Dissertação

Mestrado em Gestão de Sistemas de Informação Médica

Sensibilidade e Especificidade na Curva ROC
Um Caso de Estudo

Mariana Vitória de Menezes Bordalo Cristiano

Leiria, julho de 2017

Dissertação

Mestrado em Gestão de Sistemas de Informação Médica

Sensibilidade e Especificidade na Curva ROC
Um Caso de Estudo

Mariana Vitória de Menezes Bordalo Cristiano

Dissertação de Mestrado realizada sob a orientação da Doutora Liliana Ferreira, Professora da Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria e do Doutor Rui Santos, Professor da Escola Superior de Tecnologia e Gestão do Instituto Politécnico de Leiria.

Leiria, julho de 2017

Agradecimentos

A entrega desta dissertação é o culminar do que foi o Mestrado em Gestão de Sistemas de Informação Médica (MGSIM). Sendo este o último documento que entrego enquanto estudante do MGSIM quero deixar alguns agradecimentos não só a quem contribuiu para a execução do mesmo, mas também a quem contribui para o meu sucesso, pessoal e profissional, durante esta caminhada.

À minha mãe por todas as vezes que fraquejei e tentei desistir e ela, com o seu jeito e palavras únicas, me fazia sentir que precisava de mais dois ou três mestrados iguais a este para te tornar algo verdadeiramente difícil para mim. Por todas as chamadas que lhe fiz e nunca me deixou sem um conforto, sem um mimo, sem uma resposta, mesmo quando eram as que menos queria ouvir. Por, durante o segundo ano, me ter garantido TODOS os dias um jantar bom, completo e bem servido e o almoço do dia a seguir.

Aos meus orientadores, professora Liliana e professor Rui por me terem dado a oportunidade de conhecer dois dos melhores docentes de quem fui aluna. Por todo o profissionalismo e disponibilidade demonstrada de forma incondicional. Pela celeridade das respostas às dúvidas colocadas por mail, por serem elaboradas e completas permitindo-me continuar a avançar tranquilamente no dia a seguir.

Ao Ivo João por me ter ajudado com os seus conhecimentos mais técnicos e profundos de processamento de imagem e pelas excêntricas constantes mensagens de motivação, que depois da irritação inicial, me davam força para mais uma linha de código, mais um gráfico, uma tabela, um parágrafo. Com ele encaixei o ditado “grão a grão enche a galinha o papo”.

À minha família e à Madi por todo o apoio, incentivo, amizade, ajuda oferecida no decorrer desta caminhada e por acreditarem todos, incondicionalmente, mais em mim do que eu própria algum dia acreditei.

Obrigada a todos os professores do MGSIM.

Muito Obrigada!

Resumo

A fiabilidade de qualquer diagnóstico clínico é fundamental para o sucesso de um sistema de saúde. Deste modo, é crucial dispor de medidas que permitam aferir, de forma intuitiva, a qualidade de testes para classificar indivíduos como doentes ou saudáveis relativamente a determinada doença.

Conceitos básicos da epidemiologia, tais como acurácia, sensibilidade, especificidade, valores preditivos positivo e negativo, razões de verosimilhança positiva e negativa permitem aferir o melhor desempenho entre diferentes testes. A curva ROC, ao representar a sensibilidade e a especificidade para todos os possíveis valores para o ponto de corte, é uma das ferramentas mais utilizadas para avaliar e comparar diferentes tipos de metodologias de diagnóstico. Além disso, a área sob a curva ROC é uma medida objetiva do desempenho do teste associado.

Nesta dissertação aplica-se a curva ROC a um caso de estudo cujos dados representam os resultados de um teste de diagnóstico ao cancro da pele que, com recurso a técnicas de imagiologia, procura detetar a presença de padrão reticular de forma a diagnosticar melanomas. A base de dados utilizada contém informação sobre 158 imagens dermatoscópicas. Através de técnicas de processamento de imagem, cada imagem foi filtrada em 9 escalas diferentes e cada escala dividida em 26 medidas. Para cada um dos 234 resultados, foram aplicadas três avaliações distintas: desvio padrão, energia e entropia.

Como aplicação ao caso de estudo, foram analisadas duas metodologias distintas. A primeira procura a medida que gera a "melhor" curva, considerando a área abaixo da mesma. A segunda consiste na procura do "melhor" ponto de corte de uma dada curva, através da maximização da soma da sensibilidade e especificidade ou da minimização da distância ao ponto ideal (ausência de erros de classificação). Em ambas as metodologias utilizaram-se medidas agrupadas, através da média aritmética ou da média ponderada, recorrendo ao algoritmo Adaboost, ou simplesmente os dados originais.

Os resultados obtidos evidenciam que o agrupamento de medidas permite alcançar resultados mais fiáveis. Todavia, devem ser utilizadas técnicas que permitam uma análise prévia das medidas a utilizar, bem como das ponderações adequadas para cada uma dessas medidas, uma vez que a utilização da média aritmética (ponderação igual em todas as medidas), na maioria das situações, demonstrou mau desempenho.

Palavras-chave: curva ROC, diagnóstico, sensibilidade, especificidade, padrão reticular.

Abstract

The reliability of any clinical diagnostic is fundamental to the success of a health system. Thus, it's crucial to have available measures that can intuitively assess the quality of diagnostic tests that classify the individuals as ill or healthy regarding a particular disease.

Basic concepts of epidemiology such as accuracy, sensitivity, specificity, positive and negative predictive values, positive and negative likelihood ratios allow the assessment of the best performance between different tests. The ROC curve, representing sensitivity and specificity for all possible values for the cut-off point, is one of the most used procedures to evaluate and compare different types of diagnostic methodologies. In addition, the area under the ROC curve is an objective measure of the performance of the associated test.

In this dissertation the ROC curve is applied to a case study whose data represent the results of a diagnostic test for skin cancer that, using imaging techniques, aims to detect the presence of a reticular pattern in order to diagnose melanomas. The used database contains information on 158 dermatoscopic images. Through image processing techniques, each image was filtered on 9 different scales and each scale divided into 26 measurements. For each of the 234 results, three different evaluations were applied: standard deviation, energy and entropy.

As an application to the case study, two different methodologies were analyzed. The first one looks for the measure that generates the “best” curve, considering the area below it. The second consists in the search for the “best” cut-off point of a given curve, by maximizing the sum of the sensitivity and specificity or minimizing the distance to the ideal point (absence of classification errors). In both methodologies we used grouped measures, using the arithmetic mean or the weighted average, using the Adaboost algorithm, or simply the original data.

The obtained results show that the grouping of measures allows to obtain more reliable classification. However, techniques should be used to allow for a prior analysis of the measures to be used and the appropriate weights for each of these measures, since the use of the arithmetic mean (equal weighting in all measures) in most situations has shown poor performance.

Key-Words: ROC curve, diagnosis, sensitivity, specificity, reticular pattern.

Lista de figuras

Lista de tabelas

Lista de siglas

– – Teste negativo

+ – Teste positivo

D – Doente

e – especificidade

FN – Falso Negativo

FP – Falso Positivo

i – Taxa de incidência

p – taxa de prevalência

ROC – *Receiver Operating Characteristic;*

RVN – Razão de Verosimilhança Negativa

RVP – Razão de Verosimilhança Positiva

S – Saudável

s – sensibilidade

VN – Verdadeiro Negativo

VP – Verdadeiro Positivo

VPN – Valor Preditivo Negativo

VPP – Valor Preditivo Positivo

Índice

1. Introdução

A identificação médica da causa do sofrimento ou perturbação de que uma pessoa se queixa implica geralmente a identificação tanto do processo patológico como do agente responsável. Este processo denomina-se diagnóstico. O diagnóstico implica ciência, técnica e um pouco de intuição médica. Após um exame físico e a formulação de um diagnóstico clínico provisório, são, muitas vezes, necessárias análises de sangue, técnicas imagiológicas e outros meios complementares de diagnóstico. Assim, torna-se necessário avaliar a capacidade destes testes em classificar corretamente os indivíduos em dois subgrupos clinicamente distintos, doente ou saudável [1].

Ao longo desta dissertação faz-se a análise da curva ROC enquanto ferramenta para avaliar o desempenho de testes de diagnóstico na medicina. Por norma, quando se faz, fala, ou ouve falar de um teste de diagnóstico, instintivamente, o que se pretende saber é se o resultado é positivo ou negativo para se poder aferir se o indivíduo está doente ou saudável. Contudo, na prática não é assim tão linear.

Qualquer que seja, ou venha a ser, o teste de diagnóstico usado no despiste de uma doença, haverá sempre a possibilidade de se obter resultados falsos. Assim, um indivíduo que esteja efetivamente doente e tenha obtido resultado positivo, então, trata-se de um verdadeiro positivo (VP). Tal como se o indivíduo estiver efetivamente saudável e tiver obtido resultado negativo, então, trata-se de um verdadeiro negativo (VN). Contudo, se um indivíduo obtiver um resultado positivo, mas na realidade estiver saudável, recebeu um falso positivo (FP). Tal como se estiver doente e tiver recebido um resultado negativo, obteve um falso negativo (FN). Por isso, pode dizer-se que para qualquer teste diagnóstico podem obter-se quatro resultados, dois falsos, positivo e negativo, e dois verdadeiros, positivo e negativo.

Assim, quando se considera o resultado de um teste de diagnóstico entre duas populações, uma doente e outra saudável, dificilmente se consegue uma separação perfeita entre elas. Deste modo, considerando que a quantidade utilizada para diagnosticar uma doença é representada por uma curva de frequências em cada uma de duas subpopulações, doentes e saudáveis, há normalmente sobreposição entre estas curvas. Nessa sobreposição, encontram-se os falsos positivos e os falsos negativos, valores com os quais não é possível saber com certeza absoluta se o indivíduo está doente ou saudável. Consequentemente, qualquer que

seja o ponto escolhido que separa as duas populações – ponto de corte –, haverá sempre a possibilidade de existirem indivíduos incorretamente classificados, por menor que seja essa percentagem [2].

Neste contexto, os principais objetivos desta dissertação são apresentar pormenorizadamente os conceitos associados à curva ROC, nomeadamente a sensibilidade e a especificidade, expor o que é o ponto de corte e mostrar a relevância da área abaixo da curva ROC. Posteriormente, pretende-se aplicar diferentes metodologias e estudar qual ou quais sobressaem no que toca ao desempenho do teste de diagnóstico do cancro da pele através da deteção do padrão reticular [3].

Para atingir estes objetivos, a dissertação encontra-se dividida em cinco capítulos, além da bibliografia e anexos. Após a introdução, no segundo capítulo, Curva ROC, encontra-se a introdução à curva ROC, faz-se um enquadramento histórico explicando a origem e aplicações iniciais desta curva, exploram-se detalhadamente alguns conceitos de epidemiologia associados, tais como prevalência e incidência, acurácia, sensibilidade e especificidade, valores preditivos positivo e negativo, *odds* ou razões de verosimilhança e o ponto de corte. Ainda neste capítulo, explana-se a análise pormenorizada da curva ROC onde se inclui a apresentação da área abaixo da curva.

O terceiro capítulo desenvolve todo o caso de estudo trabalhado nesta dissertação, a deteção do cancro da pele através da deteção do padrão reticular. Começa-se por introduzir o caso de estudo, de seguida, define-se a técnica da dermatoscopia, mais tarde, desenvolve-se o processamento de imagem e, por fim, explica-se minuciosamente a base de dados utilizada. Este caso de estudo baseia-se numa base de dados de imagens dermatoscópicas cedidas pelos autores do artigo [4], que aplicaram técnicas de processamento de imagem filtrando cada uma das imagens em 9 escalas diferentes, sendo cada escala divida em 26 medidas e considerando para cada uma três avaliações diferentes (desvio padrão, energia e entropia). Com as medidas disponíveis foi feita uma análise no sentido de perceber quais as melhores opções para a deteção de padrão reticular, se a utilização de medidas isoladas ou agrupadas.

O quarto capítulo aplica os conceitos explorados no capítulo dois ao caso de estudo, contemplando a procura da "melhor" curva ROC e do "melhor" ponto de corte. São consideradas diversas técnicas/metodologias, tais como, a utilização de cada medida isolada

ou o agrupamento de diversas medidas, gerado pela média aritmética ou pela média ponderada através do algoritmo Adaboost.

No quinto e último capítulo são apresentadas as principais conclusões deste estudo.

2. Curva ROC

Neste capítulo, far-se-á a apresentação discriminativa da curva ROC através de um enquadramento funcional, fazendo a perspetiva histórica, apresentando e detalhando os conceitos básicos de prevalência e incidência, acurácia, sensibilidade e especificidade, valores preditivos negativo e positivo, *odds* ou razões de verosimilhança (chances) e ponto de corte. Expõe-se ainda a análise ROC através da representação da curva ROC, dos níveis de discriminação e critérios de decisão da área abaixo da curva e a problemática da comparação entre curvas.

2.1. Introdução

São variadíssimas as situações em que determinados conjuntos de objetos, situações ou ações podem ser classificados como pertencentes a uma de duas classes. Os procedimentos de classificação devem basear-se em informações observadas sobre cada um(a) deles(as). Contudo, estes procedimentos não são perfeitos, são cometidos, como se verá mais à frente, alguns erros, o que inviabiliza uma correta classificação da classe. Por este motivo, deve avaliar-se a qualidade da realização dos procedimentos. Assim, é possível decidir se um teste de diagnóstico é bom o suficiente, tentar melhorá-lo ou, simplesmente, substituí-lo [5].

Exemplos de problemas reais que se encaixam na dimensão de objetos [5]:

- Realização do diagnóstico médico, em que o objetivo é classificar o paciente em "doente" ou "saudável";
- Desenvolvimento de sistemas de reconhecimento de voz, em que o objetivo é classificar as palavras faladas em "reconhecida" ou "não reconhecida";
- Avaliação das aplicações financeiras de crédito, em que o objetivo é determinar se o candidato tem "padrão provável" ou "padrão não provável";
- Filtrar *emails* recebidos, em que o objetivo é determinar se são *spam* ou mensagens genuínas;
- Avaliar os candidatos a determinado curso, em que o objetivo é determinar se fica "colocado" ou "não colocado";
- Examinar as transações do cartão de crédito, em que o objetivo é decidir se são "fraude" ou "não fraude";

- Avaliar a expressão genética de dados *microarray*, em que o objetivo é determinar se correspondem a cancro, ou não.

De facto, a lista de possíveis aplicações de procedimentos de classificação é longa. Em certos casos, existem mais de duas classes de classificação. Normalmente, nestes casos agrupam-se as classes até que fiquem apenas duas. Por exemplo, no caso de quatro classes A, B, C e D, poder-se-ia manter uma delas e as outras três representariam a segunda classe. Há, ainda, métodos mais complexos que pretendem analisar várias categorias simultaneamente, mas ainda sem grande efeito prático. Veja-se, por exemplo, as referências de [6] relativamente à mamografia através da qual se pretende classificar em "Sem quisto", "Presença de quisto benigno" e "Presença de quisto maligno". Mas, na prática, as situações divididas em duas classes são as mais frequentes e estudadas, tais como doente/saudável, sim/não, certo/errado, aceite/rejeitado, condição presente/condição ausente, entre outras.

Existem várias formas de avaliar a qualidade do desempenho de sistemas de classificação que visam responder a que classe pertence cada indivíduo em estudo. Nesta dissertação, recorrer-se-á a uma abordagem extremamente importante e amplamente utilizada, a curva *Receiver Operating Characteristic* (ROC).

2.2. Perspetiva histórica

A curva ROC foi originalmente criada na área da psicologia sensorial. Tinha como objetivo comprovar a existência de uma relação empírica entre o corpo e a mente, segundo Gustav Theodor Fechner (1801-1887), filósofo alemão e médico de formação, considerado o pioneiro em psicometria [7]. Fechner sujeitava as pessoas a determinado estímulo até obter um valor consideravelmente estável de respostas positivas. Reproduziu graficamente a relação entre as respostas positivas e a medida física da intensidade do estímulo adquirindo, assim, uma função psicométrica [8].

Louis Leon Thurstone (1887-1955), pioneiro em psicometria e psicofísica, com base nos resultados de Fechner, desenvolveu novas técnicas para quantificar as qualidades mentais. Publicou diversas escalas de atitude onde procurou medir a influência de preconceitos de propaganda do homem. Tinha especial interesse pela medição da aprendizagem e tentou

exprimir, através de unidades absolutas, o desenvolvimento da aprendizagem [9]. Thurstone é o criador dos conceitos ruído, critério de decisão e ponto de corte [8].

Durante a Segunda Guerra Mundial (1939-1945), a curva ROC foi usada para quantificar a capacidade dos operadores de radares distinguirem um sinal de ruído [6]. Ou seja, quando um radar detetava algum movimento cabia ao operador determinar a veracidade e relevância do que havia sido detetado (se um míssil, avião inimigo ou, simplesmente, um bando de pássaros) [10].

Na década de 60, as curvas ROC foram usadas essencialmente em psicologia experimental e na de 70 expandiram-se pelos campos da investigação biomédica, campo cujo objetivo se tornou, basicamente, em auxiliar a classificação de indivíduos em doentes ou saudáveis [10].

Em suma, o nome ROC surge, em teoria, para a deteção de um sinal, cujo objetivo é detetar a presença, ou não, de um sinal particular.

2.3. Conceitos básicos

De modo a que mais intuitivamente se possa compreender e acompanhar a leitura desta dissertação, assumir-se-á a nomenclatura de doente *versus* saudável em vez de doente *versus* não doente. É importante a ressalva, visto que uma pessoa que não padeça da doença em estudo não tem de estar necessariamente saudável.

De seguida, apresentar-se-ão os resultados de um teste de diagnóstico, meramente ilustrativos, que visam servir de meio de introdução aos conceitos fundamentais aplicados ao longo desta dissertação.

Considere-se, então, que para um determinado teste, efetuado sob uma população total de 22 indivíduos, 8 estão doentes e 14 estão saudáveis. Dos doentes, 6 foram corretamente classificados como doentes e 2 foram incorretamente classificados como saudáveis. Dos saudáveis, 4 foram incorretamente classificados como doentes e 10 foram corretamente classificados como saudáveis. Este cenário encontra-se esquematizado na tabela de contingência representada na Tabela 1.

	Teste +	Teste -	Total
Doente (D)	6	2	8
Saudável (S)	4	10	14
Total	10	12	22

Tabela 1 – Tabela de contingência de um teste de diagnóstico ilustrativo.

Os conceitos básicos estão diretamente relacionadas com os quatro possíveis resultados obtidos num teste de diagnóstico, tal como apresentado na Tabela 2.

	Teste +	Teste -
Doente (D)	**V**erdadeiros **P**ositivos	**F**alsos **N**egativos
Saudável (S)	**F**alsos **P**ositivos	**V**erdadeiros **N**egativos

Tabela 2 – Possíveis resultados de um teste de diagnóstico.

2.3.1 Prevalência e incidência

A prevalência (p) ou taxa de prevalência é a medida que permite aferir se a condição em estudo é frequente ou esporádica numa determinada população, ou seja, é a proporção de indivíduos doentes na população em estudo [11]. Assim, determina-se segundo a fórmula, $p = \frac{VP+FN}{VP+FN+FP+VN}$ = 8/22 = 0.36. Portanto, para este caso em concreto, 36% da população é doente [11].

A incidência (i) ou taxa de incidência permite quantificar ou determinar a proporção de novos casos, na condição em estudo, num determinado período. Deste modo, tem-se que $i = \frac{número\ de\ novos\ casos}{número\ de\ pessoas\ analisadas}$, no período considerado. Supondo que nos últimos seis meses, dos 14 indivíduos saudáveis, surgiram 7 novos doentes, $i = \frac{7}{14}$ = 0.5. Isto significa que por cada 100 indivíduos surgem 50 novos casos de doença [11].

2.3.2 Acurácia

A acurácia (a) é a medida que traduz a precisão de um teste de diagnóstico, ou seja, permite determinar a percentagem de diagnósticos corretamente determinados [12]. Calcula-se

através da fórmula, $a = \frac{VP+VN}{VP+FN+FP+V}$ = 16/22 = 0.73. Portanto, 73% da população obtém o resultado correto [13].

A acurácia não é uma medida muito utilizada visto que varia com a prevalência. Contudo, torna-se uma medida de interesse quando se sabe, à partida, que deixar escapar diagnósticos falsos, ou seja, quer falsos positivos quer falsos negativos, pode trazer sérias repercussões num futuro próximo ou longínquo. Assim, neste caso, pretende-se um valor de acurácia elevado. Caso a ocorrência de diagnósticos falsos não for relevante, pode aceitar-se um valor de acurácia mais baixo [13].

2.3.3 Sensibilidade e especificidade

Dois conceitos basilares nesta dissertação são a sensibilidade (*s*) e a especificidade (*e*). A sensibilidade é a capacidade do teste em identificar um indivíduo doente, ou seja, corresponde à probabilidade do teste classificar corretamente um indivíduo doente. A especificidade é a capacidade do teste em identificar um indivíduo saudável, ou seja, corresponde à probabilidade do teste classificar corretamente um indivíduo saudável [10] [14].

Dada a Tabela 2 e explicações anteriores, a sensibilidade corresponde à probabilidade de ocorrer um resultado positivo sabendo que o indivíduo é doente, ou seja, à fração de positivos entre os doentes, portanto $s = \frac{VP}{VP+FN}$ = P(+|D). A especificidade corresponde à probabilidade de ocorrer um resultado negativo sabendo que o indivíduo é saudável, ou seja, à fração de negativos entre os saudáveis, logo $e = \frac{VN}{VN+FP}$ = P(−|S).

Para este teste, já se verificou que, dos 22 indivíduos, 8 estão doentes e 14 estão saudáveis. Contudo, é importante frisar que, dos doentes, 6 obtiveram resultado positivo (VP) e 2 resultado negativo (FN) e, dos saudáveis, 4 obtiveram resultado positivo (FP) e 10 resultado negativo (VN).

Logo, *s* = P(+|D) = 6 / (6+2) = 0.75. Portanto, 75% dos indivíduos doentes obtém resultado verdadeiro. Assim como, *e* = P(−|S) = 10/(10+4) = 0.71. Portanto, 71% dos indivíduos saudáveis obtém resultado verdadeiro.

Como é possível constatar, a sensibilidade e a especificidade são medidas independentes entre si, uma vez que não são calculadas sobre os mesmos indivíduos, uma foca-se apenas nos indivíduos doentes e a outra nos indivíduos saudáveis, e não sofrem efeitos de prevalência *p* da doença, ou seja, a proporção de doentes em estudo não tem influência no cálculo destas medidas. Dependendo da dinâmica e envolvência do estudo, pode atribuir-se maior importância à sensibilidade ou à especificidade [15].

Um teste é sensível quando dificilmente deixa escapar um indivíduo doente. Deste modo, deve dar-se ênfase à sensibilidade quando:

- É grave não diagnosticar a patologia;
- A patologia em causa tem cura;
- Tratar falsos negativos não cause nenhuma sequela física, psicológica ou social a curto, médio ou longo prazo ao indivíduo;
- O tratamento supera, de alguma forma, o não tratamento.

Um teste é específico quando dificilmente caracteriza um indivíduo saudável sem o estar. Assim sendo, deve prevalecer a especificidade quando:

- A patologia é difícil de curar e/ou o tratamento traz qualquer efeito secundário agravado para o indivíduo;
- É importante ter a certeza que o indivíduo está de facto doente;
- Há possibilidade de um tratamento aplicado a falsos positivos causar sequelas físicas, psicológicas ou sociais a curto, médio ou longo prazo ao indivíduo;
- O não tratamento supera, de alguma forma, o tratamento.

2.3.4 Valores preditivos positivo e negativo

Outros conceitos básicos, mas não menos importantes, são o valor preditivo positivo (VPP) e o valor preditivo negativo (VPN). O VPP é a capacidade do teste em identificar verdadeiros positivos de entre todos os indivíduos com resultado positivo. O VPN é a capacidade do teste em identificar verdadeiros negativos de entre todos os indivíduos com resultado negativo [16].

Relembrando a Tabela 2, o VPP corresponde à probabilidade de um indivíduo estar doente sabendo que o resultado do seu teste é positivo, logo, à fração de doentes entre os positivos,

ou seja, VPP = $\frac{VP}{VP+F}$ = P(D |+). O VPN corresponde à probabilidade de um indivíduo estar saudável sabendo que o resultado do teste é negativo, logo, à fração de saudáveis entre os negativos, isto é, VPN = $\frac{VN}{VN+FN}$ = P(S |−) [16].

Como se pode verificar, o VPP e o VPN são valores independentes entre si, pois não são calculados sobre os mesmos indivíduos. Um foca-se apenas nos indivíduos com resultados positivos e o outro nos indivíduos apenas com resultados negativos.

Logo, VPP = P(D |+) = 6/(6+4) = 0.60. Portanto, 60% dos indivíduos com resultado positivo estão realmente doentes. Assim como, VPN = P(S |−) = 10 / (10+2) = 0.83. Portanto, 83% dos indivíduos com resultado negativo estão realmente saudáveis.

Os valores preditivos positivos e negativos são medidas pouco usadas pelo facto do seu valor variar com a taxa de prevalência. Veja-se, se a prevalência for baixa, significa que o número de indivíduos saudáveis é superior ao número de indivíduos doentes, logo, o número de FP e VN tenderão a ser mais elevados do que o número de VP e FN, consequentemente VPP tenderá a ser baixo e VPN alto. Já se a prevalência for alta, significa que o número de indivíduos doentes é superior ao número de indivíduos saudáveis, logo, o número de VP e FN tenderá a ser superior a FP e VN, consequentemente VPP tenderá a ser elevado e VPN baixo. É por esta razão que são medidas pouco aplicadas. Todavia, são importantes medidas quando o resultado do teste de diagnóstico já for conhecido, ou seja, imaginando que se quer responder à questão "se realizar um determinado teste e der positivo qual a probabilidade do resultado estar certo?", sabendo os valores de VP, FP, FN, VN, como é ilustrado no exemplo apresentado no início do capítulo 2.3 [16].

