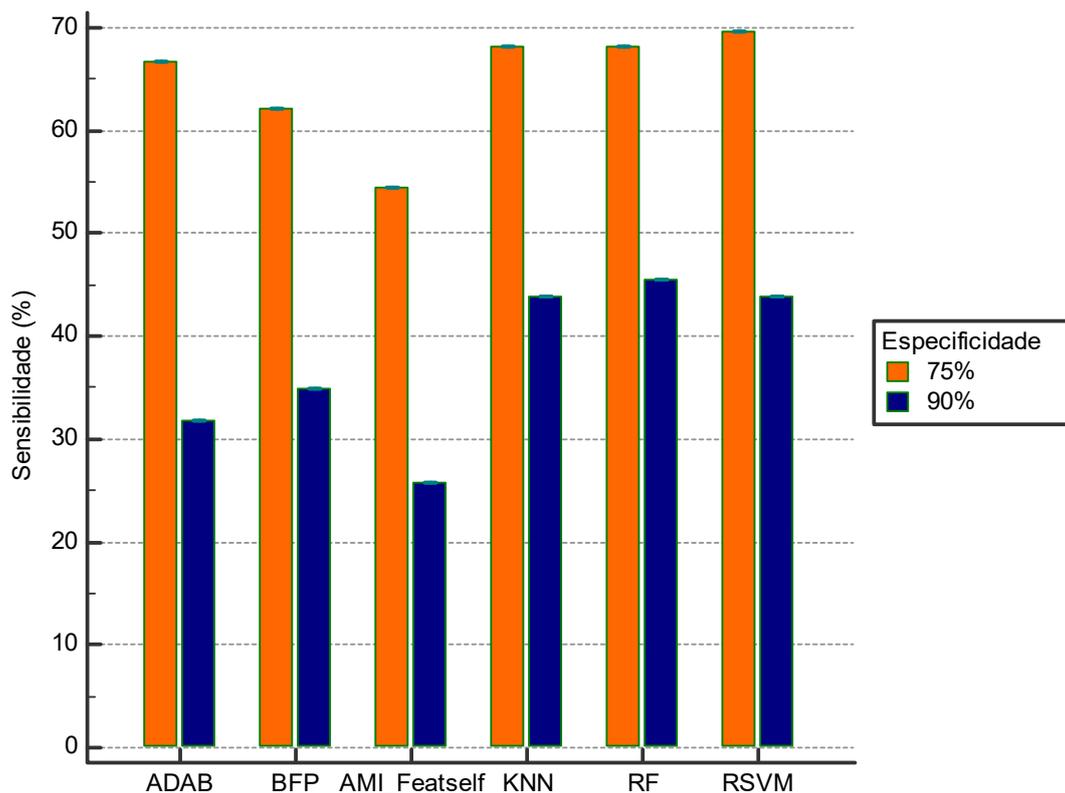


Figura 53 - Análise comparativa da sensibilidade obtida nos seis experimentos com a especificidade de 75% e 90% para os três atributos escolhidos através da seleção direta - Análise Controle *versus* Restritivo

A Figura 53 mostra que com o valor da especificidade fixado em 75% os algoritmos ADAB, KNN, RF e RSVM tiveram um aumento na sensibilidade, quando comparados com o BFP. Todos eles ficaram abaixo da faixa de sensibilidade moderada.

Na sequência, quando a especificidade foi fixada em 90%, os algoritmos KNN, RF e RSVM tiveram um aumento na sensibilidade, quando comparados com o BFP, porém continuaram abaixo da faixa de sensibilidade moderada.