Para terminar este subcapítulo é importante acentuar a importância de saber qual o universo que está a ser usado, pois só assim é que os conceitos, como falsos positivos ou falsos negativos, podem ser bem interpretados.

Deste modo, quando se pretende determinar a percentagem de falsos positivos, pode considerar-se:

- Toda a população e ter-se-á 4 FP em 22 indivíduos, logo 18% da população obteve FP;

- Apenas os saudáveis e ter-se-á 4 FP em 14 indivíduos, logo 29% dos indivíduos saudáveis obteve FP o que corresponde ao cálculo de 1-especificidade;
- Apenas resultados positivos e ter-se-á 4 FP em 10 indivíduos, logo 40% dos indivíduos com resultado positivo estão, na verdade, saudáveis o que corresponde ao cálculo de 1-VPP.

De igual modo, quando se pretende determinar a percentagem de falsos negativos, pode considerar-se:

- Toda a população e ter-se-á 2 FN em 22 indivíduos, logo 9% da população obteve FN;
- Apenas os doentes e ter-se-á 2 FN em 8 indivíduos, logo 25% dos indivíduos doentes obteve FN o que corresponde ao cálculo de 1-sensibilidade;
- Apenas resultados negativos e ter-se-á 2 FN em 12 indivíduos, logo 16.7% dos indivíduos com resultado negativo estão, na verdade, doentes o que corresponde ao cálculo de 1-VPN.

2.3.5 *Odds* ou razões de verosimilhança (chances)

Para um determinado teste é frequentemente interessante analisar as chances, *odds*, quando o resultado do teste é positivo ou negativo [17].

O objetivo desta medida é indicar quantas vezes um acontecimento é mais provável que outro [18]. Assim, para determinar quantas vezes é mais provável estar doente que estar saudável, tem-se que $odds = \frac{p}{1-p}$. Como foi visto no subcapítulo 2.3.1, p = 0.36, pelo que *odds* = 0.36/0.64 = 0.56. Pode concluir-se que estar doente é 0.56 vezes mais provável que estar saudável, logo é 44 por cento menos provável estar doente que saudável.

Um conceito semelhante das chances ou *odds* são as chamadas razões de verosimilhança. As razões de verosimilhança correspondem ao quociente entre um determinado resultado para um indivíduo doente e o mesmo resultado para um indivíduo saudável [17].

A razão de verosimilhança positiva (RVP) representa quantas vezes um diagnóstico verdadeiro positivo, entre os indivíduos doentes, é mais provável que um falso positivo, entre

os indivíduos saudáveis. Assim, $RVP = \frac{s}{1-e}$ = 0.75/0.29 = 2.6. Logo, para este teste, um verdadeiro positivo é 2.6 vezes mais provável que um falso positivo.

A razão de verosimilhança negativa (RVN) representa quantas vezes um diagnóstico falso negativo, entre os indivíduos doentes, é mais provável que um verdadeiro negativo, entre os indivíduos saudáveis. Assim, $RVN = \frac{1-s}{e}$ = 0.25/0.71 = 0.35. Logo, para este teste, um falso negativo é 65 por cento menos provável que um verdadeiro negativo.

As razões de verosimilhança são medidas de tanto interesse como a sensibilidade e a especificidade, aliás, são calculadas através destas duas últimas e dos seus complementares. O teste tende a ser melhor quanto maior for a RVP e menor for a RVN, pois quanto maior for RVP maior será a probabilidade de um resultado positivo ser VP do que FP. De forma semelhante, quanto menor for RVN maior será a probabilidade de um resultado negativo ser VN do que FN.

2.3.6 Ponto de corte

Uma grande percentagem dos testes de diagnóstico origina respostas sob a forma de variáveis qualitativas ordinais ou quantitativas discretas ou contínuas, por isso, é necessário aplicar "*uma regra de decisão baseada em encontrar o ponto de corte que resume tal quantidade numa resposta dicotómica*", valor que separa indivíduos doentes de saudáveis no teste de classificação [10].

O ponto de corte corresponde a um ponto de separação na identificação dos indivíduos como doentes ou saudáveis, considerando uma medida utilizada para fazer o diagnóstico. O ponto de separação representa um valor definido para essa medida, estabelecendo assim quais os indivíduos que estão acima ou abaixo desse ponto [19].

Ao longo da presente dissertação, de forma a simplificar a sua exposição, vai supor-se que um indivíduo é classificado como doente se na sua análise obtiver um valor superior ao valor do ponto de corte e saudável caso contrário, isto é, se na sua análise resultar um valor inferior ou igual ao valor do ponto de corte. Todavia, refira-se que em muitos diagnósticos é aplicada a desigualdade oposta, isto é, valores baixos correspondem a indivíduos doentes e valores

altos a indivíduos saudáveis. Contudo, o raciocínio a aplicar será idêntico ao desenvolvido ao longo desta investigação.

Os gráficos que se seguem não explanam uma situação real, são unicamente elucidativos.

A situação ideal é apresentada na Figura 1, em que a distinção entre saudáveis e doentes é perfeita, ou seja, neste exemplo todos os indivíduos são corretamente classificados. Neste caso, o valor do ponto de corte é, por exemplo, 10, o que determina que abaixo de 10 todos os indivíduos são saudáveis e acima de 10 todos são doentes. O código usado para gerar esta ação encontra-se no Anexo A.1.

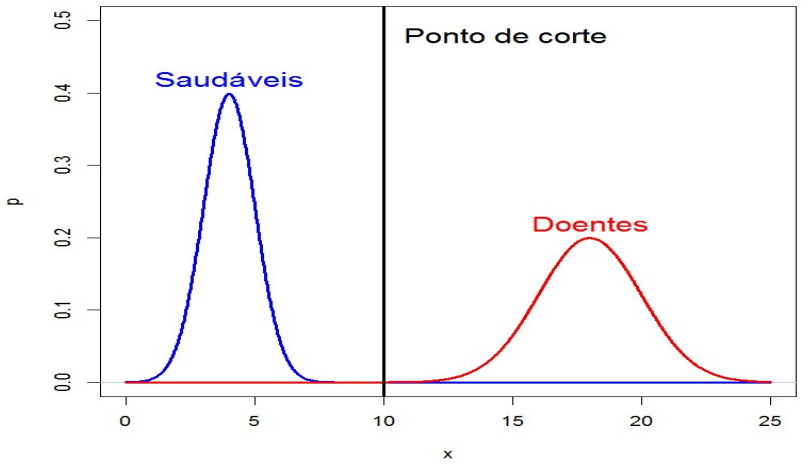


Figura 1 – Representação do ponto de corte ideal.

Todavia, é extremamente difícil encontrar um teste de diagnóstico para o qual não se obtenham falsos resultados positivos ou negativos. É nestes casos que se analisa diferentes valores para o ponto de corte de forma a encontrar o valor que minimize a ocorrência de resultados falsos [20].

Considerando o ponto de corte definido no exemplo apresentado na Figura 2, a distinção já não é perfeita, uma vez que alguns indivíduos doentes são classificados como saudáveis (os que se encontram à esquerda da reta do ponto de corte, na região a vermelho), enquanto outros indivíduos saudáveis são identificados como doentes (os que estão à direita da reta do ponto de corte, na região a azul). O código usado para gerar esta ação encontra-se no Anexo A.2.

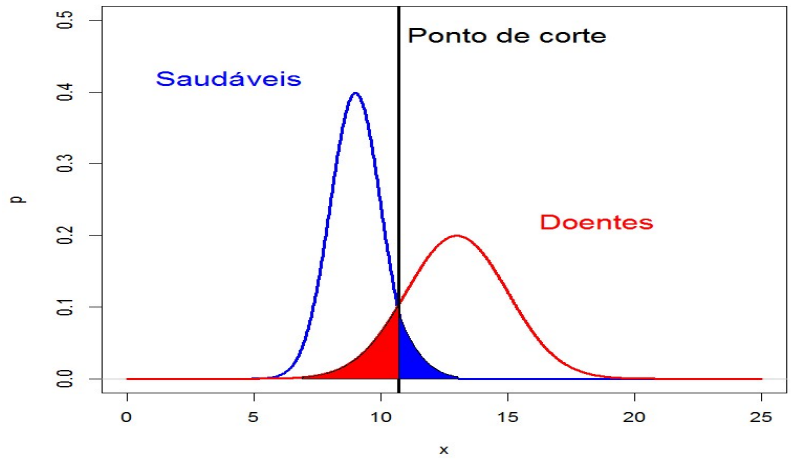


Figura 2 – Representação semelhante do ponto de corte na realidade.

De um modo geral, quanto menor o ponto de corte, maior será a sensibilidade *s*, ou seja, maior a capacidade do teste em classificar corretamente indivíduos doentes e menor será a especificidade *e*, ou seja, menor a capacidade em classificar corretamente indivíduos saudáveis. Esta ideia é ilustrada através na Figura 3, que corresponde ao mesmo teste de diagnóstico representado na Figura 2, mas diminuindo o valor do ponto de corte. Deste modo, a área da região a vermelho (FN) diminuiu e a área da região a azul (FP) aumentou. O código usado para gerar esta ação encontra-se no Anexo A3.

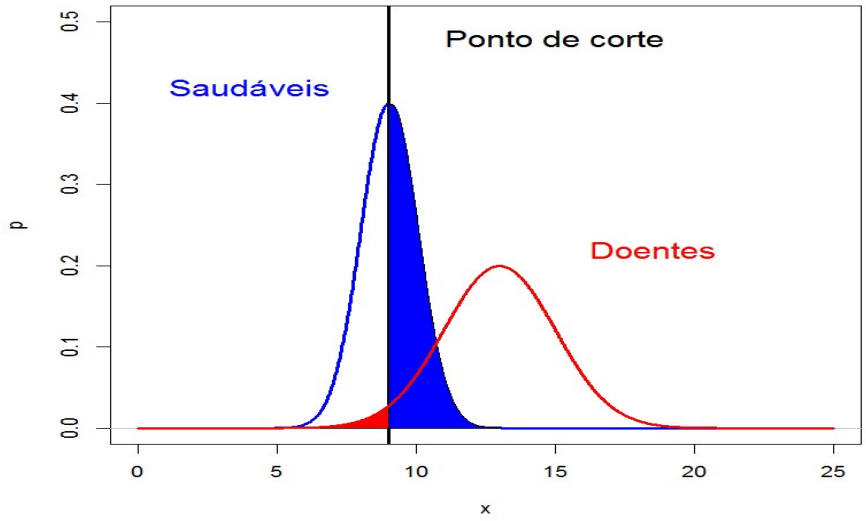


Figura 3 – Aumento da sensibilidade

Por outro lado, quanto maior for o ponto de corte, menor será a sensibilidade *s*, mas maior será a especificidade *e*. Deste modo, melhorar uma medida implica piorar a outra.

Na Figura 4, em comparação com as Figuras 2 e 3, pode verificar-se que quando o valor do ponto de corte aumenta, a área da região a vermelho (FN) aumenta, logo a sensibilidade diminui, e a área da região a azul (FP) diminui, logo a especificidade aumenta.

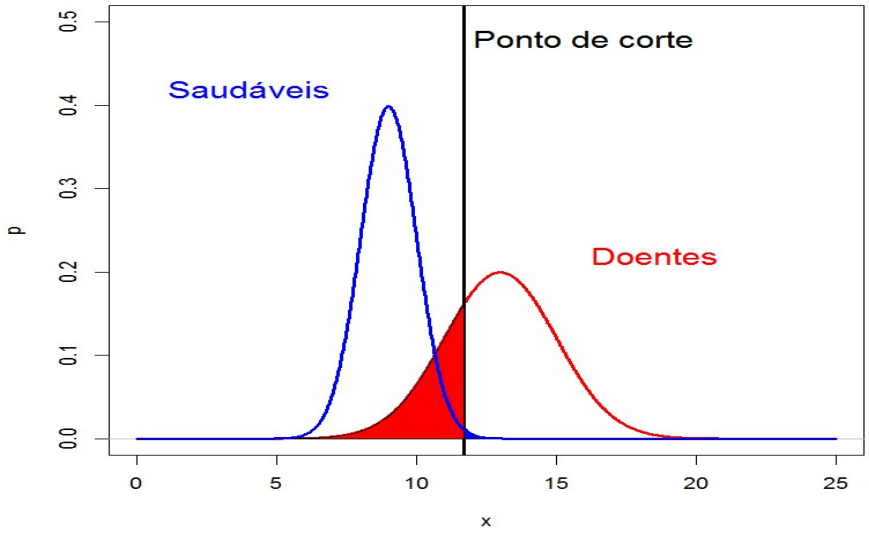


Figura 4 – Aumento da espeficidade.

Assim, é muitas vezes inevitável que alguns indivíduos sejam incorretamente classificados como doentes (FP) ou como saudáveis (FN). Contudo, o valor do ponto de corte pode ser definido de forma a que, de entre os possíveis valores da sensibilidade e da especificidade, se possa optar por aqueles que se consideram os mais adequados à doença em análise. O código usado para gerar esta ação encontra-se no Anexo A.4.

Uma escolha, que nem sempre é a mais correta, é a de escolher como ponto de corte a zona em que a soma da sensibilidade com a especificidade é maior, isto porque, quando não se pode correr o risco de não diagnosticar deve privilegiar-se a sensibilidade. Tal como, quando é mais prudente diagnosticar se e só se o indivíduo for indubitavelmente doente, deve privilegiar-se a especificidade.

No decorrer da dissertação, esta será uma situação recorrente, o desafio de saber o que privilegiar, se a sensibilidade se a especificidade, e para a qual dificilmente se chegará ao ponto de corte certo, exato, preciso e infalível, livre das repercussões da complexidade do ser humano. Uma vez que não é possível conciliar o melhor de dois mundos, tenta escolher-se o ponto de corte que expresse o melhor equilíbrio entre valores de sensibilidade e

especificidade, o que traga, cumulativamente, mais benefícios e menos riscos para a população em estudo e o que depois de devidamente estudado, analisado e ponderado pela Organização Mundial de Saúde, e/ou outras entidades nacionais como o Infarmed, tenha aval positivo das mesmas.

Tanto a estatística como a informática pretendem, juntas, gerar o melhor ponto de corte para cada teste de diagnóstico. Desta forma, para diferentes pontos de corte, pretende estimar-se os valores de sensibilidade e especificidade de forma a ser representada a curva ROC, que usa como coordenadas (1-*e*; *s*).

2.4. Análise ROC

A curva ROC é uma técnica gráfica utilizada para avaliar a capacidade de um teste de diagnóstico fazer a distinção entre doentes e não doentes. Permite fazer análises visuais entre sensibilidade e especificidade relativamente a diversos pontos de corte. A curva é obtida através do cálculo da sensibilidade e da especificidade para cada ponto de corte e representam-se graficamente os pontos de coordenadas (1-especificidade, sensibilidade). Por convenção, 1-especificidade é representada no eixo das abcissas e a sensibilidade é representada no eixo das ordenadas, variando ambas de 0 a 1 (0-100%).

2.4.1 Representação da curva ROC

A curva ROC no plano unitário resulta da representação gráfica dos índices de precisão 1-especificidade *versus* sensibilidade, oriundas da variação do ponto de corte. Assim, a curva ROC é uma descrição empírica da capacidade e, consequentemente, da qualidade de um teste de diagnóstico diferenciar duas classes num universo [10].

Ao longo deste subcapítulo, far-se-á uso do exemplo ilustrativo apresentado no subcapítulo 2.3, o que permitirá ao leitor um melhor enquadramento e, consequentemente, mais rápida compreensão de como se constrói uma curva ROC.

Acrescentando informação ao exemplo a que se recorre, suponha-se que os 22 indivíduos foram sujeitos a uma análise clínica que pretende avaliar o nível de determinada hormona

na corrente sanguínea. Suspeita-se que os indivíduos com valores inferiores a 100 são indivíduos negativos e com valores superiores a 100 são positivos. A análise em causa não é um diagnóstico, daí não se poder dizer "com valores inferiores a 100 os indivíduos são saudáveis e com valores superiores a 100 são doentes".

Na tabela seguinte, estão representados os resultados hipotéticos da análise de cada indivíduo. Considerando que o estado 1 corresponde a um indivíduo doente e o 0 corresponde a um indivíduo saudável, sabe-se que dos 22 indivíduos, 8 estão efetivamente doentes e 14 estão efetivamente saudáveis. Dos doentes, 2 obtiveram resultado negativo (valor da medida igual a 87 e 90), enquanto, dos saudáveis, 4 obtiveram resultado positivo (valor da medida igual a 112, 130, 108 e 122), conforme ilustra a Tabela 3.

Estado	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Medida	121	110	105	147	178	167	87	90	112	130	108	122	56	45	56	89	65	45	34	74	54	55

Tabela 3 – Resultados do teste de diagnóstico.

Primeiramente, começa-se por representar graficamente as duas classes em estudo, doentes e saudáveis. Pode optar-se por um gráfico linear, de barras ou de dispersão. Esta representação permite ao investigador/observador localizar a "zona crítica", ou seja, a zona onde VP se cruzam com VN e vice-versa, a que correspondem os FP e FN. Para o caso em estudo obter-se-ia algo semelhante à Figura 5. O código usado para gerar esta ação encontra-se no Anexo A.5.

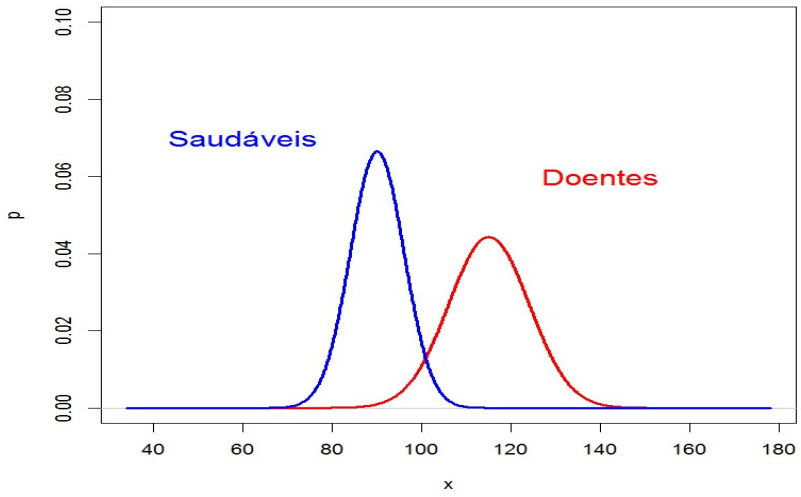


Figura 5 – Representação gráfica dos indivíduos doentes e saudáveis.

De seguida, analisam-se múltiplos valores para o ponto de corte, ao longo do eixo das abcissas e recolhem-se patamares de decisão ou *thresholds*. Um ponto de corte não repercute mais que uma reta vertical traçada sobre o gráfico cujo objetivo é registar o número de indivíduos classificados como doentes e como saudáveis nesse ponto.

Primeiramente, é importante salientar que o número de indivíduos classificados como doentes e saudáveis considerados no início do exemplo correspondem a um estado conhecido *a priori* e, por isso, inalterado. Pegando no ponto de corte 87, pretende saber-se quantos indivíduos são classificados como doentes, que são os que apresentam análise superior a esse valor, e quantos são classificados como saudáveis, que são os que apresentam análise inferior ou igual a esse valor. Após esta observação, é possível calcular a sensibilidade e a especificidade, verificando a fração de indivíduos corretamente identificados como doentes ou saudáveis. Observando a Tabela 3, tendo em conta este valor definido para o ponto de corte, é possível constatar que, no universo dos doentes, 7 são corretamente identificados como doentes, $s = 7/8 = 0.88$, e no universo dos saudáveis, 9 são corretamente identificados como saudáveis, $e = 9/14 = 0.64$, ou seja, $1\text{-}e = 0.36$. E assim sucessivamente se fará desde o valor de medida mais baixa até à medida mais alta.

Corte	<34	34	45	54	55	56	65	74	87	89	90
s	1	1	1	1	1	1	1	1	0,88	0,88	0,75
1-*e*	1	0,93	0,79	0,71	0,64	0,5	0,43	0,36	0,36	0,29	0,29
Corte	105	108	110	112	121	122	130	147	167	178	>178
s	0,63	0,63	0,5	0,5	0,38	0,38	0,38	0,25	0,13	0	0
1-*e*	0,29	0,21	0,21	0,14	0,14	0,07	0	0	0	0	0

Tabela 4 – Pontos de corte considerados.

Quanto maior for a dimensão da amostra mais pontos de corte existirão para a representação da curva e, consequentemente, mais precisa será a curva ROC representada. Os valores de x equivalem ao valor de 1-especificidade e os valores de y ao valor da sensibilidade. Conclui-se, assim, que cada ponto de corte dará origem a um ponto $(1\text{-}e; s)$ para a construção da curva ROC.

Por último, marcam-se todos os pontos (1-*e*; *s*) no gráfico, unem-se por retas e obtém-se a curva ROC, cf. Figura 6. Daí a importância de recolher o maior número de observações possíveis, para que a separação entre pontos seja menos notória. Na Figura 6, é exposta a curva ROC, obtida através do *software* R [21], para este exemplo. O código usado para gerar esta ação encontra-se no Anexo A.6, onde se recorreu ao *package* ROCR do *software* R [22], [23].

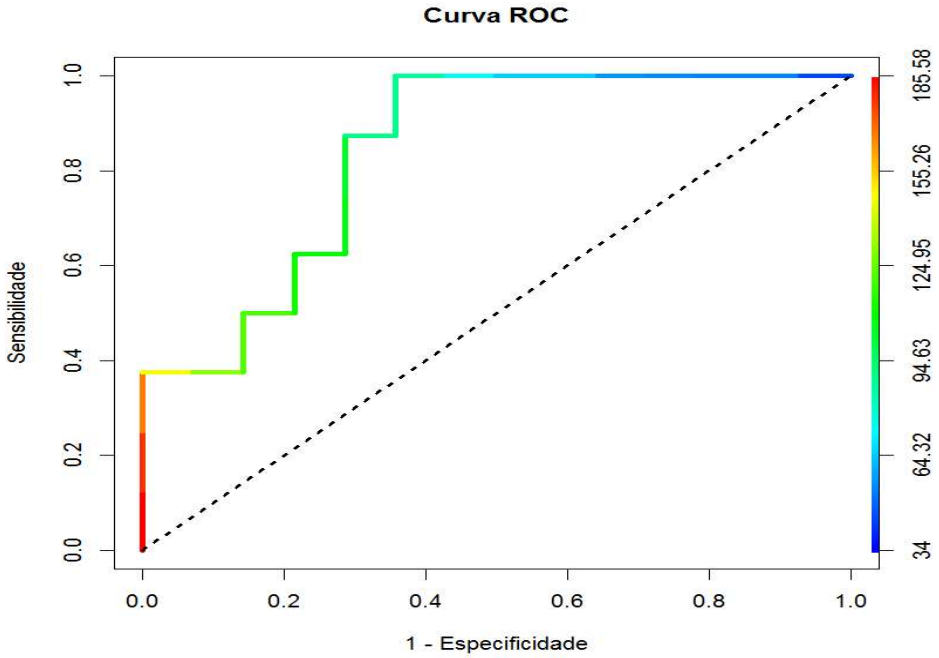


Figura 6 – Curva ROC do exemplo ilustrativo.

2.4.2 Níveis de discriminação e critérios de decisão

O ideal seria que as taxas de sensibilidade e especificidade fossem sempre iguais a 100%. Na prática há diversos testes, para certas doenças ou infeções, que são perfeitos: 100% de sensibilidade e de especificidade. Logo, estes testes não precisam da curva ROC.

Todavia, infelizmente as nossas limitações, de conhecimento ou de tecnologia, custo (na aplicação de cada teste ou no investimento necessário para o laboratório adquirir a tecnologia necessária para a realização do teste), ou à demora de esse teste fornecer os resultados, ou a agressividade do teste (pode causar lesões ao indivíduo) não permitem termos testes perfeitos para a maioria dos problemas. Para estes testes a curva ROC é uma ferramenta extremamente útil.

Nos subcapítulos 2.3.3 e 2.4.1 foram expostas algumas das razões que levam a privilegiar a sensibilidade ou a especificidade. Contudo, é importante salvaguardar que nem sempre o melhor ponto de corte teórico é o escolhido como decisão final. Isto porque:

- Pode requerer um elevado custo económico, tanto para o indivíduo, para a entidade prestadora de cuidados, para o estado ou ambas;
- Pode requerer um elevado custo físico, psicológico e/ou social para o indivíduo ou seus familiares.

Não menos frequente, também são os pontos de corte representados sob a forma de intervalos, devido ao facto de:

- Não haver unanimidade entre instituições de saúde, laboratórios, centros de investigação, profissionais de saúde, biólogos, cientistas, entre outros, para um valor fixo e todas concordarem na definição dum intervalo;
- Não ser relevante determinar a sensibilidade e a especificidade para todos os valores e fazerem-se, por exemplo, saltos de duas em duas, cinco em cinco ou mais unidades.

A visualização gráfica das estimativas de sensibilidade e 1-especificidade para diferentes pontos de corte, reflete a capacidade do teste em discriminar indivíduos doentes de saudáveis. A partir daqui cabe ao investigador decidir como melhor utilizar o teste.

Um teste que distingue perfeitamente indivíduos doentes de saudáveis é, para determinado ponto de corte, um teste tal que, a sensibilidade e a especificidade sejam iguais a 100%, ou seja, nesse caso a curva ROC está localizada no canto superior esquerdo, ao que corresponde um nível discriminação elevado. Já um teste totalmente incapaz de distinguir indivíduos doentes de saudáveis é, para determinado ponto de corte, um teste tal que, a sensibilidade e a especificidade sejam iguais, ou seja, $x=y$, ao que corresponde um nível de discriminação baixo. Estes e o outro nível de discriminação, médio, são apresentados na Figura 7.

Para avaliar o desempenho de testes de diagnóstico do mesmo tipo, e por conseguinte, compará-los, assumindo que as curvas não se cruzam, existem três níveis de descriminação: baixo, médio e elevado. Quanto mais próxima do canto superior esquerdo maior será o seu poder discriminante, ou seja, maior será a capacidade do teste em distinguir indivíduos doentes de saudáveis [8]. Na Figura 7 são apresentadas as curvas ROC dos diferentes tipos

de discriminação. Esta capacidade será tanto menor à medida que a curva se aproxima de $x=y$. O código usado para gerar esta ação encontra-se no Anexo A.7.

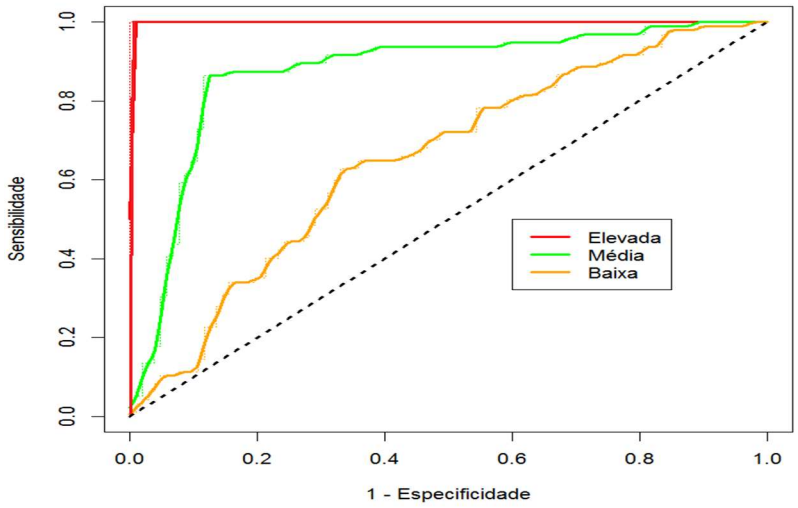


Figura 7 – Níveis de discriminação.

Conforme o que se pretenda privilegiar, sensibilidade, especificidade ou ambas, num teste de diagnóstico existem três critérios de decisão [8]:

- Critério estrito: gera um ponto na curva ROC que se situa no canto inferior esquerdo do espaço ROC. Significa que representa pequenas frações, tanto de verdadeiros positivos como de falsos positivos, logo representa situações com baixa sensibilidade mas elevada especificidade.
- Critério moderado: gera um ponto na curva ROC que se situa, aproximadamente, no meio do espaço ROC. Significa que representa fração de verdadeiros positivos relativamente superior a falsos positivos.
- Critério brando: gera um ponto na curva ROC que se situa no canto superior direito do espaço ROC. Significa que representa uma grande fração de verdadeiros positivos e uma pequena fração de falsos positivos, logo representa situações com elevada sensibilidade mas diminuta especificidade.

Estes critérios são apresentados na Figura 8 em que os pontos com abcissas 0.03, 0.2, e 0.65 correspondem, respetivamente, ao critério estrito, moderado e brando.

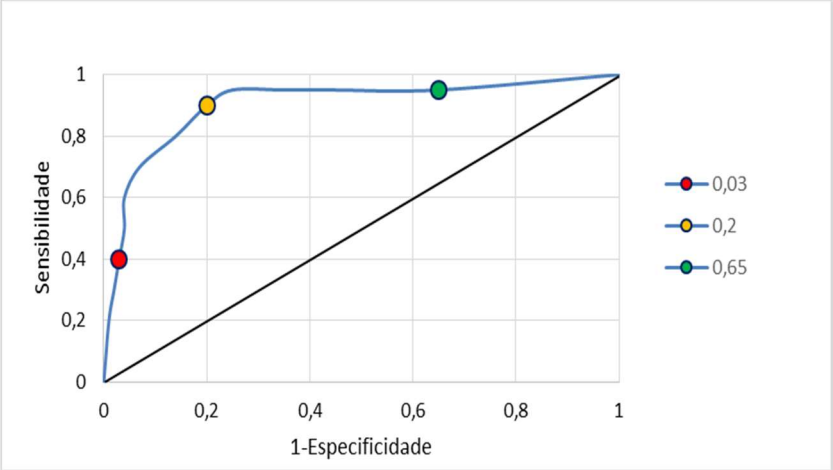


Figura 8 – Representação dos critérios de decisão.

2.4.3 Área abaixo da curva

A área abaixo da curva ROC é um dos índices de precisão mais utilizados para avaliar a qualidade da curva e o desempenho do teste de diagnóstico.

Para um dado indivíduo doente e outro saudável, ambos escolhidos ao acaso, a área abaixo da curva, é uma medida que permite aferir qual a probabilidade do indivíduo doente obter um resultado verdadeiro positivo e do indivíduo saudável obter um resultado verdadeiro negativo. Um teste totalmente inapto de discriminar indivíduos doentes de saudáveis, ou quaisquer outras duas classes, será o que apresentar uma área sob a curva de 0.5, pois significa que sensibilidade é sempre igual a 1-especificidade, ou seja, $x=y$. Como já foi previamente referido, quanto mais a curva se aproxima do canto superior esquerdo melhor é a qualidade do teste, logo, quanto maior for o valor da área isto é, quanto mais próximo estiver da unidade, maior é a capacidade para discriminar estes dois tipos de indivíduos [10].

Alguns autores preferem utilizar, no eixo das abcissas, especificidade em vez de 1-especificidade, o que na realidade acaba por ser irrelevante visto que a área será numericamente a mesma, apenas altera a concavidade da curva de baixo para cima [10].

A área sob a curva ROC resume-a como um todo, o que pode ser inadequado, uma vez que, na realidade, as decisões clínicas apenas se baseiam numa parte da curva. Pode, por isso, usar-se uma medida alternativa a esta situação, a que se denomina de área parcial sob a curva,

estimada sob a região de interesse da curva. Para mais informações sobre a área parcial sob a curva ROC consultar, por exemplo, [24][25][26].

2.4.4 Problemática da comparação entre curvas

Os gráficos que representam duas ou mais curvas ROC, na mesma escala, associadas a diferentes testes de diagnóstico permitem uma comparação rápida e direta do desempenho teórico dos mesmos.

A escolha do melhor teste diagnóstico não depende apenas do melhor desempenho, uma vez que esse desempenho pode representar mais custos, dos mais variados tipos, e mais tempo e a relação entre estes últimos não ser significativamente relevante que torne justificável a preferência de uma curva, matematicamente melhor, ao invés de outra.

Quando se comparam duas curvas de testes distintos podem ocorrer as situações representadas nas figuras seguintes [8].

Situação A: as curvas são diferentes e não se cruzam e a curva com maior área é a que afere os testes de diagnóstico com melhor desempenho. Note-se que nestes casos, como ilustra a Figura 9, para qualquer valor de especificidade o teste associado à curva ROC 1 obtém sempre melhor valor para a sensibilidade que o outro teste que deu origem à curva ROC 2.

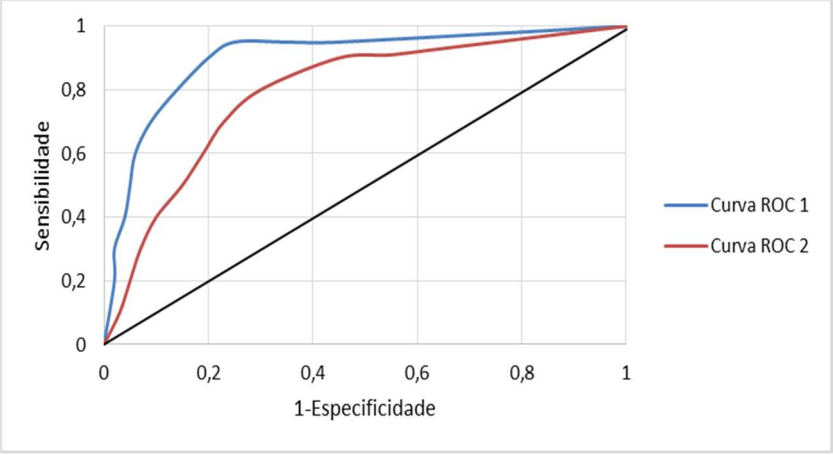


Figura 9 – Situação A.

Situação B: as curvas são diferentes num dado intervalo, mas iguais noutro, sendo a curva com maior área a que traduz os testes de diagnóstico com melhor desempenho. Assim, nesta situação, pode verificar-se um melhor desempenho teórico da curva ROC 1. Note-se que nestes casos, como ilustra a Figura 10, para valores de especificidade entre 0 e 0.4 o teste associado à curva ROC 1 obtém sempre melhor valor para a sensibilidade que o outro teste que deu origem à curva ROC 2.

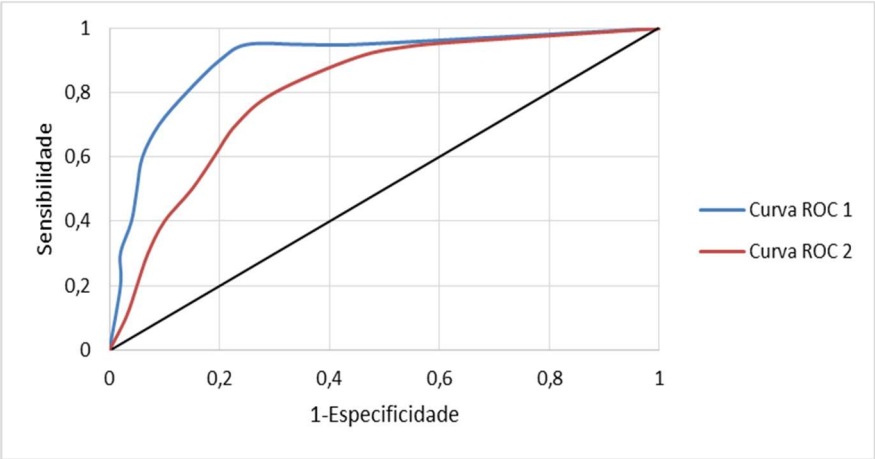


Figura 10 – Situação B.

Situação C: as curvas cruzam-se e, apesar das áreas serem próximas, os testes de diagnóstico aferem desempenhos diferentes. Nesta situação, pode-se dar o caso de se usarem os dois testes de diagnóstico, em que um complementa o outro. Tendo em conta o exemplo da Figura 11, o teste correspondente à curva ROC 1 poderia ser utilizado primeiro para identificar a maioria dos indivíduos saudáveis e, numa segunda fase, para os indivíduos que foram identificados como doentes na primeira fase, fazer o teste correspondente à curva ROC2, uma vez que este segundo teste é melhor a identificar os indivíduos que estão de facto doentes. Esta viragem é feita próxima do ponto (0.35; 0.7).

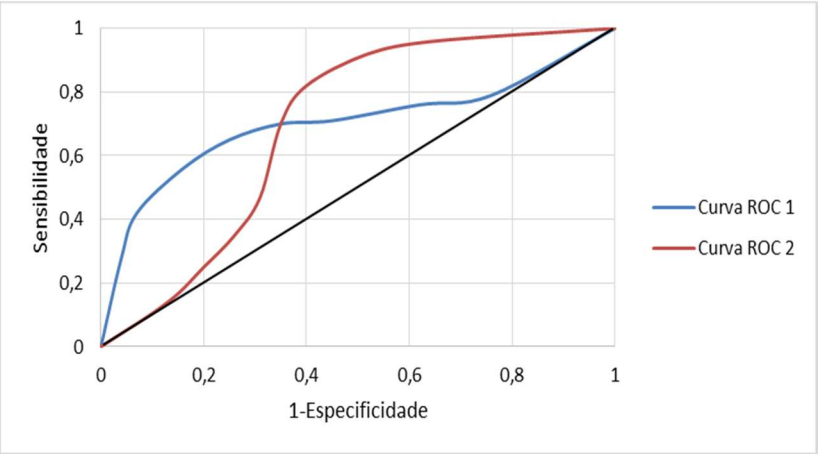


Figura 11 – Situação C.

Quando se comparam duas curvas ROC diferentes que possuam combinações de sensibilidade e 1-especificidade relevantes num intervalo de pontos de corte é, muitas vezes, mais vantajoso utilizar métodos de comparação para esse intervalo específico em vez de para toda a curva [27].

3. Caso de estudo

No decorrer deste capítulo será apresentado o caso de estudo: a deteção do cancro da pele através do padrão reticular. Primeiramente, será feita uma breve introdução do mesmo. De seguida, explana-se a técnica da dermatoscopia, introduz-se a base de dados utilizada, bem como as respetivas características e detalhes e faz-se uma sintetização sobre processamento de imagem. Para terminar, aplicar-se-á a curva ROC para indagar conclusões relativamente a este caso de estudo, onde serão aplicados alguns dos conceitos básicos supra descritos, ao longo do capítulo 2.4. Neste capítulo, será apresentado o tratamento dos dados na procura do "melhor" teste, no sentido de maximizar a área abaixo da correspondente curva ROC, e do "melhor" ponto de corte, no sentido de maximizar a soma da sensibilidade e especificidade e minimizar a distância ao ponto ótimo (0,1). Em suma, far-se-á a comparação dos resultados obtidos através da aplicação de diferentes metodologias.

3.1. Introdução

O cancro caracteriza-se por ser o grupo de doenças em que os sintomas se devem ao crescimento descontrolado de células num dos órgãos ou tecidos do corpo. Mais vulgarmente, desenvolvem-se tumores malignos nos grandes órgãos, tais como pulmões, mamas, intestinos, pele, estômago ou pâncreas, mas também podem aparecer nos seios nasais, nos testículos, nos ovários, nos lábios, na boca, etc. Os cancros podem igualmente surgir nos tecidos da medula óssea, as leucemias, no sistema linfático, nos músculos ou nos ossos. O cancro é a segunda maior causa de morte em Portugal (a seguir às doenças cardiovasculares) [1].

Ao desenvolver-se, o cancro, através da circulação sanguínea, espalha-se e infiltra-se nos tecidos que o circundam e pode bloquear canais, destruir nervos e corroer ossos. As células cancerosas podem disseminar-se, transportadas através dos vasos sanguíneos e dos canais linfáticos, para outras partes do corpo, onde as metástases ou novos tumores satélite se desenvolvem e crescem autonomamente [1].

O tumor maligno da pele constitui uma das formas mais comuns de cancro. O carcinoma basocelular, o carcinoma espinocelular e o melanoma maligno são formas comuns de cancro

de pele relacionadas com uma exposição prolongada ao sol. A doença de *Bowen*, doença cutânea rara que pode tornar-se cancerosa, pode também estar relacionada com a exposição ao sol [1].

Tipos menos frequentes de cancro de pele incluem a doença de *Paget* e a micose fungóide. Ambas apresentam um aspeto parecido com o do eczema. O sarcoma de Kaposi é um tipo de cancro de pele presente em muitos doentes com sida [1].

Ainda que, na sua maior parte, os cancros de pele possam ser facilmente curados se forem tratados numa fase inicial da doença, muitas pessoas morrem por tardarem a consultar o médico, sobretudo quando contraem carcinoma espinocelular e melanoma maligno. Qualquer ulceração ou lesão com crosta de cicatrização difícil ou alteração da forma ou aspeto de mancha ou lesão cutânea devem ser comunicadas ao médico.

As doenças da pele são, na sua maioria, diagnosticadas pelas suas características físicas. Pode proceder-se a uma biopsia da pele para auxiliar o diagnóstico da doença cutânea ou para excluir ou confirmar o cancro da pele. Contudo, os avanços científico-tecnológicos das últimas décadas têm permitido apostar numa medicina mais preventiva em prol de uma medicina curativa, que no caso dos cancros será sempre uma mais-valia, não só porque aumenta, consideravelmente, as hipóteses de sobrevivência como evita os efeitos secundários penosos da quimioterapia. Consequentemente, torna-se fundamental desenvolver meios capazes de responder a esta necessidade, meios capazes de auxiliar o diagnóstico médico.

Como é espectável, nenhum meio, especialmente os que envolvem processamento de informação biológica, pode ser posto em prática sem uma bateria de testes, aprovações e opiniões prévias por parte de diferentes profissionais de saúde.

3.2. Dermatoscopia

A dermatoscopia é uma técnica de imagem *in vivo*, não invasiva, utilizada principalmente para avaliar as cores e estrutura das lesões pigmentadas da pele, que se está a tornar, exponencialmente, num meio auxiliar da medicina preventiva. Sucintamente, é uma

ampliação microscópica de lesões, pigmentações, pintas, entre outras alterações suspeitas. Este exame avalia as seguintes características: simetria ou assimetria, homogeneidade ou heterogeneidade, distribuição dos pigmentos, queratina da superfície da pele, morfologia e padrão vascular, bordas da lesão, presença de ulcerações. Estas imagens são, portanto, uma fonte desafiadora de uma ampla gama de características digitais, frequentemente com associação clínica. Entre estes marcadores, um de interesse particular para o diagnóstico na avaliação da pele é o padrão reticular [4].

Normalmente, os melanomas têm estruturas diferenciais, tais como redes pigmentadas, estrias ou pontos, ajudando a revelar a sua origem maligna. A rede pigmentada é uma importante pista diagnóstica, representando uma marca dermatoscópica de lesões malignas, cuja presença é, em geral, independente da presença ou ausência de um processo carcinogénico. O padrão reticular aparece como uma grade de linhas castanhas finas, sobre um fundo castanho claro difuso. Trata-se de uma estrutura em forma de favo de mel, constituída por linhas redondas pigmentadas e orifícios mais leves e hipo-pigmentados, formando um padrão subtil que aparece em muitas lesões malignas [4].

A deteção automatizada do padrão reticular ou rede de pigmentos é frequentemente um problema desafiador, uma vez que nestas estruturas reticulares existe um baixo contraste entre a rede e o fundo. O tamanho dos orifícios da rede pode compreender tamanhos diferentes em imagens diferentes e também, na mesma imagem, muitas vezes existem irregularidades na sua forma e tamanho [4][28].

3.3. Processamento de imagem

A captação de imagem é desde os seus primórdios um fator de extrema importância para variadíssimas áreas de negócio e prestação de serviços. Desde cedo se percebeu que a imagem após a sua obtenção não apresentava as características ideais para recolher dados que se irão transformar em informação e mais tarde em conhecimento. O avanço da tecnologia permitiu expandir a qualidade das técnicas, das ferramentas e a qualidade dos resultados conseguidos [29].

O processamento de imagem foi inicialmente adotado nas áreas que requerem a necessidade de ver para além do que o olho "nu" consegue ver. Por exemplo, nos aeroportos para a inspeção de bagagens, na construção civil para analisar estruturas de difícil acesso através da termodinâmica dos corpos, no setor militar para a deteção de armas químicas e na medicina para o tratamento, diagnóstico e prevenção [30].

A matemática é a base de vastos fundamentos da vida humana, é aplicada em diversos princípios e ciências, sendo o processamento de imagem uma delas. Uma imagem não é mais do que uma matriz de valores com uma dimensão de *m* por *n*. Em cada posição dessa matriz é definido um píxel da imagem que pode compreender valores entre 0 (preto) e 255 (branco). O conhecido conceito RGB não é mais do que a definição das três camadas que formam uma imagem a cores, *Red Green Blue* (RGB). Uma imagem a cores é composta por um conjunto de três matrizes sobrepostas, com as mesmas dimensões, que apresentam a imagem tal como é conhecida nos dias de hoje. A diferença essencial entre uma imagem a cores e uma a preto e branco é o número de matrizes. Uma imagem a preto e branco apenas tem uma matriz (camada). Considerando que os píxeis podem ir de 0 a 255, numa imagem que apenas tenha uma matriz, todos os seus píxeis estarão compreendidos na escala de cinza, de branco a preto passando por todas as tonalidades de cinza limitadas a 254 valores distintos [31].

A algoritmia tem acompanhado os desenvolvimentos das novas tecnologias, pois estes obrigaram à constante necessidade de criar algoritmos mais robustos, otimizados e adequados aos requisitos da ciência [30].

O processamento de imagem consiste na aplicação de algoritmos matemáticos à ou às matrizes que constituem uma imagem. Os algoritmos aplicados são desenvolvidos conforme a informação que se pretende extrair de uma imagem, por exemplo: extrair contornos, realçar certas regiões de acordo com uma condição e destacar elementos pela sua forma. As diversificadas técnicas aplicadas em imagens não são mais do que fórmulas matemáticas aplicadas em cada píxel da imagem em análise. De acordo com o tipo de processamento que se pretende efetuar, os algoritmos podem ser mais ou menos complexos e robustos, mas por norma nunca analisam um píxel individualmente, pois a sua vizinhança tem sempre imensa informação a acrescentar [31].

Assim sendo, inicialmente, pode considerar-se que o processamento de imagem não é mais do que a aplicação de algoritmos de acordo com os nossos objetivos, mas existe um fator relevante até agora não mencionado. A obtenção da imagem, seja por câmaras ou aparelhos mais específicos, está sempre condicionada a um conjunto enorme de fatores cujo controlo sobre eles é, na maioria das vezes, muito limitado. Exemplos destes fatores são a intensidade da luz, o ponto a capturar e as características dos equipamentos. Com isto podemos concluir que uma imagem de um objeto idêntico pode sofrer processamentos completamente distintos e executados em ordens distintas devido às condições que influenciaram a sua obtenção [31].

3.4. Base de dados

Para este caso de estudo foi considerado um conjunto de dados correspondente a informação sobre imagens dermatológicas, no sentido de averiguar se cada imagem apresenta ou não padrão reticular [4].

Inicialmente, este conjunto contemplava dezenas de imagens que foram analisadas por três dermatologistas que registaram, de forma independente, o seu diagnóstico sobre a existência ou não de padrão reticular em cada imagem [4].

Do conjunto inicial de imagens, houve concordância dos dermatologistas em cento e cinquenta e oito imagens. Esse diagnóstico foi registado e guardado. Pretende-se, então, recorrer a métodos de processamento de imagem e a metodologias de classificação para que se consiga, sem recorrer a dermatologistas especializados, concluir o mesmo diagnóstico.

Para tal, cada imagem foi avaliada sob setecentos e dois algoritmos distintos isto é, tem-se para cada imagem setecentas e duas medidas disponíveis. Estas estão divididas em três avaliações distintas: desvio padrão, energia e entropia. Ou seja, cada avaliação contempla duzentos e trinta e quatro resultados, 702/3. O primeiro bloco de resultados corresponde à avaliação baseada no desvio padrão, o segundo bloco de resultados corresponde à avaliação referente à energia e o último bloco de resultados corresponde à avaliação relativa à entropia. Dentro de cada um destes três tipos de avaliações, os resultados estão divididos em nove escalas (k=1,…,9) que não são mais que diferentes frequências da mesma imagem. Assim,

para cada escala têm-se vinte e seis resultados distintos, 234/9, que se irá denotar por m=1,…,26.

Deste modo, para cada avaliação (desvio padrão, energia ou entropia) há 234 resultados, os quais são representados pelo par (k, m) com k =1,..., 9 a identificar a escala e m=1,..., 26 a identificar cada resultado.

Estes resultados já foram igualmente analisados no artigo [4], tendo estas medidas sido determinadas pelos seus autores com recurso ao *software* MATLAB. Deste modo, reitera-se o agradecimento aos autores do referido artigo pela cedência destes dados. Será igualmente importante salientar que não é objetivo da presente dissertação explicar as técnicas de processamento de imagem utilizadas, nem os algoritmos utilizados para a obtenção das 702 medidas, nem o seu significado, mas antes aplicar estes valores recorrendo aos conceitos desenvolvidos ao longo do capítulo 2.

Como foi descrito anteriormente, no subcapítulo 3.3, uma imagem não é mais que uma matriz com uma determinada dimensão. Sendo uma matriz composta por números inteiros entre 0 e 255, há todo um conjunto de operações matemáticas que podem ser aplicadas.

3.4.1 Desvio padrão

O desvio padrão de uma imagem é uma medida de dispersão dos valores dos píxeis em relação à sua média. Assim, quanto maior for a dispersão entre todos os píxeis que constituem a imagem em análise, maior será o valor do desvio padrão, sendo nulo quando os píxeis assumem todos o mesmo valor.

3.4.2 Energia

A energia é uma medida de textura aplicada a imagens, ou seja, é uma medida que pretende avaliar a quantidade de informação presente na imagem, que corresponde à soma dos píxeis da imagem ao quadrado.

3.4.3 Entropia

Uma outra medida muito aplicada às imagens é o cálculo da entropia dessa imagem. A entropia de uma imagem é um valor obtido através da aplicação de uma fórmula matemática que terá em conta todos os píxeis da mesma. Este valor quantifica a aleatoriedade da imagem, ou seja, quanto maior o valor da entropia, mais atípica, irregular ou não padronizada será a imagem analisada. Por exemplo, caso seja determinada a entropia de uma imagem em *greyscale* que apresente em todos os seus píxeis o mesmo tom de cinza, ter-se-á uma entropia muito próxima de zero ou mesmo zero. Caso uma imagem se pareça com um típico código de barras o seu valor de entropia será bastante elevado, pois os píxeis da imagem apenas apresentam os valores 0 ou 255.

3.4.5 A estreita relação entre desvio padrão e entropia

O desvio padrão, energia e a entropia de uma imagem são valores distintos calculados a partir de fórmulas bem diferenciadas, no entanto, representam informação que se relaciona. Obtendo um valor de desvio padrão elevado, devemos também esperar uma entropia elevada, pois o desvio padrão elevado representará a existência de um número significativo de píxeis que se encontram distantes do valor médio da imagem e esta discrepância entre valores de píxeis da imagem resultará num valor significativo da entropia.

3.4.6 Diagrama da obtenção da base de dados

O Diagrama 1 resume o processo de obtenção dos dados, isto é, das 702 medidas de cada imagem que serão analisadas no capítulo seguinte, na tentativa de encontrar um teste fiável para a deteção de um padrão reticular.

Exemplo de uma das 158 imagens

Figura 12 – Cancro da pele [32]

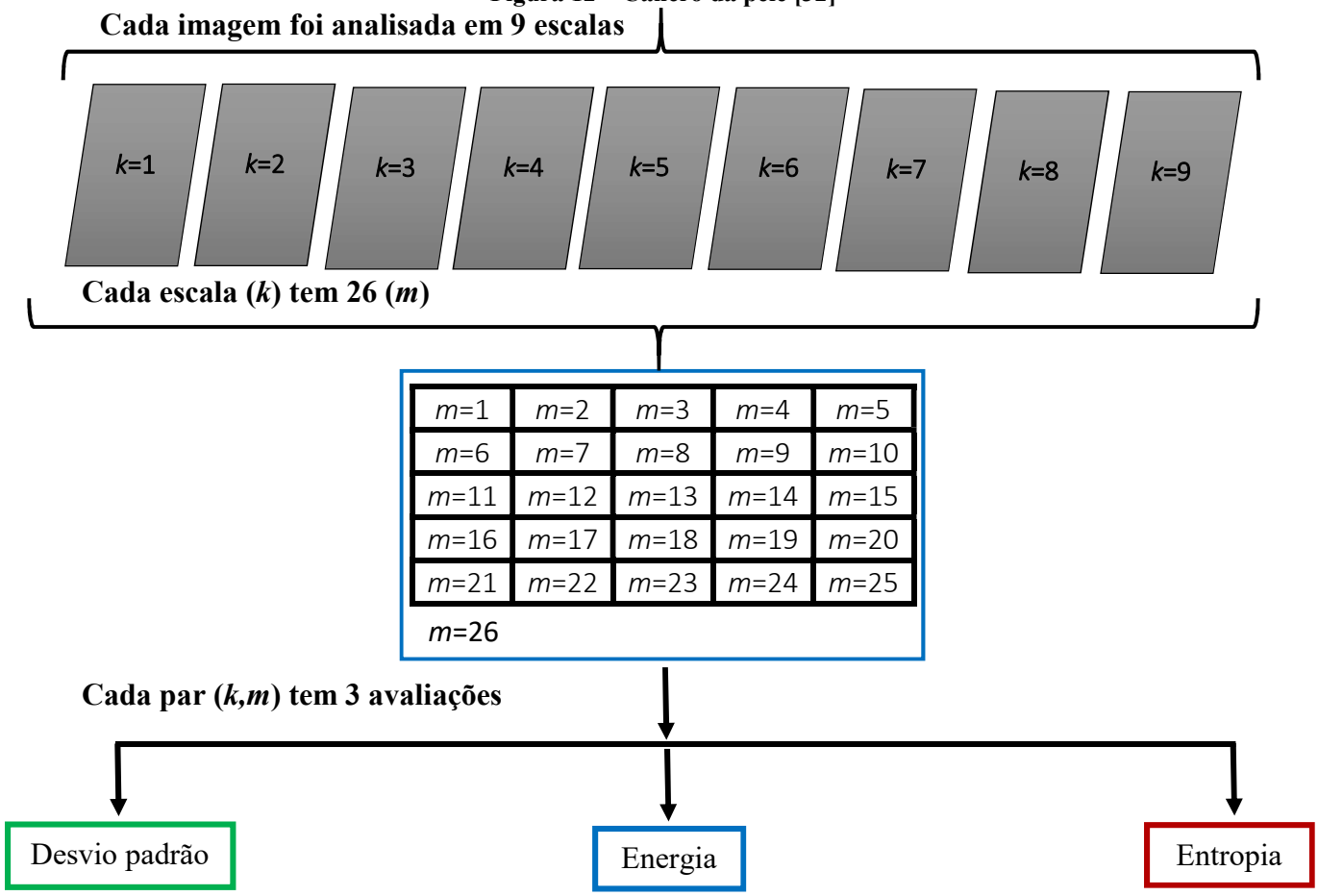


Diagrama 1 – Esquematização da obtenção das 702 medidas para cada imagem.

4. Aplicação da curva ROC ao caso de estudo

Conforme referido, o objetivo da utilização desta base de dados de imagens dermatológicas é aferir a existência ou não de padrão reticular em cada imagem. No âmbito desta dissertação, pretende-se analisar, através da aplicação da curva ROC, até que ponto a aplicação do processamento de imagem no diagnóstico produz resultados úteis e fiáveis.

Uma vez que os dados foram obtidos através da aplicação de três medidas distintas, denominadas por avaliações, desvio padrão, energia e entropia, tornou-se importante *standardizá-los* a fim de tornar lógicas comparações de grandezas entre eles.

A *standardização* de dados consiste na alteração dos dados, de modo a que os resultados obtidos dessa transformação sejam mais facilmente comparados, uma vez que todas as variáveis passam a ter a mesma localização (média igual a zero) e a mesma escala (desvio padrão igual a um). Note-se que, com a aplicação desta transformação, as observações mantêm a mesma ordem e, como tal, dão origem à mesma curva ROC. Desta forma, no cálculo da média entre duas medidas, é atribuído o mesmo peso a todas as variáveis, evitando que uma medida com valores elevados suprima outra medida com valores próximos de zero, apesar de cada variável passar a variar entre valores distintos dos iniciais. Portanto, todo o estudo e análise realizados serão efetuados sob dados *standardizados*, Anexo A.8.

Antes de avançar com o estudo em si, tornaram-se simétricos os valores das colunas cuja curva ROC está maioritariamente abaixo da reta $y=x$ (área sob a curva ROC inferior a 0.5), isto porque a classificação apenas está inversa ao que se pretende estudar, Anexo A.9.

Inicialmente, com o conjunto das cento e cinquenta e oito imagens e setecentas e duas medidas, optou-se por construir as curvas ROC para cada uma das medidas. Obtiveram-se assim setecentas e duas curvas ROC distintas, concluindo-se de imediato que algumas medidas dariam um contributo bastante melhor que outras na deteção do padrão reticular, conforme se pode observar na Figura 12. O código usado para gerar estas curvas ROC é apresentado no Anexo A.10.

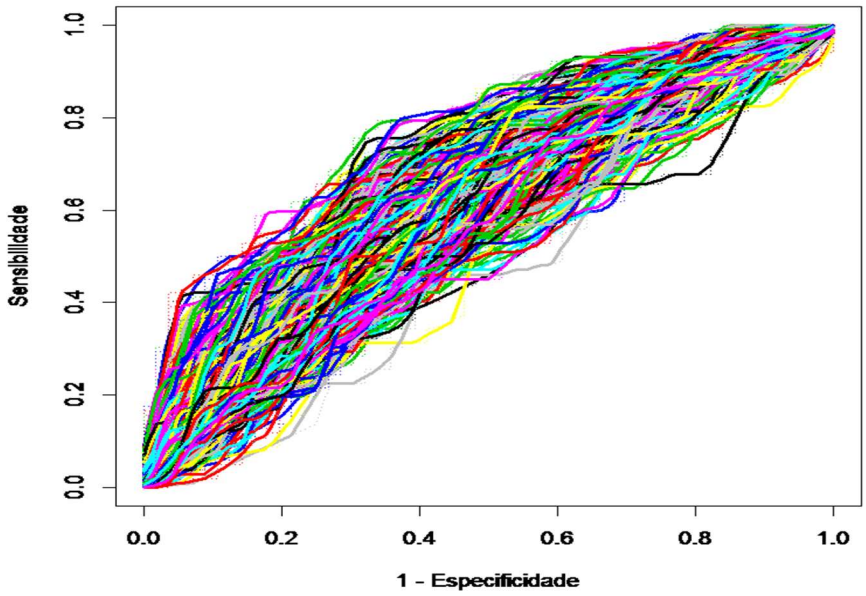


Figura 13 – Representação gráfica das 702 curvas ROC do caso de estudo.

De seguida, foi-se investigar, através do código apresentado no Anexo A.11, qual a curva com maior área abaixo da curva ROC, chegando-se à conclusão que era a coluna 486 (avaliação pela entropia, primeira escala, m=18), com 0.753 de área. Posteriormente, representaram-se as setecentas e duas medidas com a coluna 486 em destaque. O código que a gerou pode ser consultado no Anexo A.12.

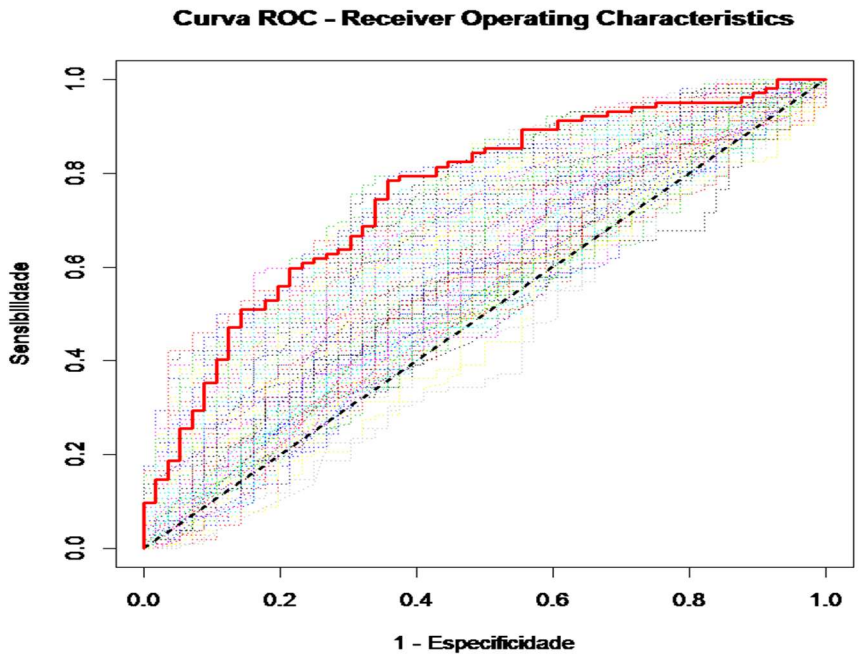


Figura 14 – Destaque da curva ROC com maior área.

Uma vez que seriam demasiadas curvas para analisar individualmente, o que inviabiliza a tomada de conclusões, foi necessário começar a particularizar o estudo das curvas.

Recordando o último período do subcapítulo 3.3, existem nove escalas por cada tipo de avaliação, cada uma com vinte e seis medidas. Para facilitar a escrita e compreensão, representar-se-á, analogamente, cada escala pela letra *k*.

Deste modo, dividiu-se a análise em blocos e representaram-se graficamente as curvas ROC dos vinte e sete *k* totais. Isto originou vinte e sete gráficos, cada um com vinte e seis curvas, correspondentes às vinte e seis medidas analisadas em cada escala.

Na Figura 14, encontra-se o gráfico com *k* =1 para o desvio padrão, na Figura 15, encontra--se o gráfico com *k* =1 para a energia e, na Figura 16, encontra-se o gráfico com *k* =1 para a entropia. Os restantes são exibidos no anexo B e o código que lhes deu origem é apresentado no Anexo A.13.

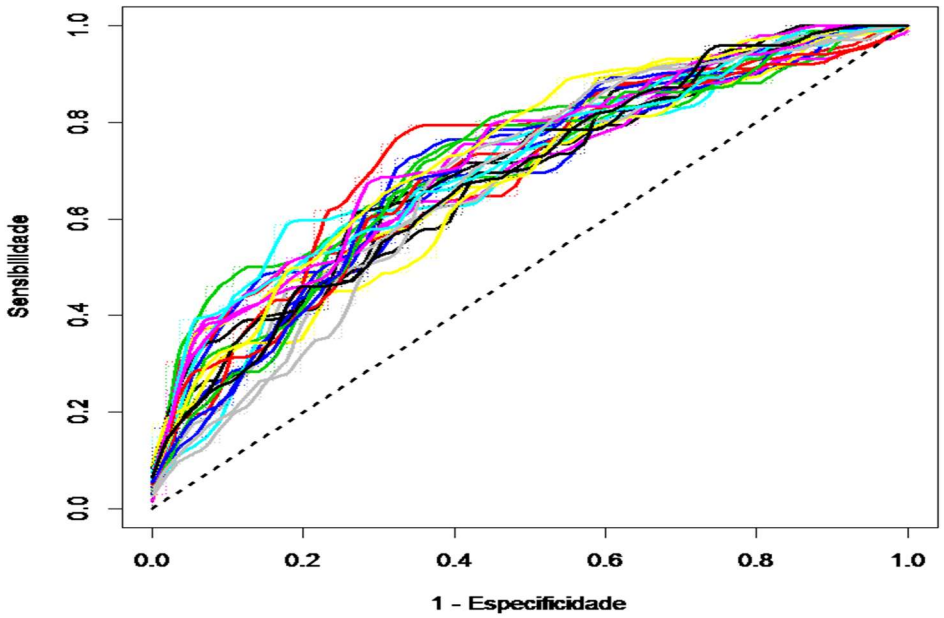


Figura 15 – Curvas ROC com *k*=1 para o desvio padrão.

Considerando a Figura 14, é possível constatar que todas as curvas estão acima do eixo *x*=*y* e têm curvaturas bastante semelhantes. Aparentemente, poderão ser bons indicativos da capacidade da deteção do padrão reticular considerando a utilização da escala *k*=1 e o desvio padrão como avaliação.

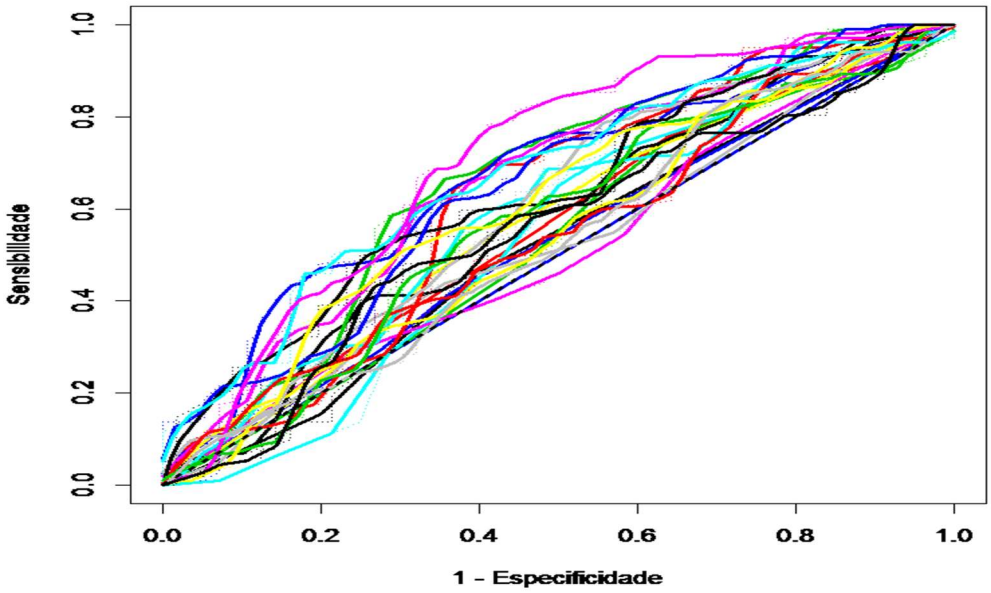


Figura 16 – Curvas ROC com k=1 para a energia.

Note-se que, na Figura 15, as curvas têm maior dispersão e embora estejam, de um modo geral, acima do eixo $x=y$, estão bastante próximas deste o que não perspetiva vir a ser uma boa escala para a deteção do padrão reticular, utilizando k=1 e a energia como avaliação.

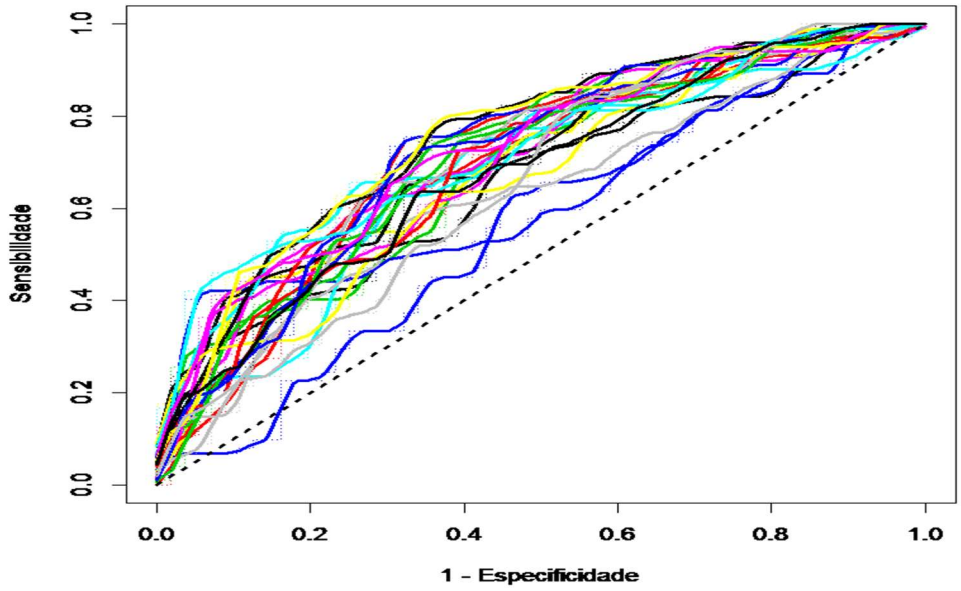


Figura 17 – Curvas ROC com k=1 para a entropia.

Tal como a Figura 14, a Figura 16 aparenta retratar uma boa escala na deteção do padrão reticular. Todavia, as curvas ROC apresentam áreas ligeiramente inferiores e, consequentemente, mais próximas do eixo $x=y$, o que não abona a seu favor.

Comparando as Figuras 14, 15 e 16 com a Figura 12, é facilmente deduzível que se torna bastante mais prático e razoável fazer o estudo de dados por k, ou seja, por vinte e sete grupos com vinte e seis medidas cada, do que com a totalidade das setecentas e duas medidas. Assim sendo, todo o estudo e análise de dados que se seguem, serão agrupados pelo valor de k em cada uma das avaliações.

No Anexo B, é possível verificar que as curvas ROC das medidas calculadas para o desvio padrão (figuras de B.1 a B.9) são as que aparentam alcançar um melhor resultado, pois são as que estão, com maior frequência, mais afastadas da linha $x=y$. Contudo, este continuaria a ser um estudo muito superficial e pouco crítico, caso a análise fosse restrita à observação das curvas ROC.

4.1. Procura da "melhor" curva ROC

Assim, com o objetivo de tornar a investigação mais fidedigna e mensurável, decidiu-se iniciar um estudo mais crítico, através de resultados numéricos. Por isso, deu-se início à procura da "melhor" curva ROC, aplicando três estratégias diferentes. A primeira consiste em achar a melhor curva por escala, a segunda pretende analisar a curva gerada pela medida que resulta da aplicação da média aritmética e a terceira aplica uma média ponderada a cada escala, de modo a obter a curva daí resultante. Recorde-se que a importância da área abaixo da curva encontra-se explanada no subcapítulo 2.4.3.

4.1.1 Medidas não agrupadas

Na prática, calculou-se para as vinte e seis medidas a respetiva área abaixo da curva ROC. Posteriormente, determinou-se a área máxima e a respetiva medida, dentro das vinte e seis medidas, para o k introduzido. O código para o cálculo da melhor área e respetiva medida encontra-se no Anexo A.14 e os resultados são apresentados na Tabela 5.

Desvio padrão			Energia			Entropia		
Área	*m*	*k*	Área	*m*	*k*	Área	*m*	*k*
0.7382703	**24**	**1**	0.5539216	**260**	**1**	0.7533263	**486**	**1**
0.7186625	**40**	**2**	0.6209734	**284**	**2**	0.7193627	**508**	**2**
0.6353291	**57**	**3**	0.6412815	**312**	**3**	0.6388305	**526**	**3**
0.6451331	**81**	**4**	0.6435574	**315**	**4**	0.6456583	**549**	**4**
0.6754202	**106**	**5**	0.6857493	**341**	**5**	0.6750700	**574**	**5**
0.6797969	**134**	**6**	0.6801471	**385**	**6**	0.6806723	**602**	**6**
0.6843487	**169**	**7**	0.6936275	**392**	**7**	0.6843487	**637**	**7**
0.6327031	**195**	**8**	0.6491597	**427**	**8**	0.6327031	**663**	**8**
0.6339286	**227**	**9**	0.6544118	**460**	**9**	0.6342787	**695**	**9**

Tabela 5 - Maior área para cada *k* e medida correspondente.

Note-se que para o desvio padrão e para a entropia as duas primeiras escalas são as que apresentam maior valor da área. Já para a energia destacam-se as escalas cinco, seis e sete.

A primeira escala do desvio padrão e da entropia são as que apresentam maiores valores de área, sendo que a da entropia se destaca. A segunda escala, também do desvio padrão e da entropia, estão na frente relativamente às que lhes sucedem e a qualquer uma da energia.

Pode aferir-se que, entre as três avaliações distintas, a energia é a que revela pior desempenho na determinação do padrão reticular e o desvio padrão e a entropia refletem, aparentemente, um desempenho semelhante, quando a *performance* é medida pela análise da área sob a curva ROC de cada medida.

Com o intuito de encontrar um padrão, ou pelo menos identificar as medidas ou os *k* que demonstram pior desempenho na deteção do padrão reticular, foram identificadas as curvas cujas áreas são inferiores a 0.65. Na Tabela 6, estão representadas as nove escalas por cada avaliação, bem como o número de medidas com área inferior 0.65. O código que retornou estes resultados encontra-se no Anexo A.18.

Através da análise da Tabela 6, é possível validar que a energia é a avaliação com mais medidas cuja área abaixo da curva é inferior a 0.65. Para as nove escalas, três têm todas as medidas e outras quatro praticamente todas as medidas com áreas inferiores a 0.65.

Desvio padrão		Energia		Entropia	
k	**Nº.Medidas**	***k***	**Nº.Medidas**	***k***	**Nº.Medidas**
1	2	**1**	20	**1**	4
2	17	**2**	26	**2**	17
3	26	**3**	26	**3**	26
4	24	**4**	24	**4**	24
5	15	**5**	6	**5**	15
6	10	**6**	7	**6**	10
7	19	**7**	20	**7**	19
8	26	**8**	26	**8**	26
9	26	**9**	23	**9**	23

Tabela 6 – Número de medidas, por *k*, com áreas inferiores a 0.65.

Em suma, usando esta metodologia conclui-se que a escala um para a entropia é a melhor escala, seguindo-se a escala um do desvio padrão e comparando a três avaliações constata-se que a energia é a que apresenta, globalmente, valores mais baixos para a área abaixo da curva ROC. Esta é mais uma evidência que sustenta as suspeitas da energia ser a menos eficiente das três avaliações para detetar o padrão reticular.

Pode, também, constatar-se que as escalas entre dois e nove para o desvio padrão e entropia coincidem exatamente no mesmo número de medidas menos capazes de achar o padrão reticular. As melhores, para ambas as avaliações, são as escalas um e seis.

Concluiu-se, então, que a energia é a avaliação com pior desempenho na deteção do padrão reticular, assim como as escalas três, quatro, oito e nove do desvio padrão e da entropia.

4.1.2 Medidas agrupadas

Para aprofundar a análise sobre qual das medidas ou qual das escalas melhor se adequa à deteção do padrão reticular, é necessário explorar e tratar os dados. Assim sendo, é necessário aplicar diferentes metodologias a fim de que uma, ou mais, permitam identificar de forma objetiva a melhor medida e respetivas escalas, bem como fazer comparações entre elas e tirar conclusões.

Deste modo, foram utilizadas metodologias com dois objetivos. Em primeiro lugar, na procura de obter curvas com áreas associadas superiores, agruparam-se medidas. Este agrupamento é feito utilizando uma função de várias medidas de forma a obter-se uma única medida para a elaboração da curva ROC. Foi utilizada a média aritmética em cada escala, bem como a média das medidas que isoladamente demonstraram melhor *performance*. Posteriormente, foi utilizado o método Adaboost, um processo iterativo que procura a ponderação a atribuir a cada medida de forma a maximizar a área abaixo da curva ROC, permitindo assim o cálculo da medida correspondente à média ponderada das medidas selecionadas e o cálculo da área abaixo da respetiva curva ROC.

4.1.2.1 Medidas agrupadas através da média aritmética

Esta metodologia consiste em, primeiramente, utilizar a média aritmética para cada escala (cada valor de *k*) dentro da mesma avaliação na determinação da curva ROC. E, seguidamente, depois de determinar as medidas correspondentes às médias das vinte e seis medidas para cada escala *k* e para cada avaliação, representaram-se as curvas ROC correspondentes a estas novas medidas e calculou-se a área abaixo das respetivas curvas.

Sabe-se que uma das vantagens da média aritmética é resultar num equilíbrio entre todas as medidas da amostra, até porque, conforme previamente referido, todas as medidas estão *standardizadas*. Todavia, uma desvantagem da utilização desta média é que a presença de valores muito discrepantes numa das medidas pode ser dissimulada ao juntar os valores das restantes 25 medidas. Assim, as diferenças nos valores das medidas associadas aos indivíduos infetados e aos não infetados podem ficar atenuadas, o que poderá resultar numa diminuição de eficácia na deteção do padrão reticular.

Foram, então, construídas as curvas ROC correspondentes à média para cada uma destas vinte e sete escalas e calculadas as respetivas áreas abaixo da curva. Deste modo, com esta metodologia, obteve-se uma curva ROC por valor de *k,* a qual contém informação das vinte e seis medidas que lhe estão associadas.

Os resultados obtidos são apresentados na Tabela 7. Um código para o cálculo das áreas pode ser consultado no Anexo A.15.

Desvio padrão		Energia		Entropia	
Área	*k*	Área	*k*	Área	*k*
0.7766106	1	0.6582633	1	0.6708683	1
0.6519608	2	0.6589636	2	0.6731443	2
0.6267507	3	0.6621148	3	0.6719188	3
0.6290266	4	0.6640406	4	0.6715686	4
0.6439076	5	0.6668417	5	0.6706933	5
0.6510854	6	0.6666667	6	0.6722689	6
0.6566877	7	0.6656162	7	0.6708683	7
0.6551120	8	0.6649160	8	0.6680672	8
0.6524860	9	0.6629902	9	0.5122549	9

Tabela 7 – Área das medidas correspondentes à média aritmética por cada escala *k*.

Em suma, é possível concluir, através da análise da Tabela 7, que a escala um para o desvio padrão é a melhor, sucedendo-lhe a escala dois da entropia. No entanto, e muito embora a melhor escala pertença ao desvio padrão, esta avaliação revela-se a pior das três, tendo os valores mais baixos para a área abaixo da curva ROC.

Continuando a explorar os dados, decidiu-se utilizar as médias das medidas que isoladamente obtiveram melhor *performance* em três situações distintas: quando a área das medidas é superior a 0.65, superior a 0.70 e superior a 0.73. Estas áreas foram calculadas com recurso ao código apresentado no Anexo A.16.

Medidas com área	Área agrupada	Nº de medidas
> 0,65	0.7330182	24
> 0,70	0.7972689	12
> 0,73	0.7766106	4

Tabela 8 – Área das medidas correspondentes à media das medidas iniciais com áreas superiores a 0.65, 0.70 e 0.73.

De acordo com os valores apresentados na Tabela 8, é vantajoso considerar em conjunto as 12 medidas iniciais cuja área abaixo das respetivas curvas ROC é superior a 0.7. A medida correspondente à média aritmética das medidas acima referidas pode proporcionar um melhor teste de diagnóstico.

Na Figura 17, encontram-se representadas as três curvas acima descritas. O código que originou este gráfico pode ser consultado no Anexo A.17.

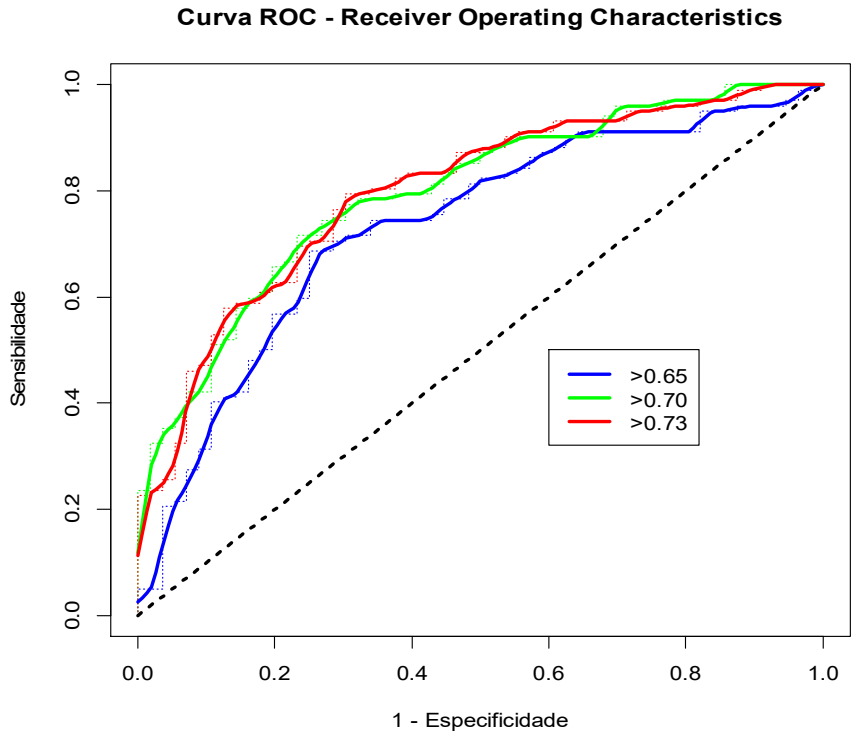


Figura 18 – Curvas das médias das medidas iniciais com áreas superiores a 0.65, 0.70 e 0.73.

Apresentou melhor área a média das medidas com área superior a 0.70, que inclui doze medidas. Note-se que corresponde ao caso intermédio, pois têm-se poucas medidas quando utilizadas unicamente as medidas que individualmente originam áreas superiores a 0.73, e estar-se-iam a utilizar muitas mais medidas no caso de se utilizarem todas as medidas que resultam em áreas superiores a 0.65.

4.1.2.2 Medidas agrupadas através da média ponderada (Adaboost)

Nesta secção, é aplicada uma metodologia que, através de um procedimento iterativo, procura obter a melhor média ponderada de todas as medidas, de modo a encontrar uma média com maior área possível.

O estudo da inteligência artificial e suas técnicas tem trazido grandes resultados para a evolução da tecnologia em diversas áreas. Técnicas como as redes neuronais ou árvores de decisão têm sido aprimoradas por técnicas de Boosting como o Adaptive Boosting (AdaBoost), introduzido pela primeira vez por Yoav Freund e Robert Schapire [33], com seu algoritmo Adaboost.

Esta técnica é uma das que apresenta maior tendência de crescimento devido ao seu potencial, flexibilidade e simplicidade para ser implementada em diferentes cenários. Existem diversas áreas que podem beneficiar do potencial desta técnica, tais como, o mercado de sensores e o tratamento de imagens para reconhecimento de padrões.

O Adaboost é um algoritmo de aprendizagem meta-heurístico e pode ser utilizado para melhorar o desempenho de outros algoritmos. Entende-se por algoritmo meta-heurístico aquele que se utiliza quando não se conhece um algoritmo eficiente, mas que utiliza a combinação de escolhas aleatórias e o conhecimento histórico dos resultados. Para a utilização do algoritmo Adaboost recorreu-se ao *package* adabag do *software* R [34], [35].

No caso particular desta dissertação, o Adaboost calcula a média ponderada das medidas selecionadas, ou seja, atribui pesos a cada medida conforme a importância que lhes dá. O algoritmo Adaboost começa por realizar um "treino" de aprendizagem de classificação, atribuindo ponderações aleatórias e determinando o correspondente desempenho. Posteriormente, com base na informação dos pontos com melhor desempenho, altera as ponderações atribuídas de forma a melhorar a sua *performance*. Deste modo, através do algoritmo Adaboost, às medidas que conduzem a um resultado final melhor são atribuídos pesos superiores, ao passo que se prejudicarem a *performance* do resultado final é-lhes atribuído um peso menor ou mesmo nulo.

Assim, a principal diferença entre a média ponderada e a média aritmética é que na média aritmética todas as medidas contribuem com igual peso, enquanto na média ponderada há a consideração dos pesos de cada termo, uma vez que existem medidas que contribuem mais que outras para atingir o objetivo proposto.

Na Tabela 9, estão apresentados os pesos atribuídos às medidas da escala um para o desvio padrão ao aplicar o algoritmo Adaboost. O código usado para obter estes resultados encontra-se no Anexo A.21.

k	*m*	Peso	*m*	Peso
1	1	3.409888	14	4.377233
	2	1.468545	15	0.000000
	3	2.894876	16	8.647012
	4	2.671707	17	0.000000
	5	2.843573	18	8.978511
	6	1.928388	19	4.318373
	7	4.362647	20	1.598112
	8	0.000000	21	0.000000
	9	0.000000	22	26.932936
	10	2.860595	23	3.229778
	11	8.400530	24	4.429977
	12	0.000000	25	0.000000
	13	6.647317	26	0.000000

Tabela 9 – Pesos atribuídos às medidas da escala um para o desvio padrão, em percentagem.

Após a determinação dos pesos, foi calcular-se a média ponderada por escala. A determinação das medidas correspondentes às médias ponderadas permitiu calcular as áreas abaixo das curvas ROC, apresentadas na Tabela10 e resultantes do código presente no Anexo A.22.

Desvio padrão		Energia		Entropia	
Área	*k*	Área	*k*	Área	*k*
0.7939426	1	0.7713585	1	0.8112745	1
0.6682423	2	0.7025560	2	0.6710434	2
0.6225490	3	0.7025560	3	0.6139706	3
0.6211485	4	0.6554622	4	0.6220238	4
0.6605392	5	0.6790966	5	0.6628151	5
0.6703431	6	0.6808473	6	0.6670168	6
0.6587885	7	0.6649160	7	0.6579132	7
0.6141457	8	0.6281513	8	0.6153711	8
0.6101190	9	0.6285014	9	0.6153711	9

Tabela 10 – Área abaixo das curvas ROC das medidas correspondentes às médias ponderadas por *k*.

Em suma, as escalas um são as melhores para todas as avaliações e de forma destacada comparativamente com as restantes escalas. Com esta metodologia, a avaliação sob a energia é a que apresenta melhor *performance* na deteção do padrão reticular, pois é a que genericamente tem melhores valores para a área abaixo da curva ROC.

Comparando as Tabelas 7 e 10, é possível constatar que os resultados das áreas das médias ponderadas são na sua maioria superiores aos das médias aritméticas. Este resultado faz sentido, visto que as medidas com pesos pouco ou nada revelantes para o estudo deixam de influenciar os resultados obtidos.

4.1.2.3 Agrupamento de escalas pelo método Adaboost

Após a apresentação e exposição dos resultados das áreas abaixo da curva ROC das médias, por *k*, optou-se por considerar todas as 702 medidas e aplicar o método Adaboost.

Recordando os conceitos apresentados no capítulo 2.3, vão calcular-se alguns dos conceitos básicos tais como acurácia, sensibilidade e especificidade, valores preditivos negativo e positivos, razão de verosimilhança positiva e negativa, considerando todas as medidas agrupadas e alguns agrupamentos de escalas.

Nas tabelas seguintes, Tabela 11 a 14, estão expostos os VP, VN, FP e FN, necessários para obter os valores dos conceitos acima descritos.

Na Tabela 11, aplicou-se o método Adaboost para todas as medidas das três avaliações e o valor da área é de 0.76. De notar que este valor é inferior aos obtidos na escala um para as três avaliações (cf. Tabela 10).

	Teste +	Teste -	Total
Doente (D)	95	7	102
Saudável (S)	4	52	56
Total	99	59	158

Tabela 11 – Agrupamento das três avaliações.

O algoritmo Adaboost além das ponderações associadas à melhor curva ROC, determina ainda o ponto ótimo. Na Tabela 11, são apresentados os resultados referentes à classificação com base nesse ponto. Da Tabela 11 podem extrair-se os seguintes resultados: a acurácia é 0.93, portanto, 93% da população obtém o resultado correto; a sensibilidade é 0.93, logo, 93% dos indivíduos doentes obtém resultado verdadeiro; a especificidade é 0.93. Logo, 93%

dos indivíduos saudáveis obtém resultado verdadeiro; o VPP é 0.96, ou seja, 96% dos indivíduos com resultado positivo estão realmente doentes; o VPN é 0.88. Ou seja, 88% dos indivíduos com resultado negativo estão realmente saudáveis; a RVP é 13.3, assim, um verdadeiro positivo é 13.3 vezes mais provável que um falso positivo e a RVN é 0.08. Assim, um falso negativo é 92 por cento menos provável que um verdadeiro negativo.

De seguida, é apresentada a Tabela 12 onde se aplicou o método Adaboost à avaliação efetuada sob o desvio padrão, sendo o valor da área para esta avaliação de 0.75, considerando a curva gerada pela média ponderada das respetivas 234 medidas.

	Teste +	Teste -	Total
Doente (D)	96	6	102
Saudável (S)	10	46	56
Total	106	52	158

Tabela 12 – Agrupamento das escalas da avaliação desvio padrão.

Da Tabela 12 podem extrair-se os seguintes resultados: a acurácia é 0.90, portanto, 90% da população obtém o resultado correto; a sensibilidade é 0.94, logo, 94% dos indivíduos doentes obtém resultado verdadeiro; a especificidade é 0.82, logo, 82% dos indivíduos saudáveis obtém resultado verdadeiro; o VPP é 0.91, ou seja, 91% dos indivíduos com resultado positivo estão realmente doentes; o VPN é 0.88, ou seja, 88% dos indivíduos com resultado negativo estão realmente saudáveis; a RVP é 5.2, assim, um verdadeiro positivo é 5.2 vezes mais provável que um falso positivo e a RVN é 0.20, assim, um falso negativo é 80 por cento menos provável que um verdadeiro negativo.

O método Adaboost foi também aplicado à avaliação energia. O valor da área sob a curva ROC correspondente à média ponderada das 234 medidas desta avaliação é de 0.75.

	Teste +	Teste -	Total
Doente (D)	95	7	102
Saudável (S)	7	49	56
Total	102	56	158

Tabela 13 – Agrupamento das escalas da avaliação energia.

Da Tabela 13 podem extrair-se os seguintes resultados: a acurácia é 0.91, portanto, 91% da população obtém o resultado correto; a sensibilidade é 0.93, logo, 93% dos indivíduos doentes obtém resultado verdadeiro; a especificidade é 0.88, logo, 88% dos indivíduos saudáveis obtém resultado verdadeiro; o VPP é 0.93, ou seja, 93% dos indivíduos com resultado positivo estão realmente doentes; o VPN é 0.88, ou seja, 88% dos indivíduos com resultado negativo estão realmente saudáveis; a RVP é 7.8, assim, um verdadeiro positivo é 7.8 vezes mais provável que um falso positivo e a RVN é 0.13, assim, um falso negativo é 87 por cento menos provável que um verdadeiro negativo.

Na Tabela 14, apresentam-se os resultados da aplicação do método de Adaboost à avaliação efetuada sob a entropia. Para esta avaliação obteve-se uma média ponderada com uma área sob a curva ROC de 0.76.

	Teste +	Teste -	Total
Doente (D)	97	5	102
Saudável (S)	12	44	56
Total	109	49	158

Tabela 14 – Agrupamento das escalas da avaliação entropia.

Da Tabela 14 podem extrair-se os seguintes resultados: a acurácia é 0.89, portanto, 89% da população obtém o resultado correto; a sensibilidade é 0.95, logo, 95% dos indivíduos doentes obtém resultado verdadeiro; a especificidade é 0.79, logo, 79% dos indivíduos saudáveis obtém resultado verdadeiro; o VPP é 0.89, ou seja, 89% dos indivíduos com resultado positivo estão realmente doentes; o VPN é 0.90, ou seja, 90% dos indivíduos com resultado negativo estão realmente saudáveis; a RVP é 13.6, assim, um verdadeiro positivo é 13.6 vezes mais provável que um falso positivo e a RVN é 0.07, assim, um falso negativo é 93 por cento menos provável que um verdadeiro negativo.

Pode, portanto, concluir-se que os melhores resultados encontram-se quando se usam todas as escalas das três avaliações: desvio padrão, energia e entropia. É claramente melhor a identificar indivíduos saudáveis, tendo um desempenho praticamente idêntico ao dos restantes casos analisados relativamente à classificação de indivíduos doentes. Dentro das três avalições todas as escalas apresentam resultados bastantes bons na deteção do padrão

reticular. Ainda assim, é possível verificar que a avaliação através da energia é a que, em geral, traduz piores resultados. Por outro lado, a entropia é a avaliação que gera uma média ponderada das suas medidas que dá origem ao maior valor da área sob a curva ROC.

4.2. Procura do "melhor" ponto de corte

Enquanto na secção anterior foi dada ênfase à procura da curva correspondente à melhor medida (isolada ou agrupada) para o diagnóstico, nesta secção, vão ser utilizadas duas metodologias que visam determinar o ponto de corte ótimo em cada diagnóstico, nomeadamente através da maximização da soma da sensibilidade e da especificidade e através da minimização da distância ao ponto ideal (0,1).

4.2.1 Maximização da soma da sensibilidade e especificidade

Esta metodologia consiste em determinar o ponto ótimo a partir da soma da sensibilidade e especificidade para cada ponto em cada curva ROC. Quanto maior o valor desta soma, tendencialmente, maior será a proximidade ao ponto (0,1). Como foi visto no capítulo 2.4.2, quanto mais a curva se aproximar do canto superior esquerdo, maior será o seu nível de discriminação e, consequentemente, maior será a soma da sensibilidade com a especificidade, sendo que, o ideal seria que fossem ambas iguais a 100%. Note-se que este método corresponde a procurar o valor que maximiza o índice de Youden (que corresponde à sensibilidade mais a especificidade menos 1), conforme [20], [36], [37].

4.2.1.1 Dados não agrupados

Em primeiro lugar, foi aplicada esta metodologia aos dados não agrupados. Para cada avaliação, dentro de cada escala *k*, procurou-se a curva com o ponto cuja soma da sensibilidade e da especificidade é máxima. Os valores da soma e das respetivas sensibilidade e especificidade foram registados, tal como a área sob a curva ROC e a medida correspondente. Os resultados obtidos são apresentados nas Tabelas 15, 16 e 17. O código para o cálculo desta soma pode ser consultado no Anexo A.19.

k	Soma	Sensibilidade	Especificidade	*m*	Área
1	1.454832	0.7941176	0.6607143	11	0.7305672
2	1.380602	0.8627451	0.5178571	31	0.6885504
3	1.280112	0.6372549	0.6428571	76	0.6164216
4	1.332633	0.7254902	0.6071429	96	0.6286765
5	1.344188	0.7549020	0.5892857	125	0.6650910
6	1.364846	0.6862745	0.6785714	156	0.6675420
7	1.345238	0.6666667	0.6785714	162	0.6437325
8	1.294118	0.7941176	0.5000000	191	0.6190476
9	1.295868	0.8137255	0.4821429	227	0.6339286

Tabela 15 – Soma da sensibilidade e especificidade dos pontos ótimos por escala utilizando como avaliação o desvio padrão.

Considerando a Tabela 15, é possível constatar que os melhores pontos de corte das escalas dois e nove têm uma sensibilidade de 0.86 e 0.81, respetivamente, contudo, a especificidade é, aproximadamente, 0.5. Isto significa que são testes de diagnóstico com elevada *performance,* caso o objetivo seja singularizar os diagnósticos de doença em detrimento dos saudáveis. Também a escala oito apresenta um ponto com um valor de sensibilidade elevado, próximo de 0.80, no entanto, volta a acontecer o mesmo cenário do caso anterior, a especificidade traduz uma baixa capacidade de poder discriminativo na distinção entre indivíduos doentes e saudáveis. Já a escala um apresenta maior soma (aproximadamente 1.45), correspondente a um ponto com valor de sensibilidade relativamente alto e valor de especificidade melhor que as escalas anteriores que têm melhor sensibilidade que esta, 0.66. Considerando a escala um, tem-se que, aproximadamente, 80% dos indivíduos doentes é corretamente classificado e 66% dos indivíduos saudáveis é corretamente classificado.

Note-se que os valores da sensibilidade são, na sua maioria (sete em nove), relativamente superiores aos da especificidade, o que significa que com a metodologia do ponto ótimo, usando a soma da sensibilidade com a especificidade, o desvio padrão é, genericamente, melhor a identificar verdadeiros positivos do que verdadeiros negativos.

De ressalvar que nos períodos anteriores foi feita uma análise detalhada aos valores de sensibilidade e especificidade, pois está-se perante uma análise agrupada de vinte e seis medidas por um tipo de avaliação, o desvio padrão, logo, apenas se têm nove resultados distintos. No entanto, a soma tornar-se-ia extremamente útil no caso de se terem dezenas, centenas ou milhares de escalas. Nestes casos, a seleção inicial para análise de resultados

poderia ser feita pelas somas mais elevadas, ou seja, pelas somas mais próximas de 2. Assim, para esta tabela, ter-se-iam as escalas números um, dois e seis, por ordem decrescente. Por fim, pode concluir-se que a melhor escala é a um e, dentro desta escala, a medida associada à coluna 11, pois quer a soma quer a área assumem valores bem superiores aos valores apresentados para as restantes escalas.

k	Soma	Sensibilidade	Especificidade	*m*	Área
1	1.373599	0.7843137	0.5892857	257	0.7032563
2	1.294468	0.5980392	0.6964286	282	0.6202731
3	1.306723	0.7352941	0.5714286	312	0.6412815
4	1.320378	0.5882353	0.7321429	315	0.6435574
5	1.382703	0.6862745	0.6964286	341	0.6857493
6	1.402311	0.7058824	0.6964286	387	0.6782213
7	1.380952	0.6666667	0.7142857	402	0.6657913
8	1.302171	0.7843137	0.5178571	425	0.6327031
9	1.313025	0.7058824	0.6071429	457	0.6278011

Tabela 16 - Soma da sensibilidade e especificidade dos pontos ótimos por escala utilizando como avaliação a energia.

Considerando a Tabela 16, é possível comprovar que todos os valores das somas dos pontos da sensibilidade e especificidade situam-se no intervalo]1.29, 1.41[, o que se traduz numa análise dúbia caso se pretenda escolher a ou as melhores escalas para a avaliação efetuada sob a energia. Não significa que não haja valores de sensibilidade ou especificidade elevados e que alguma escala não se repercuta num bom teste de diagnóstico, no entanto, esta análise terá de ser feita para cada escala visto que pela soma não se podem tirar grandes conclusões. Ainda assim, mesmo que pouco díspares, os maiores valores da soma encontram-se nas escalas seis, cinco e sete, por ordem decrescente.

Note-se que os valores da sensibilidade e da especificidade são, de um modo geral, relativamente próximos, o que indica que as medidas são tão boas a diagnosticar verdadeiros positivos como verdadeiros negativos. Como vantagem, tem-se o facto de diagnosticar sempre corretamente uma elevada percentagem da população. Como desvantagem, há a certeza em como existe a ocorrência quer de falsos positivos quer de falsos negativos.

Cinco das nove escalas apresentam valores de sensibilidade que superam os da especificidade, o que conduz a um teste de diagnóstico que dá primazia aos diagnósticos corretos de verdadeiros doentes. Ou seja, tal como na Tabela 16, é maior a capacidade em identificar corretamente um indivíduo doente do que um saudável.

Em suma, a avaliação sob o ponto de vista da energia é melhor a identificar verdadeiros positivos do que verdadeiros negativos. Contudo, é nesta que se encontram os cinco mais elevados valores de especificidade entre os três tipos de avaliação,]0.69, 0.74[. Estes valores encontram-se nas escalas dois, quatro, cinco, seis e sete que são, então, as melhores medidas a identificar corretamente indivíduos saudáveis. Por fim, pode concluir-se que, nesta avaliação, não há uma escala que se destaque em relação às restantes, pois as escalas um, cinco, seis e sete têm comportamentos similares onde, apesar da escala um apresentar maior área, as escalas cinco, seis e sete obtiveram maior valor na soma da sensibilidade e da especificidade.

k	Soma	Sensibilidade	Especificidade	*m*	Área
1	1.436975	0.7941176	0.6428571	492	0.7526261
2	1.372549	0.8725490	0.5000000	499	0.6901261
3	1.289916	0.6470588	0.6428571	544	0.6162465
4	1.322829	0.7156863	0.6071429	564	0.6283263
5	1.344188	0.7549020	0.5892857	593	0.6654412
6	1.364846	0.6862745	0.6785714	624	0.6670168
7	1.345238	0.6666667	0.6785714	630	0.6432073
8	1.294118	0.7941176	0.5000000	659	0.6192227
9	1.295868	0.8137255	0.4821429	695	0.6342787

Tabela 17 – Soma da sensibilidade e especificidade dos pontos ótimos por escala utilizando como avaliação a entropia.

Observando a Tabela 17, é possível verificar que é a primeira escala a destacar-se no que refere ao valor da soma com 1.44. Seguidamente, os maiores valores pertencem à escala dois, seis e sete com 1.37, 1.36 e 1.35, respetivamente. As restantes rondam um intervalo relativamente próximo,]1.29, 1.35[. De qualquer modo, os valores das somas mais elevadas podem ser considerados baixos atendendo a que são valores distantes de 2 que é o valor máximo que a soma pode tomar quando a sensibilidade e especificidade são ambas ótimas.

As escalas dois, nove, um e oito são as que apresentam maiores valores para a sensibilidade, por ordem decrescente, com os valores 0.87, 0.81, 0.79 e 0.79, respetivamente. Também as escalas quatro e cinco apresentam valores elevados para a sensibilidade, acima dos 0.71.

As escalas dois, oito e nove têm, valores de especificidade de aproximadamente 0.5, valores diminutos relativamente ao pretendido.

Assim, usando o máximo da soma da sensibilidade com a especificidade, a entropia é, genericamente, melhor a identificar verdadeiros positivos do que verdadeiros negativos. Por fim, pode concluir-se que a melhor escala é a um, pois é a que atinge maior soma bem como maior área.

Comparando, agora, estas três tabelas (Tabela 14, 15 e 16) conclui-se que as avaliações efetuadas sob o desvio padrão e entropia são as que apresentam maior número de semelhanças, pois os três valores com maior soma e área encontram-se nas escalas um, dois e seis.

4.2.1.2 Dados agrupados: Adaboost

Este método de pesquisa pelo melhor ponto foi também aplicado para as curvas resultantes das médias ponderadas determinadas pelo método Adaboost.

Para cada avaliação e para cada uma das escalas (k) aplicou-se o algoritmo Adaboost de forma a determinar a melhor média ponderada das 26 medidas associadas. Da curva ROC correspondente, determinou-se o ponto cuja soma da sensibilidade com a especificidade era máxima e calculou-se a área abaixo das respetivas curvas.

Os resultados obtidos são apresentados nas Tabelas 18, 19 e 20 para o desvio padrão, energia e entropia, respetivamente. O código para o cálculo desta soma pode ser consultado no Anexo A.23.

k	Soma	Sensibilidade	Especificidade	Área
1	1.472689	0.7941176	0.6785714	0.7928922
2	1.246499	0.4607843	0.7857143	0.6701681
3	1.206583	0.8137255	0.3928571	0.6200980
4	1.220938	0.7745098	0.4464286	0.6108193
5	1.330882	0.7058824	0.6250000	0.6579132
6	1.297269	0.5294118	0.7678571	0.6629902
7	1.303221	0.6960784	0.6071429	0.6542367
8	1.231092	0.5882353	0.6428571	0.6137955
9	1.226541	0.6372549	0.5892857	0.6188725

Tabela 18 - Soma da sensibilidade e da especificidade no ponto ótimo, utilizando o desvio padrão como avaliação e o algoritmo AdaBoost.

Recordando que a sensibilidade é a capacidade de um teste identificar corretamente um indivíduo doente e a especificidade é a capacidade de um teste identificar corretamente um indivíduo saudável, o teste de diagnóstico ideal é aquele cujos valores se aproximem o mais possível da unidade (100%).

Observando a Tabela 18, é possível constatar que os maiores valores da soma da sensibilidade com a especificidade estão presentes nas escala um e cinco com 1.47 e 1.33 e na escala número sete com 1.30. As restantes seis escalas oscilam no intervalo]1.20, 1.30[. As três escalas com maior soma apresentam, respetivamente, capacidade de, aproximadamente, 79%, 71% e 70% para diagnosticar corretamente indivíduos doentes e 68%, 63% e 61% de capacidade para diagnosticar corretamente indivíduos saudáveis.

Os três maiores valores de área encontram-se nas escalas um, dois e seis onde, curiosamente, só a escala um apresenta maior valor de soma ao mesmo tempo que maior valor de área abaixo da curva. Para a escala dois, 46 em cada 100 indivíduos doentes são corretamente classificados e 79 em cada 100 indivíduos saudáveis são corretamente classificados como tal. Já para a escala seis, 53 em cada 100 indivíduos doentes são corretamente classificados e 77 em cada 100 indivíduos saudáveis são corretamente classificados.

Recorde-se que a área analisa todos os valores possíveis para o ponto de corte, avaliando assim a fiabilidade do diagnóstico em todos estes pontos, enquanto a soma unicamente analisa a fiabilidade num único ponto (o que maximiza a soma).

Em suma, e depois de analisar pormenorizadamente os resultados da tabela pode concluir-se que a melhor escala é a um.

k	Soma	Sensibilidade	Especificidade	Área
1	1.515756	0.6764706	0.8392857	0.7949930
2	1.278011	0.8137255	0.4642857	0.6733193
3	1.362745	0.8627451	0.5000000	0.6824230
4	1.264356	0.4607843	0.8035714	0.6400560
5	1.386204	0.7254902	0.6607143	0.6839986
6	1.392507	0.6960784	0.6964286	0.6803221
7	1.374650	0.6960784	0.6785714	0.6717437
8	1.279062	0.7254902	0.5535714	0.6265756
9	1.292367	0.7745098	0.5178571	0.6323529

Tabela 19 - Soma da sensibilidade e da especificidade no ponto ótimo, utilizando a energia como avaliação e o algoritmo AdaBoost.

Analisando a Tabela 19, é possível comprovar que os maiores valores da soma da sensibilidade com a especificidade encontram-se nas escala um, seis e cinco com 1.52, 1.39 e 1.39. As restantes situam-se no intervalo]1.26, 1.38[. As três escalas com maior soma apresentam, respetivamente, capacidade de, aproximadamente, 68%, 70% e 73%, respetivamente, para diagnosticar corretamente indivíduos doentes e 84%, 70% e 66% de capacidade para diagnosticar corretamente indivíduos saudáveis.

Contudo, analisando detalhadamente os resultados obtidos, é possível constatar que existem valores tanto de sensibilidade como de especificidade maiores que os pertencentes às escalas com maiores valores da soma. Assim, destacam-se os valores 0.86 e 0.81 referentes às escalas três e dois para a sensibilidade, apesar de depois apresentarem valores baixos, 0.5 e 0.46, na especificidade, e 0.80 para a especificidade relativo à escala quatro.

Assim, também para esta tabela a escala que apresenta concomitantemente maiores valores de soma e área abaixo da curva ROC é a escala um, por isso, pode concluir-se que esta é a melhor escala desta avaliação, tendo em consideração esta metodologia.

k	Soma	Sensibilidade	Especificidade	Área
1	1.519258	0.7156863	0.8035714	0.8205532
2	1.280462	0.4411765	0.8392857	0.6789216
3	1.206933	0.6176471	0.5892857	0.6113445
4	1.214636	0.8039216	0.4107143	0.6130952
5	1.324580	0.7352941	0.5892857	0.6552871
6	1.303221	0.6960784	0.6071429	0.6619398
7	1.315826	0.6372549	0.6785714	0.6559874
8	1.256653	0.7745098	0.4821429	0.6158964
9	1.218487	0.6470588	0.5714286	0.6169468

Tabela 20 – Soma da sensibilidade e da especificidade no ponto ótimo, utilizando a entropia como avaliação e o algoritmo AdaBoost.

Considerando a Tabela 20, é possível verificar que os maiores valores da soma da sensibilidade com a especificidade estão presentes nas escalas um, cinco e sete com 1.52, 1.32 e 1.32. As restantes situam-se no intervalo]1.20, 1.31[. As três escalas com maior soma apresentam, respetivamente, capacidade de, aproximadamente, 72%, 74% e 64% para diagnosticar corretamente indivíduos doentes e 80%, 59% e 68% de capacidade para diagnosticar corretamente indivíduos saudáveis.

Porém, existem escalas que obtêm isoladamente valores de sensibilidade ou de especificidade melhores que os acima descritos. Assim, destacam-se os valores 0.80 e 0.77 referentes às escalas quatro e oito para a sensibilidade, as quais têm associados valores de especificidade de 0.41 e 0.77. Salienta-se ainda o ponto obtido para a escala 2 com 0.84 para a especificidade e 0.44 na sensibilidade.

Por fim, e mais uma vez, para esta tabela a escala que apresenta concomitantemente maiores valores de soma e área abaixo da curva ROC é a escala um, por isso pode ser selecionada como a melhor escala desta avaliação, considerando a soma da sensibilidade e da especificidade e o algoritmo Adaboost.

Comparando, agora, estas três tabelas (Tabela 18, 19 e 20) conclui-se que as avaliações efetuadas sob o desvio padrão e entropia são as que apresentam maior número de semelhanças, pois dois dos três valores com maior soma encontram-se nas escalas um e sete e os três valores com maior área encontram-se nas escalas um, dois e seis.

Em comum as três tabelas têm a escala um destacadamente com maior soma, única com valores superiores a 1.4 (1.47, 1.52 e 1.52), e maior área (única com áreas superiores a 0.7, verificando-se 0.79, 0.79 e 0.82).

Na Tabela 21, são expostos os resultados da soma, sensibilidade, especificidade e área por agrupamento de escalas.

	Soma	Sensibilidade	Especificidade	Área
Desvio padrão	1.403361	0.6176471	0.7857143	0.7405462
Energia	1.390756	0.6764706	0.7142857	0.7177871
Entropia	1.405812	0.7450980	0.6607143	0.7585784
Todas	1.485294	0.7352941	0.7500000	0.7864146

Tabela 21 – Soma da sensibilidade e da especificidade no ponto ótimo e o algoritmo AdaBoost.

Destacam-se os valores da soma e da área considerando as três avaliações juntas, a sensibilidade para a entropia, a especificidade para o desvio padrão.

Notemos que com a aplicação do algoritmo Adaboost a todas as 234 medidas de cada avaliação, os resultados não são melhores que os previamente obtidos utilizando unicamente a primeira escala, como se pode concluir da comparação dos valores apresentados na Tabela 21 com os apresentados nas Tabelas 18, 19 e 20. Na aplicação do algoritmo Adaboost às 702 medidas, o resultado é melhor que os obtidos dentro de cada avaliação, mas, mesmo assim, ainda inferior aos valores obtidos na primeira escala da energia e da entropia.

4.2.2 Minimização da distância ao ponto ótimo

Conforme referido previamente, quanto mais próximo do ponto ótimo, no qual não existe erros de classificação, melhor será o desempenho do teste diagnóstico. Deste modo, uma forma intuitiva de escolher o "melhor" ponto de corte é optar pelo ponto que tem distância mínima relativamente ao ponto ótimo, cf. [36], [37]. Deste modo, esta metodologia assemelha-se bastante à anterior no que toca ao princípio base, pois quanto menor for a distância ao ponto (0,1) mais a curva se aproxima do canto superior esquerdo, maior será o seu nível de discriminação e, tendencialmente, maior será a soma da sensibilidade com a especificidade.

4.2.2.1 Dados não agrupados

Numa primeira análise, foi calculado o ponto com distância mínima ao ponto ideal (0,1) considerando todas as medidas para cada escala. Posteriormente, selecionou-se para cada escala, a curva que contém o ponto com menor distância.

Os resultados encontram-se nas Tabela 18, 19 e 20. O código para o cálculo desta soma pode ser consultado no Anexo A.20.

k	Distância	Sensibilidade	Especificidade	*m*	Área
1	0.3901649	0.7549020	0.6964286	11	0.7305672
2	0.4685425	0.7745098	0.5892857	31	0.6885504
3	0.5090531	0.6372549	0.6428571	76	0.6164216
4	0.4792623	0.7254902	0.6071429	96	0.6286765
5	0.4773718	0.6470588	0.6785714	124	0.6652661
6	0.4491548	0.6862745	0.6785714	156	0.6675420
7	0.4630631	0.6666667	0.6785714	162	0.6437325
8	0.5292757	0.7156863	0.5535714	192	0.6235994
9	0.5160851	0.6274510	0.6428571	226	0.6257003

Tabela 22 – Distância minima ao ponto (0,1) considerando a avaliação desvio padrão.

Considerando os resultados apresentados na Tabela 22, os menores valores de distância encontram-se nas escalas um, seis e sete por ordem crescente. Estas escalas são também as que apresentam maior área, isto significa que traduzem as curvas mais próximas do canto superior esquerdo para a avaliação efetuada sobre o desvio padrão.

Analisando-se as escalas um, seis e sete pode comprovar-se que para a escala um, com distância 0.39, 75% dos indivíduos doentes são corretamente classificados e, aproximadamente, 70% dos indivíduos saudáveis são corretamente classificados. Para a escala seis, com distância 0.45, 69% dos indivíduos doentes são corretamente classificados e, aproximadamente, 68% dos indivíduos saudáveis são corretamente classificados. Para a escala sete, com distância 0.46, aproximadamente, 67% dos indivíduos doentes são corretamente classificados e, aproximadamente, 68% dos indivíduos saudáveis são corretamente classificados.

Os valores da sensibilidade são, na sua maioria (cinco em nove), relativamente superiores aos da especificidade, o que significa que com a metodologia da distância ao ponto (0,1), a avaliação sob o desvio padrão é, genericamente, melhor a identificar doentes do que saudáveis.

Por fim, pode concluir-se que a melhor escala é a um, uma vez que apresenta uma menor distância ao ponto ótimo bem como uma maior área sob a curva ROC associada.

k	Distância	Sensibilidade	Especificidade	*m*	Área
1	0.4621022	0.6862745	0.6607143	257	0.7032563
2	0.5037143	0.5980392	0.6964286	282	0.6202731
3	0.5037288	0.7352941	0.5714286	312	0.6412815
4	0.4912205	0.5882353	0.7321429	315	0.6435574
5	0.4365539	0.6862745	0.6964286	341	0.6857493
6	0.4226829	0.7058824	0.6964286	387	0.6782213
7	0.4390259	0.6666667	0.7142857	402	0.6657913
8	0.4966878	0.6372549	0.6607143	427	0.6491597
9	0.4907565	0.7058824	0.6071429	457	0.6278011

Tabela 23 – Distância minima ao ponto (0,1) considerando a avaliação energia.

Observando a Tabela 23, é possível verificar que todas as distâncias ao ponto (0, 1) estão contidas no intervalo]0.42, 0.51[, o que se repercute numa análise ambígua caso se pretenda escolher uma ou mais escalas para efetuar a avaliação sob a energia. Não obstante, analisando detalhadamente, encontram-se valores de sensibilidade e especificidade relativamente bons.

As escalas seis, cinco e sete são as que apresentam menor distância ao ponto (0, 1). A escala seis com 0.42 apresenta a capacidade de diagnosticar corretamente indivíduos doentes em 71% e, aproximadamente, 70% em diagnosticar corretamente indivíduos saudáveis. As escalas cinco e sete com, aproximadamente, 0.44 de distância ao ponto (0, 1), apresentam a capacidade de diagnosticar corretamente indivíduos doentes em, aproximadamente 69% e 67%, respetivamente, e 70% e 71% em diagnosticar corretamente indivíduos saudáveis, respetivamente.

Esta avaliação traduz valores muito competitivos entre sensibilidade e especificidade. Não há nenhum padrão ou valores matematicamente elevados que façam sobressair a energia como forma de avaliação mais apta a diagnosticar corretamente indivíduos doentes em detrimento de diagnosticar corretamente indivíduos saudáveis, ou vice-versa.

Os três maiores valores de sensibilidade encontram-se nas escalas três, seis, nove e um com 74%, aproximadamente, 71% e 69%, respetivamente, de capacidade em diagnosticar corretamente os indivíduos doentes. Já os dois maiores valores de especificidade encontram-se nas escalas quatro e sete (73% e 71%), bem como as escalas dois, cinco e seis com 67% de capacidade em diagnosticar corretamente os indivíduos saudáveis. Por fim, pode concluir-se que a escala seis apresenta o ponto de corte mais próximo do ponto ótimo, mas a escala um obtém a maior área sob a curva ROC.

Em suma, pode concluir-se que este teste, sob a metodologia da menor distância ao ponto (0, 1) e sob a avaliação da energia é, genericamente, tão bom a identificar doentes como saudáveis, pois os maiores valores de sensibilidade e especificidade são bastante próximos.

k	Distância	Sensibilidade	Especificidade	*m*	Área
1	0.4027714	0.7352941	0.6964286	479	0.7223389
2	0.4683073	0.7450980	0.6071429	499	0.6901261
3	0.5021140	0.6470588	0.6428571	544	0.6162465
4	0.4849444	0.7156863	0.6071429	564	0.6283263
5	0.4773718	0.6470588	0.6785714	592	0.6657913
6	0.4491548	0.6862745	0.6785714	624	0.6670168
7	0.4630631	0.6666667	0.6785714	630	0.6432073
8	0.5292757	0.7156863	0.5535714	660	0.6234244
9	0.5160851	0.6274510	0.6428571	694	0.6255252

Tabela 24 – Distância minima ao ponto (0,1) considerando a avaliação entropia.

Atendendo aos valores patentes na Tabela 24, pode observar-se que as distâncias ao ponto (0,1) estão contidas no intervalo]0.40, 0.53[e que apesar de intuitivamente se escolher a escala com menor distância, 0.40, é necessário analisar os valores da sensibilidade e da especificidade para cada escala, pois pode haver uma outra que seja mais adequada como teste diagnóstico, por ter um valor de sensibilidade ou de especificidade mais ou menos relevante para a situação que se pretenda avaliar.

Os três menores valores de distância ao ponto (0,1) pertencem às escalas um, seis e sete com 0.40, 0.45 e 0.46, respetivamente. A escala número um tem, aproximadamente, 74% de capacidade em diagnosticar corretamente um indivíduo doente e 70% de capacidade em diagnosticar corretamente um indivíduo saudável. A escala seis tem, aproximadamente, 69% de capacidade em diagnosticar corretamente um indivíduo doente e 68% de capacidade em diagnosticar corretamente um indivíduo saudável. A escala número sete tem, aproximadamente, 67% de capacidade em diagnosticar corretamente um indivíduo doente e 68% de capacidade em diagnosticar corretamente um indivíduo saudável.

A escala um é a que apresenta maior valor de sensibilidade e maior valor de especificidade, portanto, a escolher alguma escala para vir a servir de teste de diagnóstico deveria ser esta.

Para terminar, pode concluir-se que este teste, sob a metodologia da menor distância ao ponto (0,1) e sob a avaliação da entropia é, genericamente, melhor a identificar doentes do que saudáveis, pois seis em nove valores da sensibilidade são superiores aos da especificidade.

Comparando agora estas três tabelas confirma-se, uma vez mais, que as avaliações efetuadas sob o desvio padrão e entropia são as que apresentam maior número de semelhanças: os três valores com menores distâncias ao ponto ótimo (0,1) encontram-se nas escalas um, seis e sete e os três valores com maior área encontram-se nas escalas um, dois e seis.

Considerando esta metodologia com os dados não agrupados, a escala um é, uma vez mais, identificada como a mais adequada, por ter menor distância, bem como maior área associada, quer para o desvio padrão quer para a entropia, com única exceção na avaliação energia na qual a escala seis apresenta um ponto mais próximo do ponto ótimo.

4.2.2.2 Dados agrupados: Adaboost

Para cada avaliação e para cada uma das escalas (k) aplicou-se o algoritmo Adaboost às 26 medidas associadas de forma a determinar a melhor média ponderada e, da curva ROC obtida com essa média ponderada, determinou-se o ponto cuja distância ao ponto (0,1) é menor.

Os resultados encontram-se nas Tabelas 12 (desvio padrão), 13 (energia) e 14 (entropia). O código para o cálculo desta distância pode ser consultado no Anexo A.24.

k	Distância	Sensibilidade	Especificidade	Área
1	0.3501332	0.7745098	0.7321429	0.7867647
2	0.5346061	0.7058824	0.5535714	0.6766457
3	0.5790179	0.5588235	0.6250000	0.6190476
4	0.5751820	0.7156863	0.5000000	0.6125700
5	0.4940059	0.7254902	0.5892857	0.6551120
6	0.4907565	0.7058824	0.6071429	0.6640406
7	0.4765818	0.7058824	0.6250000	0.6682423
8	0.5600297	0.5686275	0.6428571	0.6122199
9	0.5232063	0.6176471	0.6428571	0.6172969

Tabela 25 – Distância mínima ao ponto (0,1) utilizando o desvio padrão como avaliação e o algoritmo AdaBoost.

Observando a tabela 25, é possível constatar que os menores valores de distância ao ponto ótimo estão presentes nas escala um e sete com 0.35 e 0.48 e na escala número seis com 0.49. As restantes seis escalas oscilam no intervalo]0.49, 0.58[. As três escalas com menor distância apresentam, respetivamente, capacidade de, aproximadamente, 77%, 71% e 71% para diagnosticar corretamente indivíduos doentes e 73%, 63% e 61% de capacidade para diagnosticar corretamente indivíduos saudáveis.

Olhando agora para os resultados obtidos dos valores da área abaixo da curva ROC, tem-se que os três maiores valores de área encontram-se nas escalas um, dois e sete, onde as escalas um e sete apresentam menores valores de distância ao mesmo tempo que maiores valores de área abaixo da curva. Para a escala dois, 71 em cada 100 indivíduos doentes são corretamente classificados e 55 em cada 100 indivíduos saudáveis são corretamente classificados.

Em suma, e depois de analisar detalhadamente os resultados da tabela, pode concluir-se que a melhor escala é a um, tendo uma distância ao ponto ótimo significativamente inferior à das restantes escalas, bem como uma área claramente superior.

k	Distância	Sensibilidade	Especificidade	Área
1	0.4306723	0.8235294	0.6071429	0.7615546
2	0.5618172	0.5392157	0.6785714	0.6383053
3	0.4737150	0.7352941	0.6071429	0.6850490
4	0.5678028	0.4901961	0.7500000	0.6286765
5	0.4356851	0.7058824	0.6785714	0.6787465
6	0.4426614	0.7156863	0.6607143	0.6822479
7	0.4555031	0.6960784	0.6607143	0.6752451
8	0.4886261	0.7352941	0.5892857	0.6250000
9	0.4886261	0.7352941	0.5892857	0.6299020

Tabela 26 - Distância mínima ao ponto (0,1) utilizando a energia como avaliação e o algoritmo AdaBoost.

Analisando a Tabela 26, é possível constatar que os menores valores de distância ao ponto ótimo estão presentes na escala um com 0.43 e nas escalas números cinco e seis com 0.44. Nas restantes escalas, as distâncias oscilam no intervalo]0.45, 0.57[. As três escalas com menor distância apresentam, respetivamente, capacidade de, aproximadamente, 82%, 71% e 72% para diagnosticar corretamente indivíduos doentes e 61%, 68% e 66% de capacidade para diagnosticar corretamente indivíduos saudáveis.

Analisando os resultados obtidos dos valores abaixo da curva ROC, tem-se que os três maiores valores de área encontram-se nas escalas um, três e seis. Deste modo, as escalas um e seis apresentam menores valores de distância ao mesmo tempo que maiores valores de área abaixo da curva. Para a escala três, aproximadamente 74 em cada 100 indivíduos doentes são corretamente classificados e 61 em cada 100 indivíduos saudáveis são corretamente classificados.

Em suma, e depois de analisar detalhadamente os resultados da tabela, pode concluir-se que a melhor escala é a um, quando se aplica o método da distância mínima à medida gerada pelo algoritmo Adaboost para a avaliação da energia.

k	Distância	Sensibilidade	Especificidade	Área
1	0.3357410	0.7156863	0.8214286	0.8072479
2	0.5482066	0.6176471	0.6071429	0.6855742
3	0.5790179	0.5588235	0.6250000	0.6130952
4	0.5790179	0.5588235	0.6250000	0.6095938
5	0.4940059	0.7254902	0.5892857	0.6577381
6	0.5027528	0.6862745	0.6071429	0.6591387
7	0.5160851	0.6274510	0.6428571	0.6481092
8	0.5678611	0.6274510	0.5714286	0.6120448
9	0.5479692	0.6372549	0.5892857	0.6137955

Tabela 27 – Distância mínima ao ponto (0,1) utilizando a entropia como avaliação e o algoritmo AdaBoost.

Aplicando a metodologia à avaliação da entropia, pela Tabela 27, é possível constatar que os menores valores de distância ao ponto ótimo estão presentes nas escala um, cinco e seis com 0.34, 0.49 e 050. As distâncias associadas às restantes escalas oscilam no intervalo]0.51, 0.58[. Assim, as três escalas com menor distância apresentam, respetivamente, capacidade de, aproximadamente, 72%, 73% e 69% para diagnosticar corretamente indivíduos doentes e 82%, 59% e 61% de capacidade para diagnosticar corretamente indivíduos saudáveis.

Comparando estes resultados com as áreas abaixo da curva ROC, tem-se que os três maiores valores de área encontram-se nas escalas um, dois e seis onde as escalas um e seis apresentam menores valores de distância ao mesmo tempo que maiores valores de área abaixo da curva. Para a escala dois, aproximadamente 62 em cada 100 indivíduos doentes são corretamente classificados e 61 em cada 100 indivíduos saudáveis são corretamente classificados.

Em suma, e depois de analisar detalhadamente os resultados da tabela pode concluir-se que a melhor escala é, igualmente nesta avaliação, a escala um.

Em comum as três tabelas têm a escala um com menor distância (com valores 0.35, 0.43 e 0.34, enquanto todas as restantes escalas assumem valores superiores a 0.43 em todas as avaliações) e com maior área (com valores 0.79, 0.76 e 0.81, quando todas as outras escalas apresentam valores inferiores a 0.69).

Na Tabela 28, são expostos os resultados da soma, sensibilidade, especificidade e área por agrupamento de escalas.

	Distância	Sensibilidade	Especificidade	Área
Desvio padrão	0.4042145	0.7549020	0.6785714	0.7561275
Energia	0.4436518	0.6764706	0.6964286	0.7498249
Entropia	0.4375738	0.7745098	0.6250000	0.7524510
Todos	0.4092814	0.7254902	0.6964286	0.7720588

Tabela 28 – Distância mínima ao ponto (0,1) e o algoritmo AdaBoost.

Destacam-se os valores da distância para o desvio padrão, da especificidade para a avaliação sob a energia, da sensibilidade para a avaliação sob a entropia e a área quando são agrupadas as medidas das três avaliações.

Uma vez mais se conclui que, efetivamente, o algoritmo Adaboost não atinge o melhor valor quando se agrupam todas as escalas de uma avaliação, uma vez que, por exemplo, os valores obtidos para a distância pela escala um de cada uma das avaliações (0.35, 0.43 e 0.34) são melhores que os obtidos utilizando todas as escalas dentro de cada avaliação. (0.40, 0.44 e 0.44). Relativamente à área sob a curva ROC ocorre o mesmo, tendo obtido áreas de 0.79, 0.76 e 0.81 utilizando unicamente as medidas da escala um, enquanto unicamente se obtém 0.76, 0.75 e 0.75 com todas as escalas dentro de cada avaliação.

4.3. Comentários gerais aos resultados obtidos

Neste capítulo serão feitas comparações entre os resultados obtidos anteriormente.

4.3.1 Dados não agrupados *versus* média aritmética

Nesta secção, faz-se a comparação dos resultados de dados não agrupados com os dados agrupados através da média aritmética.

Relativamente à avaliação no âmbito do desvio padrão, pode averiguar-se, através das Tabelas números 5 e 7, que as áreas abaixo das curvas ROC da média aritmética melhoram em relação às escalas números um, oito e nove. Já as escalas de dois a sete, inclusive,

apresentam melhores valores de área quando usadas na sua forma isolada e não aplicando a média.

Atendendo à avaliação no âmbito da energia, pode verificar-se que as áreas abaixo das curvas ROC, da Tabela 7, para as escalas compreendidas entre um a quatro, oito e nove são melhores que as áreas das escalas correspondentes na Tabela 5. Ou seja, para estas é mais eficiente usar a média das áreas das vinte e seis medidas em cada escala em vez da maior área de cada escala. Já as escalas de cinco a nove, inclusive, apresentam melhores valores de área quando usadas na sua forma isolada e não aplicando a média.

Considerando a avaliação no âmbito da entropia, pode concluir-se que apenas as escalas três e quatro saem beneficiadas ao usar a média das medidas de cada k. Já as escalas um e dois e de cinco a nove, inclusive, apresentam melhores valores de área quando usadas na sua forma isolada e não aplicando a média.

Em suma, a avaliação baseada no desvio padrão tem uma melhoria em três escalas, a energia em seis e a entropia em duas. Pode concluir-se que a aplicação da média aritmética, de um modo geral, não melhora as áreas abaixo da curva. Contudo, beneficia a maioria das escalas da energia. Pode dizer-se que a ter de usar a energia como medida de avaliação para a deteção do padrão reticular seria preferível usar a média das escalas.

4.3.2 Dados não agrupados *versus* média ponderada (Adaboost)

Nesta secção, faz-se a comparação dos resultados de dados não agrupados com os dados agrupados através da média ponderada.

Analisando as Tabelas 5 e 10, correspondentes aos dados não agrupados e aos dados agrupados através da média ponderada, recorrendo ao algoritmo Adaboost, é possível concluir que, de modo geral, a procura da melhor curva ROC a partir dos dados iniciais traduz valores de área abaixo da curva superiores aos dos valores resultantes da agregação pelo Adaboost.

No que refere à avaliação no âmbito do desvio padrão, apenas a escala um apresenta um valor da área abaixo da curva superior quando usada a média ponderada, passando de 0.74 para 0.79.

Observando a avaliação no âmbito da energia, as escalas de um a quatro são as que têm uma melhoria nos valores das áreas abaixo das curvas quando usado o Adaboost.

Notando a avaliação no âmbito da entropia, apenas a escala um apresenta um valor da área abaixo da curva superior quando usada a média ponderada, tal como na avaliação sob o desvio padrão, transitando de 0.75 para 0.81.

Por fim, a ter de optar por uma ou outra metodologia, o não agrupamento de dados revela melhor *performance* que o agrupamento. No entanto, para a escala um, quando é usado o Adaboost, há uma melhoria significativa no valor da área abaixo da curva para as três avaliações.

4.3.3 Soma: dados não agrupados *versus* dados Adaboost

Neste subcapítulo, faz-se a comparação dos resultados dos valores de maior soma de sensibilidade e especificidade dos dados não agrupados e dos dados agrupados pela média ponderada (Adaboost).

Relativamente à avaliação no âmbito do desvio padrão, pode averiguar-se, através das Tabelas números 15 e 18, que os maiores valores de soma encontram-se nas escalas um, seis e sete, para a soma dos dados não agrupados, e nas escalas um, cinco e sete, para os dados agrupados com o algoritmo Adaboost.

Examinando a avaliação no âmbito da energia, pode verificar-se, através das Tabelas números 16 e 19, que os maiores valores de soma encontram-se nas escalas um, cinco e seis, tanto com os dados não agrupados como agrupados pelo método Adabost.

Analisando a avaliação no âmbito da entropia, pode concluir-se, através das Tabelas números 17 e 20, que os maiores valores de soma encontram-se nas escalas um, seis e sete,

para a soma dos dados não agrupados, e nas escalas um, cinco e sete, para os dados agrupados com o algoritmo Adaboost. Repare-se que, tal como na avaliação sob o desvio padrão, as escalas são as mesmas para ambas as metodologias.

Por fim, pode concluir-se que a escala um é a que apresenta melhor desempenho tanto no valor da soma como na área abaixo da curva e que o Adaboost é o que apresenta melhores resultados para ambos os valores. Por outro lado, a metodologia dos dados não agrupados apresenta melhores valores de soma para quatro escalas, um, cinco, seis e sete, sendo que a cinco apenas ocorre quando utilizada a avaliação sob a energia. Pode ainda afirmar-se que a metodologia dos dados agrupados pelo Adaboost aponta sempre para as mesmas três escalas como sendo as melhores, um, cinco e sete. Por fim, refira-se que o maior valor de maximização da soma encontra-se na avaliação efetuada sob a entropia, na escala um, com o valor de 1.52.

4.3.4 Distância: dados não agrupados *versus* dados Adaboost

Neste subcapítulo, faz-se a comparação dos resultados dos valores de menor distância ao ponto ótimo (0,1) dos dados não agrupados e dos dados agrupados pela média ponderada (Adaboost).

Examinando a avaliação no âmbito do desvio padrão, pode averiguar-se, através das Tabelas números 22 e 25, que os menores valores de distância encontram-se nas escalas um, seis e sete, para ambas as tabelas.

Notando a avaliação no âmbito da energia, pode verificar-se, através das Tabelas números 23 e 26, que os menores valores de distância encontram-se nas escalas um, cinco e seis para ambas as metodologias.

Observando a avaliação no âmbito da entropia pode concluir-se, através das Tabelas números 24 e 27, que os menores valores de distância encontram-se nas escalas um, seis e sete, para a Tabela 24 e nas escalas um, cinco e seis para a Tabela 27.

Pode concluir-se que a escala um é a que apresenta melhor desempenho tanto na distância como na área abaixo da curva e que o Adaboost é o que apresenta melhor valor para a área

e menor valor de distância ao ponto de corte ótimo (0,1). Para ambas as metodologias usadas, as melhores escalas são as um, cinco, seis e sete, sendo que apenas na avaliação sob a entropia as escalas diferem entre as metodologias. O menor valor da minimização da distância encontra-se na escala um da entropia, quando é usado o método Adaboost, com o valor de 0.34.

4.3.5 Soma e distância dados não agrupados

Se seguida, apresentam-se algumas comparações e conclusões entre a metodologia da maior soma da sensibilidade com a especificidade e da menor distância ao ponto ótimo (0,1) para os dados não agrupados.

Atendendo às avaliações nos âmbitos do desvio padrão e da entropia, pode averiguar-se, através das Tabelas números 15 e 22, e das Tabelas números 17 e 24, que os três melhores valores de área abaixo da curva encontram-se nas mesmas escalas para ambas as metodologias, escalas um, dois e seis.

Considerando a avaliação no âmbito da energia, pode verificar-se, através das Tabelas números 16 e 23, que, mais uma vez, os três melhores valores de área abaixo da curva encontram-se nas mesmas escalas para ambas as metodologias, escalas um, cinco e seis.

Em suma, concluiu-se que as três avaliações, desvio padrão, energia e entropia têm em comum as escalas um, dois, cinco e seis como sendo as que apresentam melhor valor de área abaixo da curva ROC. Por outro lado, a escala cinco apenas ocorre na avaliação sob a energia, enquanto a escala um é a que apresenta melhor desempenho em ambas as metodologias e o melhor valor de área pertence à maximização da soma da sensibilidade com a especificidade sob a entropia com 0.75 de área abaixo da curva.

4.3.6 Soma e distância dados agrupados pela média ponderada

De seguida, apresentam-se algumas comparações e conclusões entre a metodologia da maior soma da sensibilidade com a especificidade e da menor distância ao ponto ótimo (0,1) para os dados agrupados pela média ponderada (Adaboost).

Analisando a avaliação no âmbito do desvio padrão, pode verificar-se, através das Tabelas números 18 e 25, que os três melhores valores de área abaixo da curva encontram-se nas escalas um, dois e seis e nas escalas um, dois e sete, respetivamente.

Observando a avaliação no âmbito da entropia, pode verificar-se, através das Tabelas números 19 e 26, que os três melhores valores de área abaixo da curva encontram-se nas escalas um, três e cinco e nas escalas um, três e seis, respetivamente.

Olhando a avaliação no âmbito da entropia, pode verificar-se, através das Tabelas números 20 e 27, que os três melhores valores de área abaixo da curva, para ambas as metodologias encontram-se nas escalas um, dois e seis.

Deste modo, pode afirmar-se que as três avaliações, desvio padrão, energia e entropia têm em comum as escalas um, dois, três, cinco e seis como sendo as que apresentam melhor valor de área abaixo da curva ROC. A avaliação sob a entropia é a única que contempla as mesmas escalas para ambas as metodologias. A escala um é a que apresenta melhor desempenho em ambas as metodologias, sendo o melhor valor de área obtido com a avaliação da entropia aquando da maximização da soma da sensibilidade com a especificidade, com 0.82 de área abaixo da curva.

4.3.7 Os melhores desempenhos em cada medida

Nesta subsecção vão ser indicados, entre os pontos ótimos determinados pelas metodologias previamente apresentadas, os pontos que apresentam melhor desempenho na sensibilidade, especificidade, soma de sensibilidade e especificidade e distância ao ponto ideal, tal como a área sob a curva ROC.

O maior valor de sensibilidade, 0.873, encontra-se na análise da maximização da soma da sensibilidade e da especificidade em dados não agrupados, escala dois da entropia.

O maior valor de especificidade, 0.839, observa-se na análise da maximização da soma da sensibilidade e da especificidade em dados agrupados pelo algoritmo Adaboost, na escala um da energia e na escala dois da entropia.

O maior valor da área, 0.821, alcança-se na análise da maximização da soma da sensibilidade e da especificidade em dados agrupados pelo algoritmo Adaboost, escala um da entropia.

O maior valor de soma dos pontos da sensibilidade com a especificidade, 1.519, obtém-se na análise da maximização da soma da sensibilidade e da especificidade em dados agrupados pelo algoritmo Adaboost, escala um da entropia.

Na análise da minimização da distância ao ponto ótimo (0,1), o menor valor é 0.336, obtido com os dados agrupados pelo algoritmo Adaboost, escala um da entropia.

Assim, tal como salientado em análises anteriores, destaca-se a escala um como as escala onde se obtêm, na maioria das vezes, melhores resultados para todas as avaliações. A avaliação da entropia também demonstra ter um bom desempenho quando se olha isoladamente para os valores obtidos numa das medidas de fiabilidade. O algoritmo Adaboost revela que é possível, por vezes, melhorar o desempenho através do agrupamento de medidas aplicando a média ponderada ao atribuir pesos maiores às medidas que têm maior contributo na deteção do padrão reticular.

5. Conclusão

Esta dissertação mostra, recorrendo a um caso de estudo, como é que a utilização da curva ROC pode ser útil na análise da fiabilidade de um diagnóstico médico. São igualmente apresentados os demais conceitos de epidemiologia associados, tais como, prevalência, acurácia, sensibilidade, especificidade, valores preditivos positivo e negativo, razões de verosimilhança positiva e negativa, área abaixo da curva ROC ou ponto de corte, ferramentas úteis na avaliação de testes de diagnóstico e na visualização desse mesmo desempenho. A precisão de um teste de diagnóstico pode ser avaliada pela acurácia que traduz a percentagem de diagnósticos corretamente determinados. Por outro lado, as principais medidas de fiabilidade de um teste de diagnóstico são a sensibilidade, que traduz a capacidade do teste em identificar um indivíduo doente, e a especificidade, que corresponde à probabilidade de se obter um resultado negativo entre os indivíduos saudáveis. Outras medidas igualmente utilizadas na prática, embora com menor frequência, são o VPP, que identifica a proporção de doentes entre os indivíduos que recebem resultado positivo, e o VPN, que corresponde à probabilidade de um indivíduo, que recebeu resultado negativo, estar efetivamente saudável. Naturalmente, quanto mais próximos os valores destas medidas estiverem de 100% melhor será a fiabilidade do teste.

Muitos dos testes de diagnóstico são realizados com base na informação de variáveis quantitativas associadas ao indivíduo. Por este motivo, é fundamental definir o valor a partir do qual um indivíduo deve ser classificado como doente. Neste contexto, o ponto de corte corresponde ao ponto de separação na identificação dos indivíduos como doentes ou saudáveis. A curva ROC, ao representar todos os possíveis valores para a sensibilidade e especificidade em função do ponto de corte, insere a informação relevante para a análise de fiabilidade do teste de diagnóstico associado. Em particular, a área abaixo da curva é uma medida que permite aferir qual a probabilidade do indivíduo doente obter um resultado verdadeiro positivo e do indivíduo saudável obter um resultado verdadeiro negativo. Um teste totalmente inapto de discriminar indivíduos doentes de saudáveis, ou quaisquer outras duas classes, é o que apresentar uma área sob a curva de 0.5, pois significa que sensibilidade é sempre igual a 1-especificidade, ou seja, $x=y$. Quanto mais a curva se aproxima do canto superior esquerdo (do ponto ideal (0,1), que corresponde à ausência de erros de classificação) melhor é a qualidade do teste, logo, quanto maior for o valor da área, isto é, quanto mais

próximo estiver da unidade, maior é a capacidade para discriminar estes dois tipos de indivíduos.

Como aplicação, foi explorado este contexto no caso de estudo da deteção automatizada do padrão reticular ou rede de pigmentos, através da dermatoscopia. A dermatoscopia é uma técnica de imagem, não invasiva, usada para avaliar as cores e estrutura das lesões pigmentadas da pele, tornando-se, cada vez mais, uma ferramenta ativa para a medicina preventiva. Com recurso a técnicas de processamento de imagem, foi utilizada uma base de dados com 158 imagens, analisadas em 9 escalas, em que cada escala está dividida em 26 medidas. Para cada medida foram calculadas três avaliações: desvio padrão, energia e entropia.

O objetivo da utilização da base de dados de imagens dermatológicas é aferir a existência ou não de padrão reticular em cada imagem. Para tal, foram usadas duas metodologias: a procura da "melhor" curva ROC e a procura do "melhor" ponto de corte. Na primeira, procurou-se a "melhor" curva através dos dados não agrupados e através dos dados agrupados, onde se aplicou o agrupamento por média aritmética e média ponderada, otimizada pelo algoritmo Adaboost. Na segunda, procurou-se o "melhor" ponto de corte através da maximização da soma da sensibilidade com a especificidade e através da minimização da distância ao ponto ótimo (0,1), quer para os dados não agrupados e quer para os dados agrupados pela média ponderada.

Com recurso às metodologias acima descritas, verificou-se que a aplicação do processamento de imagem no diagnóstico médico produz efetivamente resultados úteis e fiáveis. Neste sentido, comparando os resultados obtidos, na procura da "melhor" curva entre o agrupamento aritmético e as medidas isoladas, é possível constatar que a aplicação da média aritmética é imprevisível, pois tanto melhora como piora os valores das áreas abaixo da curva. Considerando apenas as escalas número um para as três avaliações, desvio padrão, energia e entropia, pois é a que sobressai entre as diferentes escalas, a média aritmética melhora os resultados para as avaliações do desvio padrão e da energia. Contudo, piora o da entropia, pois a média aritmética atribui o mesmo peso a todos os valores, sejam eles bons ou maus para a classificação pretendida. Além de imprevisível, pode também ser arriscado usar esta média sem analisar aprofundadamente os dados que se têm em mãos. Por exemplo, quando se estão a estudar os dados de forma não agrupada, os dados na sua forma original,

tem-se que a pior avaliação é a energia, uma vez que é a que tem maior quantidade de medidas com áreas inferiores a 0.65 e, consequentemente, a que possui áreas abaixo da curva mais baixas. Aquando da aplicação da média arimética, a energia dá o lugar de pior avaliação ao desvio padrão (exceto no primeira escala), embora o desvio padrão e a entropia sejam as avaliações que apresentam menos medidas com áreas abaixo da curva ROC inferiores a 0.65.

Confrontando os resultados obtidos para os dados não agrupados e para os dados agrupados pela média ponderada na procura do "melhor" ponto ótimo, é possível tirar algumas conclusões, quer recorrendo à soma da sensibilidade com a especificidade, quer usando a distância ao ponto ótimo (0,1). Quando se considera a escala um em cada uma das três avaliações, a escala com melhor *performance*, há uma melhoria dos valores das áreas abaixo da curva ROC, dos valores da soma e dos valores da distância ao ponto ideal (0,1), com a aplicação do algoritmo Adaboost. Contudo, ao contrário do que era expectável, a média ponderada otimizada pelo algoritmo Adaboost nem sempre revela os melhores resultados. Exemplo disso são as escalas dois, quatro, oito e nove do desvio padrão, três, quatro, cinco e oito da energia e dois, quatro, cinco e nove da entropia, quer para a soma quer para a distância.

Apesar do Adaboost não apresentar sempre os melhores resultados, é com este algoritmo que se obtêm os valores mais elevados para a área abaixo da curva ROC e para a soma da sensibilidade e especificidade, bem como os menores valores de distância ao ponto ideal (0,1). Exemplo disso é a escala um da avaliação sob o desvio padrão, energia e entropia, tanto nas metodologias usadas na procura da "melhor" curva, como nas metodologias usadas na procura do "melhor" ponto de corte.

Algo que acontece com alguma frequência na procura do "melhor" ponto de corte é o facto da escala que apresenta melhor área abaixo da curva ROC não ser a mesma que apresenta o "melhor" ponto. Tal ocorre porque maior área sob a curva ROC não implica necessariamente melhor desempenho no "melhor" ponto de corte.

Por fim, os resultados mostram que, quer usando medidas isoladas quer usando medidas agrupadas, há boas opções para a escolha do teste de diagnóstico, sendo que o método Adaboost revelou em algumas situações uma melhoria significativa dos resultados alcançados.

Ao longo desta dissertação, assumiu-se apenas a possibilidade da população pertencer a um dos grupos, doente ou saudável, que é a forma básica e *standard* de trabalhar com a curva ROC.

Contudo, existe a curva ROC *fuzzy*. Num conjunto *fuzzy* um indivíduo pode pertencer parcialmente ao grupo dos doentes ou ao grupo dos saudáveis. Esta definição permite atenuar a restrição rígida da escolha de um dos grupos. Com a metodologia *fuzzy*, um indivíduo com um determinado resultado de uma análise pode não ser imediatamente classificado como doente ou como saudável, podendo ao profissionais de saúde tomar algumas precauções caso o utente se encontre nos limiares do intervalo considerado. A principal vantagem da metodologia *fuzzy* é permitir uma melhor representação da realidade quando não há um ponto de corte exato, pois nenhum ponto tem sensibilidade ou especificidade iguais a 1, valor que seria obtido se fosse ideal. Todos os conceitos inerentes à curva ROC descritos ao longo da dissertação são também os da curva ROC *fuzzy*, a diferença reside no facto desta última ser mais flexível que a primeira. Assim, poderia ser levado um estudo semelhante ao desenvolvido, mas onde o foco fosse a análise da curva ROC *fuzzy* no caso em que se consideram medidas isoladas ou agrupadas [38].

Outra questão que poderia ser analisada num trabalho futuro seria a análise do valor da área parcial em vez do valor da área total. Este tipo de área concentra-se na região mais vantajosa de sensibilidade e especificidade, ignorando os valores de menor interesse. A área sob a curva ROC resume-a como um todo e uma medida baseada em toda esta extensão pode tornar-se inadequada. Uma alternativa para esta situação foi introduzida por Thompson e Zucchini que a denominaram de área parcial sob a curva ROC, estimada sobre alguma região de interesse da curva [10].

Bibliografia

[1] A. António, A. Balsa, *et al.*, *Enciclopédia de Medicina*, Printer Portuguesa. Lisboa: Seleções do Reader's Digest, 2000.

[2] M. T. S. Backes, L. M. da Rosa, G. C. M. Fernandes, S. G. Becker, B. H. S. Meirelles, and S. M. de A. dos Santos, "Concepts of heallth/disease along history under the light of epidemiology and anthropology" *Rev. enferm. UERJ, Rio Janeiro*, pp. 1–7, 2009.

[3] T. L. Skare, I. Nakano, D. L. Escuissiato, R. Batistetti, T. de O. Rodrigues, and M. B. Silva, "Pulmonary changes on high-resolution computed tomography of patients with rheumatoid arthritis and their association with clinical, demographic, serological and therapeutic variables" *Rev Bras Reum.*, vol. 51(4), pp. 325–337, 2011.

[4] M. Machado, J. Pereira, and R. Fonseca-Pinto, "Reticular pattern detection in dermoscopy: an approach using Curvelet Transform" *Biomed. Eng. (NY).*, vol. 32, no. 2, pp. 129–136, 2016.

[5] Wojtek J., Krzanowski, and D. J. Hand, *ROC Curves for Continuous Data*, CRC Press Book, 2013.

[6] C. E. Metz, "ROC analysis in medical imaging: a tutorial review of the literature" *Radiol Phys Technol*, vol. 1, pp. 2–12, 2008.

[7] "Gustav Theodor Fechner." [Online]. Available:
http://www.biografiasyvidas.com/biografia/f/fechner.htm. [Accessed: 16-Oct-2016].

[8] A. C. Silva Braga, *Curvas ROC: aspectos funcionais e aplicações,* Universidade do Minho, 2000.

[9] "Louis Leon Thurstone." [Online]. Available:
http://www.biografiasyvidas.com/biografia/t/thurstone.htm. [Accessed: 16-Oct-2016].

[10] E. Z. Martinez, F. Louzado-Neto, and B. de B. Pereira, "A curva ROC para testes diagnósticos" in *Cadernos Saúde Coletiva*, Rio de Janeiro,, pp. 7–9, 2003.

[11] M. S. Uva, P. Victorino, R. Roquette, and C. M. D. Ausenda Machado, "Investigação epidemiológica sobre prevalência e incidência de hipertensão arterial na população portuguesa - uma revisão de âmbito", 2014.

[12] V. da S. Caldeira, C. C. D. Starling, R. R. Britto, J. A. Martins, R. F. Sampaio, and Verônica Franco Parreira, "Precisão e acurácia da cirtometria em adultos saudáveis" *J Bras Pneumol*, pp. 219–529, 2007.

[13] F. Matos e D. da Cruz, "Construção de instrumento para avaliar a acurácia diagnóstica" *Rev. Enfermagem*, 2009.

[14] V. Chiaraa, R. Sichierib, and e P. D. Martins, "Sensitivity and specificity of overweight classification of adolescents, Brazil" *Rev Saúde Pública,* 37(2):226-31, 2003.

[15] F. Pitanga, "Sensibilidade e especificidade do índice de conicidade como discriminador do risco coronariano de adultos em Salvador, Brasil" *Rev. Bras. Epidemiol.*, pp. 259–269, 2004.

[16] D. Junior, S. Piato, *et al.*,"Valores preditivos das categorias 3, 4 e 5 do sistema BI-RADS em lesões mamárias nodulares não-palpáveis avaliadas por mamografia, ultra-sonografia e ressonância magnética", *Radiol Bras*, 40(2):93-98, 2007.

[17] R. Silva, *et al.*, "Bem-estar psicológico e adolescência: fatores associados" *Repos. Inst. da Univ. Fed. do Rio Gd.*, 2007.

[18] R. Santos, *Probabilidades e Conceitos Associados*, pp. 36–40, 2016.

[19] H. C. P. Morana, *Identificação do ponto de corte para a escala PCL-R (Psychopathy Checklist Revised) em população forense brasileira: caracterização de dois subtipos de personalidade; transtorno global e parcial*, Faculdade de Medicina, 2003.

[20] B. Fluss, D. Faraggi, "Estimation of the Youden Index and its associated cutoff poin" *Biom J*, pp. 458–72, 2005.

[21] "What is R." [Online]. Available: https://www.r-project.org/about.html.

[22] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, "ROCR: visualizing classifier performance in R." *Bioinformatics*, vol. 21, pp. 78–81, 2005.

[23] T. Sing, O. Sander, N. Beerenwinkel, and T. Lengauer, "Package 'ROCR", 2015.

[24] H. Ma, A. Bandos, H. Rockette, and D. Gur, "On use of partial area under the ROC curve for evaluation of diagnostic performance" *Stat Med*, vol. 32, pp. 3449-3458, 2013.

[25] M. H., A. Bandos, and D. Gur, "On the use of partial area under the ROC curve for comparison of two diagnostic tests" *Biom J*, vol. 57, pp. 304-320, 2015.

[26] D. Walter, "The partial area under the summary ROC curve," *Stat Med*, vol. 24, pp. 2025-2040, 2005.

[27] F. Pitanga e I. Lessa, "Indicadores Antropométricos de Obesidade como Instrumento de Triagem para Risco Coronariano Elevado em Adultos na Cidade de Salvador – Bahia" *Arq. Bras. Cardiol.*, vol. 85, pp. 26–31, 2005.

[28] F. Nunes, "Introdução ao Processamento de Imagens Médicas para Auxílio ao Diagnóstico – Uma Visão Prática" vol. 2, pp. 73–124, 2006.

[29] R. Fonseca-Pinto, "Processamento de Imagem Digital e Cancro de Pele: uma abordagem interdisciplinar", pp. 1–11, 2014.

[30] M. J. Schneider, *Processamento digital de imagens tecnológicas na resolução em sensoriamento remoto: o caso da fusão de imagens de diferentes resoluções espaciais*, Universidade Federal do Paraná, 2001.

[31] J. Cordeiro, *et al.*, "Álgebra de Geo-Campos e suas Aplicações" *Simpósio Bras. Sensoriamento Remoto*, pp. 691–696, 1996.

[32] Advance Care, "Cancro da pele." [Online]. Available:
https://advancecare.pt/glossario/cancro-da-pele/. [Accessed: 05-May-2017].

[33] Y. Freund and R. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting." in *Computational Learning Theory: Eurocolt 95*, Springer-Verlag, pp. 23–37, 1999.

[34] E. Alfaro, M. Gamez, and N. Garcia, "Package 'adabag'" 2015.

[35] E. Alfaro, M. Gamez, and N. Garcia, "Adabag: An R Package for Classification with Boosting and Bagging" *J. Stat. Softw.*, vol. 54, no. 2, pp. 1–35, 2013.

[36] X. Liu, "Classification accuracy and cut point selection" *Stat Med*, vol. 31, pp. 2676–2686, 2012.

[37] N. Perkins and E. Schisterman,"The inconsistency of "optimal" cut-points using two ROC based criteria" *Am J Epidemiol*, vol. 163, pp. 670–675, 2006.

[38] M. Castanho, A. Yamakami, L. Barros, and L. Vendite, “Fuzzy Receiver Operating Characteristic Curve: An Option to Evaluate Diagnostic Tests”, IEEE Transactions on Information Technology in Biomedicine, vol. 11(3): 244–250, 2007.

Anexo A - Código

1.

```
# Ponto de corte perfeito
x <- seq(0, 25, length.out=1000)
plotDistr(x, dnorm(x, mean=4, sd=1), lwd =3,ylim=c(0,0.5))
lines(x, dnorm(x, mean=4, sd=1), col="blue", lwd =3)
text(4, 0.42, "Saudáveis", cex = 1.5, col="blue")
lines(x, dnorm(x, mean=18, sd=2),col="red",lwd =3)
text(18, 0.22, "Doentes", cex = 1.5, col="red")
abline(v =10, lwd =3)
text(14.75, 0.48, "Ponto de corte", cex = 1.5, col="black")
```

2.

```
# O que acontece na realidade + ponto de corte
x <- seq(0, 25, length.out=1000)
plotDistr(x, dnorm(x, mean=9, sd=1), lwd =3,ylim=c(0,0.5))
lines(x, dnorm(x, mean=9, sd=1), col="blue", lwd =3)
text(4, 0.42, "Saudáveis", cex = 1.5, col="blue")
lines(x, dnorm(x, mean=13, sd=2),col="red",lwd =3)
text(18.5, 0.22, "Doentes", cex = 1.5, col="red")
abline(v =10.7, lwd =3)
text(15, 0.48, "Ponto de corte", cex = 1.5, col="black")
cord.x <- c(10.7,seq(10.7,13,0.01),13)
cord.y <- c(0,dnorm(seq(10.7,13,0.01),8.5,1.5),0)
polygon(cord.x, cord.y, col="blue")
cord.x <- c(6.9,seq(6.9,10.7,0.01),10.7)
cord.y <- c(0,dnorm(seq(6.9,10.7,0.01),13,2),0)
polygon(cord.x, cord.y, col="red")
```

3.

```
# Aumento da sensibilidade (ponto de corte --> direita)
x <- seq(0, 25, length.out=1000)
plotDistr(x, dnorm(x, mean=9, sd=1), lwd =3,ylim=c(0,0.5))
lines(x, dnorm(x, mean=9, sd=1), col="blue", lwd =3)
text(4, 0.42, "Saudáveis", cex = 1.5, col="blue")
lines(x, dnorm(x, mean=13, sd=2),col="red",lwd =3)
text(18.5, 0.22, "Doentes", cex = 1.5, col="red")
abline(v =11.7, lwd =3)
text(16, 0.48, "Ponto de corte", cex = 1.5, col="black")
cord.x <- c(11.7,seq(11.7,13,0.01),13)
cord.y <- c(0,dnorm(seq(11.7,13,0.01),9.1,1),0)
polygon(cord.x, cord.y, col="blue")
cord.x <- c(3,seq(3,11.7,0.01),11.7)
cord.y <- c(0,dnorm(seq(3,11.7,0.01),13,2),0)
polygon(cord.x, cord.y, col="red")
lines(x, dnorm(x, mean=9, sd=1), col="blue", lwd =3)
```

4.

```
* Aumento da especificidade (ponto de corte --> esquerda)
x <- seq(0, 25, length.out=1000)
plotDistr(x, dnorm(x, mean=9, sd=1), lwd =3,ylim=c(0,0.5))
lines(x, dnorm(x, mean=9, sd=1), col="blue", lwd =3)
text(4, 0.42, "Saudáveis", cex = 1.5, col="blue")
lines(x, dnorm(x, mean=13, sd=2),col="red",lwd =3)
text(18.5, 0.22, "Doentes", cex = 1.5, col="red")
abline(v =9, lwd =3)
text(15, 0.48, "Ponto de corte", cex = 1.5, col="black")
cord.x <- c(9,seq(9,13,0.01),13)
cord.y <- c(0,dnorm(seq(9,13,0.01),9.1,1),0)
polygon(cord.x, cord.y, col="blue")
cord.x <- c(3,seq(3,9,0.01),9)
cord.y <- c(0,dnorm(seq(3,9,0.01),13,2),0)
polygon(cord.x, cord.y, col="red")
lines(x, dnorm(x, mean=13, sd=2),col="red",lwd =3)
```

5.

```
# Representação exemplo ilustrativo
x <- seq(34, 178, length.out=1000)
plotDistr(x, dnorm(x, mean=115, sd=9), lwd =3,ylim=c(0,0.1))
lines(x, dnorm(x, mean=115, sd=9), col="red", lwd =3)
text(140, 0.06, "Doentes", cex = 1.5, col="red")
lines(x, dnorm(x, mean=90, sd=6),col="blue",lwd =3)
text(60, 0.07, "Saudáveis", cex = 1.5, col="blue")
```

6.

```
# Curva ROC com dados do exemplo ilustrativo
pred <- prediction(mariana$medida, mariana$estado)
perf <- performance(pred,"tpr","fpr")
plot(perf,colorize=TRUE,lty=1,lwd=4,main="Curva ROC", xlab="1 -
Especificidade", ylab="Sensibilidade")
segments(0,0,1,1, lty=3, lwd=3, col="black")
abline(h=seq(0,1),v=seq(0,1),lty=3,lwd=2, col="black")
```

7.

```
# Níveis de discriminação
pred.e1 <- prediction(ROCR.simple$predictions, ROCR.simple$labels)
perf.e1 <- performance(pred.e1, 'tpr', 'fpr')
pred.e3 <- prediction(ROCR.simple$predictions, ROCR.simple$labels3)
perf.e3 <- performance(pred.e3, 'tpr', 'fpr')
pred.e5 <- prediction(ROCR.simple$predictions, ROCR.simple$labels5)
perf.e5 <- performance(pred.e5, 'tpr', 'fpr')
plot(perf.e1, lty=3, col="red",main="Curvas ROC", xlab="1 - Especificidade",
ylab="Sensibilidade")
plot(perf.e3, lty=3, col="green",add=TRUE)
```

```
plot(perf.e5, lty=3, col="orange",add=TRUE)
plot(perf.e1, avg="vertical", lwd=3, col="red",
spread.estimate="stderror",plotCI.lwd=2,add=TRUE)
plot(perf.e3, avg="vertical", lwd=3, col="green",
spread.estimate="stderror",plotCI.lwd=2,add=TRUE)
plot(perf.e5, avg="vertical", lwd=3, col="orange",
spread.estimate="stderror",plotCI.lwd=2,add=TRUE)
legend(0.6,0.5,c('Elevada','Média','Baixa'),col=c('red','green','orange'),lwd=3)
segments(0,0,1,1, lty=3, lwd=3, col="black")
```

8.

```
# Standardizar todas as colunas
for (i in 4:705) {BD[[i]] = (BD[[i]]-mean(BD[[i]]))/sd(BD[[i]])}
```

9.

```
# Torna simétricos os valores das colunas cuja curva ROC está abaixo da curva x=y
for (i in 4:705) {
pred <- prediction(BD[i], BD[2])
if (unlist(slot(performance(pred,"auc"),"y.values"))<0.5) {BD[[i]] = - BD[[i]]}
}
```

10.

```
# Gráfico com as 702 colunas
pred <- prediction(BD[4], BD[2])
perf <- performance(pred,"tpr","fpr")
plot(perf, lty=3, col=4, xlab="1 - Especificidade", ylab="Sensibilidade")
plot(perf, avg="vertical", lwd=3, col=4,
spread.estimate="stderror",plotCI.lwd=2,add=TRUE)
for (n in 2:702) {pred <- prediction(BD[n+3], BD[2])
perf <- performance(pred,"tpr","fpr")
plot(perf, lty=3, col=n,add=TRUE)
plot(perf, avg="vertical", lwd=3, col=n,
spread.estimate="stderror",plotCI.lwd=2,add=TRUE)}
segments(0,0,1,1, lty=3, lwd=3, col="black")
```

11.

```
# Maior área e respetiva coluna
area=c()
for (k in 1:27)
{
pred <- prediction(BD[4+(k-1)*26], BD[2])
perf <- performance(pred,"tpr","fpr")
#plot(perf, lwd=2, col=4,main="Curva ROC - Receiver Operating Characteristics",
xlab="1 - Especificidade", ylab="Sensibilidade")
auc<- performance(pred,"auc")
auc<- unlist(slot(auc, "y.values"))
area[(k-1)*26+1]<-auc
```

```
        for (n in 2:26)
        {
        pred <- prediction(BD[n+3+(k-1)*26], BD[2])
        perf <- performance(pred,"tpr","fpr")
        #plot(perf, lty=3, col=n,add=TRUE)
        auc<- performance(pred,"auc")
        auc<- unlist(slot(auc, "y.values"))
        area[(k-1)*26+n]<-auc
        }
}
#coluna
which.max(area)
#área
max(area)
```

12.

```
# Todas e coluna 486 destacada
pred <- prediction(BD[4], BD[2])
perf <- performance(pred,"tpr","fpr")
plot(perf, lty=3, col=4,main="Curva ROC - Receiver Operating Characteristics",
xlab="1 - Especificidade", ylab="Sensibilidade")
for (n in 4:705) {
pred <- prediction(BD[n+3], BD[2])
perf <- performance(pred,"tpr","fpr")
plot(perf, lty=3, col=n,add=TRUE)
}
segments(0,0,1,1, lty=3, lwd=3, col="black")
pred_aucmax <- prediction(BD$M486, BD[2])
perf_aucmax <- performance(pred_aucmax,"tpr","fpr")
plot(perf_aucmax, lwd=3, col="red",add=TRUE)
auc_1<- performance(pred_aucmax,"auc")
auc_1<- unlist(slot(auc_1, "y.values"))
auc_1<-round(auc_1, digits = 2)
auc_1<- paste(c("auc_1 = "),auc_1,sep="")
```

13.

```
# Gráfico com 26 curvas dos k=[1...27]
k=1
pred <- prediction(BD[26*(k-1)+4], BD[2])
perf <- performance(pred,"tpr","fpr")
plot(perf, lty=3, col=4, xlab="1 - Especificidade", ylab="Sensibilidade")
plot(perf, avg="vertical", lwd=3, col=4,
spread.estimate="stderror",plotCI.lwd=2,add=TRUE)
for (n in 1:25) {pred <- prediction(BD[26*(k-1)+n+4], BD[2])
perf <- performance(pred,"tpr","fpr")
plot(perf, lty=3, col=n,add=TRUE)
```

```
plot(perf, avg="vertical", lwd=3, col=n,
spread.estimate="stderror",plotCI.lwd=2,add=TRUE)}
segments(0,0,1,1, lty=3, lwd=3, col="black")
```

14.

```
# Maior área de cada k e respetiva coluna
area=c()
for (k in 1:702)
{pred <- prediction(BD[3+k], BD[2])
area[k]<- unlist(slot(performance(pred,"auc") , "y.values"))}
# k=[1...27]
k=1
max(area[((k-1)*26+1):(k*26)])
which.max(area[((k-1)*26+1):(k*26)])+(k-1)*26
```

15.

```
# Cria um vector com a área obtida em cada uma das 702 colunas
area=c()
for (k in 1:702)
{pred <- prediction(BD[3+k], BD[2])
area[k]<- unlist(slot(performance(pred,"auc") , "y.values"))}
# Qual é a área máxima no vetor area?
max(area)
# Em que casa está a área máxima no vetor area?
which.max(area)
```

16.

```
# Cria vetor com as posições do vetor area que têm auc  >= 0.65, 0.70 e 0.73
VM=0
posicao=c(which(area>=0.65))
for (i in posicao) {VM=VM+BD[i+3]}
pred <- prediction(VM, BD[2])
unlist(slot(performance(pred,"auc") , "y.values"))
```

17.

```
# Cria 3 vetores com as posições do vetor area que têm auc  >= 0.65, 0.7, 0.73
VM65=0
posicao=c(which(area>=0.65))
for (i in posicao) {VM65=VM65+BD[i+3]/length(posicao)}
VM70=0
posicao=c(which(area>=0.70))
for (i in posicao) {VM70=VM70+BD[i+3]/length(posicao)}
VM75=0
posicao=c(which(area>=0.73))
for (i in posicao) {VM75=VM75+BD[i+3]/length(posicao)}
# Curvas ROC
pred.M1 <- prediction(VM65, BD[2])
```

```
perf.M1 <- performance(pred.M1,"tpr","fpr")
pred.M2 <- prediction(VM70, BD[2])
perf.M2 <- performance(pred.M2,"tpr","fpr")
pred.M3 <- prediction(VM75, BD[2])
perf.M3 <- performance(pred.M3,"tpr","fpr")
plot(perf.M1, lty=3, col="blue",main="Curva ROC - Receiver Operating
Characteristics", xlab="1 - Especificidade", ylab="Sensibilidade")
plot(perf.M2, lty=3, col="green",add=TRUE)
plot(perf.M3, lty=3, col="red",add=TRUE)
plot(perf.M1, avg="vertical", lwd=3, col="blue",
spread.estimate="stderror",plotCI.lwd=2,add=TRUE)
plot(perf.M2, avg="vertical", lwd=3, col="green",
spread.estimate="stderror",plotCI.lwd=2,add=TRUE)
plot(perf.M3, avg="vertical", lwd=3, col="red",
spread.estimate="stderror",plotCI.lwd=2,add=TRUE)
legend(0.6,0.5,c('>0.65','>0.70','>0.73'),col=c('blue','green','red'),lwd=3)
segments(0,0,1,1, lty=3, lwd=3, col="black")
```

18.

```
# Determina o número de valores inferiores a 0.65
area=c()
for (k in 1:702)
{pred <- prediction(BD[3+k], BD[2])
area[k]<- unlist(slot(performance(pred,"auc") , "y.values"))}
# k=[1...27]
k=1
length(c(which(area[((k-1)*26+4):(k*26+3)]<0.75)))
```

19.

```
# Calcular o ponto ótimo sensibilidade vs especificidade para cada k
# k=[1...27]
espmax=c()
sensmax=c()
somamax=c()
for (i in 1:26)
{pred <- prediction(BD[(k-1)*26+3+i], BD[2])
perf<- performance(pred,"sens","spec")
esp= unlist(slot(perf, "x.values"))
sens= unlist(slot(perf, "y.values"))
soma=esp+sens
somamax[i]=max(soma)
espmax[i]=esp[which.max(soma)]
sensmax[i]=sens[which.max(soma)]}
which.max(somamax)+(k-1)*26 #O máximo verifica-se na variável M…
max= which.max(somamax)
somamax[max] #valor máximo observado
espmax[max] #especificidade associada ao valor máximo
```

```
sensmax[max] #sensibilidade associada ao valor máximo
pred <- prediction(BD[max+(k-1)*26+3], BD[2])
unlist(slot(performance(pred,"auc") , "y.values")) # área abaixo da curva ROC
associada ao valor máximo
```

20.

```
# Calcular o ponto ótimo sensibilidade vs especificidade para cada k
# k=[1...27]
espmax=c()
sensmax=c()
distmin=c()
for (i in 1:26)
{pred <- prediction(BD[(k-1)*26+3+i], BD[2])
perf<- performance(pred,"sens","spec")
esp= unlist(slot(perf, "x.values"))
sens= unlist(slot(perf, "y.values"))
dist=sqrt((1-esp)^2+(sens-1)^2)
distmin[i]= min(dist)
espmax[i]=esp[which.min(dist)]
sensmax[i]=sens[which.min(dist)]}
which.min(distmin)+(k-1)*26 #O máximo verifica-se na variável M…
min= which.min(distmin)
distmin[min] #valor máximo observado
espmax[min] #especificidade associada ao valor máximo
sensmax[min] #sensibilidade associada ao valor máximo
pred <- prediction(BD[min+(k-1)*26+3], BD[2])
unlist(slot(performance(pred,"auc") , "y.values")) # área abaixo da curva ROC
associada ao valor máximo
```

21.

```
# calcula pesos
# k=[1...27]
BDada=cbind(BD[2],BD[((k-1)*26+4):((k-1)*26+29)])
BDada$Caso<-factor(BDada$Caso)
adaboost<-boosting(Caso~., data=BDada, boos=TRUE, mfinal=5)
summary(adaboost)
errorevol(adaboost,BDada)
Adaboost<-predict(adaboost,BDada)
Adaboost
Adaboost=as.numeric(as.character(Adaboost$class))
Adaboost
pesos_ada=cbind(adaboost$importance)
names(pesos_ada)=names(adaboost$importance)
pesos_ada=pesos_ada[pesos_ada>0]
pesos_ada
```

22.

```
# calcula áreas
# k=[1...27]
BDada=cbind(BD[2],BD[((k-1)*26+4):((k-1)*26+29)])
BDada$Caso<-factor(BDada$Caso)
adaboost<-boosting(Caso~., data=BDada, boos=TRUE, mfinal=5)
summary(adaboost)
errorevol(adaboost,BDada)
Adaboost<-predict(adaboost,BDada)
Adaboost=as.numeric(as.character(Adaboost$class))
pesos_ada=cbind(adaboost$importance)
names(pesos_ada)=names(adaboost$importance)
pesos_ada=pesos_ada[pesos_ada>0]
names(pesos_ada)=chartr("M"," ",names(pesos_ada))
a=as.numeric(names(pesos_ada))
P=pesos_ada/100
Medida_ada=BD[a[1]+3]*P[1]
for(i in 2:length(a))
{
Medida_ada=Medida_ada+BD[a[i]+3]*P[i]
}
pred_ada <- prediction(Medida_ada, BD[2])
perf_ada <- performance(pred_ada,"tpr","fpr")
auc_3<- performance(pred_ada,"auc")
auc_3<- unlist(slot(auc_3, "y.values"))
auc_3
```

23.

```
# k=[1...27]
BDada=cbind(BD[2],BD[((k-1)*26+4):((k-1)*26+29)])
BDada$Caso<-factor(BDada$Caso)
adaboost<-boosting(Caso~., data=BDada, boos=TRUE, mfinal=5)
summary(adaboost)
errorevol(adaboost,BDada)
Adaboost<-predict(adaboost,BDada)
Adaboost=as.numeric(as.character(Adaboost$class))
pesos_ada=cbind(adaboost$importance)
names(pesos_ada)=names(adaboost$importance)
pesos_ada=pesos_ada[pesos_ada>0]
names(pesos_ada)=chartr("M"," ",names(pesos_ada))
a=as.numeric(names(pesos_ada))
P=pesos_ada/100
Medida_ada=BD[a[1]+3]*P[1]
for(i in 2:length(a))
{
Medida_ada=Medida_ada+BD[a[i]+3]*P[i]
}
```

```
pred_ada <- prediction(Medida_ada, BD[2])
perf_ada <- performance(pred_ada,"tpr","fpr")


#esp e sens
perf_ada2<- performance(pred_ada,"sens","spec")
esp= unlist(slot(perf_ada2, "x.values"))
sens= unlist(slot(perf_ada2, "y.values"))
soma=esp+sens
max(soma)
espmax=esp[which.max(soma)]
sensmax=sens[which.max(soma)]
sensmax
espmax
auc_3<- performance(pred_ada,"auc")
auc_3<- unlist(slot(auc_3, "y.values"))
auc_3
```

24.

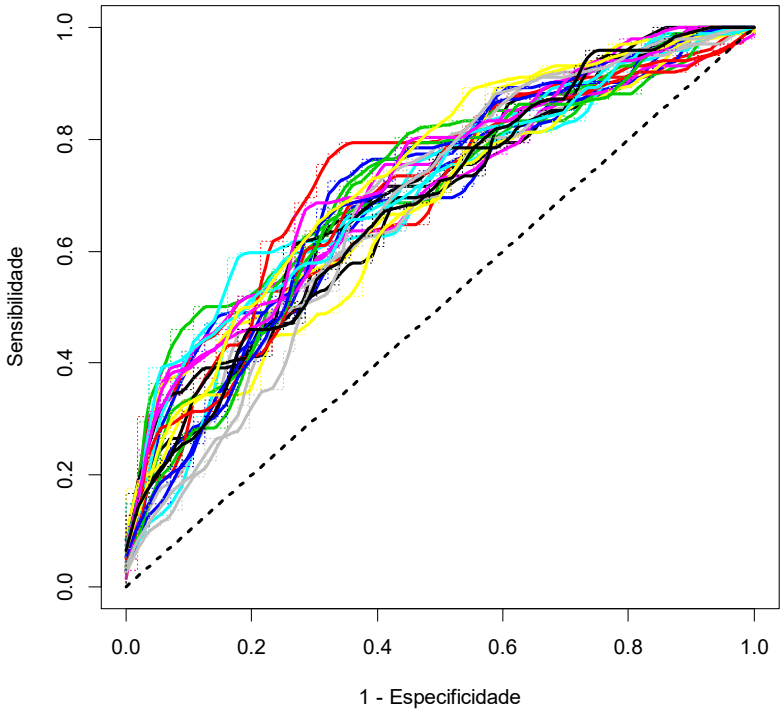
```
# k=[1...27]
BDada=cbind(BD[2],BD[((k-1)*26+4):((k-1)*26+29)])
BDada$Caso<-factor(BDada$Caso)
adaboost<-boosting(Caso~., data=BDada, boos=TRUE, mfinal=5)
summary(adaboost)
errorevol(adaboost,BDada)
Adaboost<-predict(adaboost,BDada)
Adaboost=as.numeric(as.character(Adaboost$class))
pesos_ada=cbind(adaboost$importance)
names(pesos_ada)=names(adaboost$importance)
pesos_ada=pesos_ada[pesos_ada>0]
names(pesos_ada)=chartr("M"," ",names(pesos_ada))
a=as.numeric(names(pesos_adaP=pesos_ada/100
Medida_ada=BD[a[1]+3]*P[1]
for(i in 2:length(a))
{
Medida_ada=Medida_ada+BD[a[i]+3]*P[i]
}
pred_ada <- prediction(Medida_ada, BD[2])
perf_ada <- performance(pred_ada,"tpr","fpr")


#esp e sens
perf_ada2<- performance(pred_ada,"sens","spec")
esp= unlist(slot(perf_ada2, "x.values"))
sens= unlist(slot(perf_ada2, "y.values"))
dist=sqrt((1-esp)^2+(sens-1)^2)
min(dist)
```

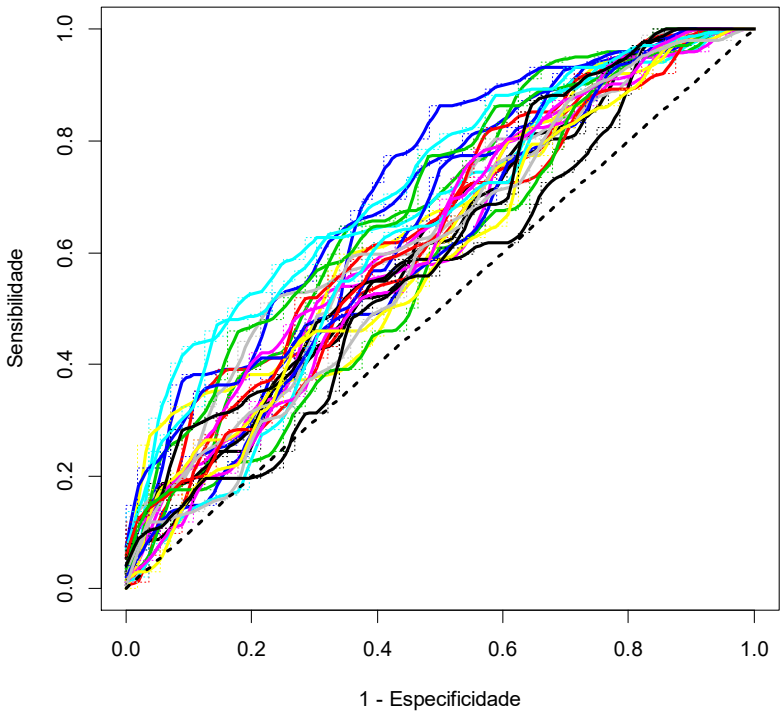
```
espmin=esp[which.min(dist)]
sensmin=sens[which.min(dist)]
sensmin
espmin
auc_3<- performance(pred_ada,"auc")
auc_3<- unlist(slot(auc_3, "y.values"))
auc_3
```

Anexo B - Gráficos

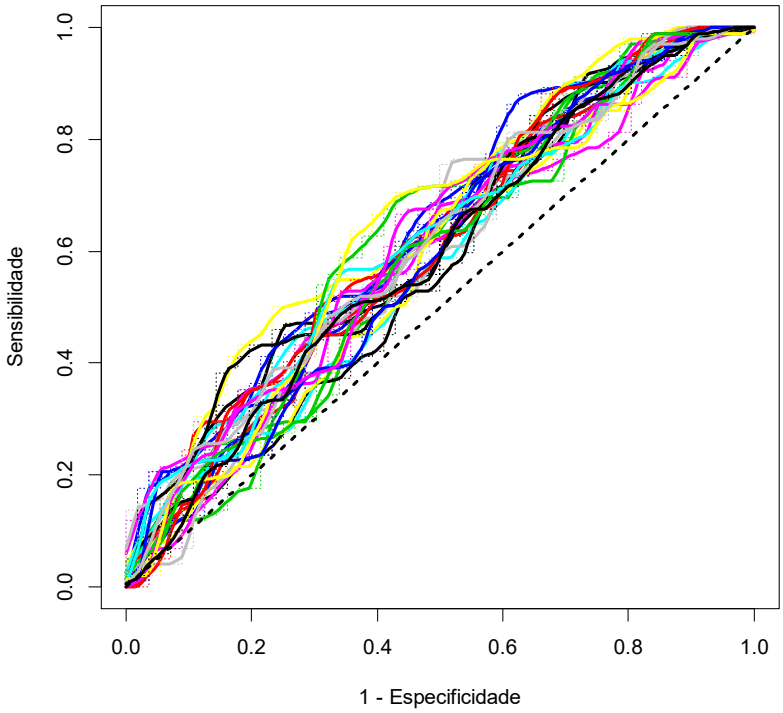
1.


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1 - Especificidade

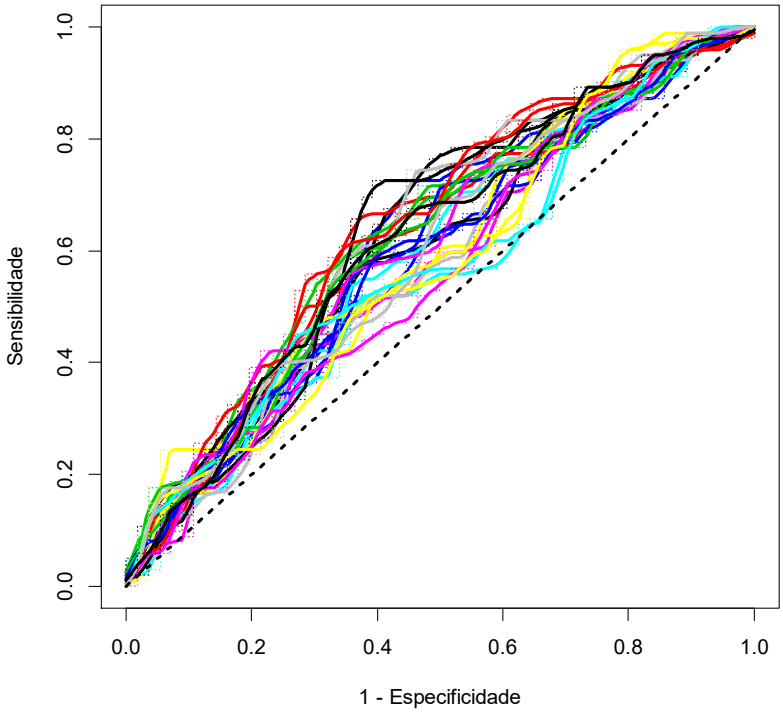
2.


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1 - Especificidade

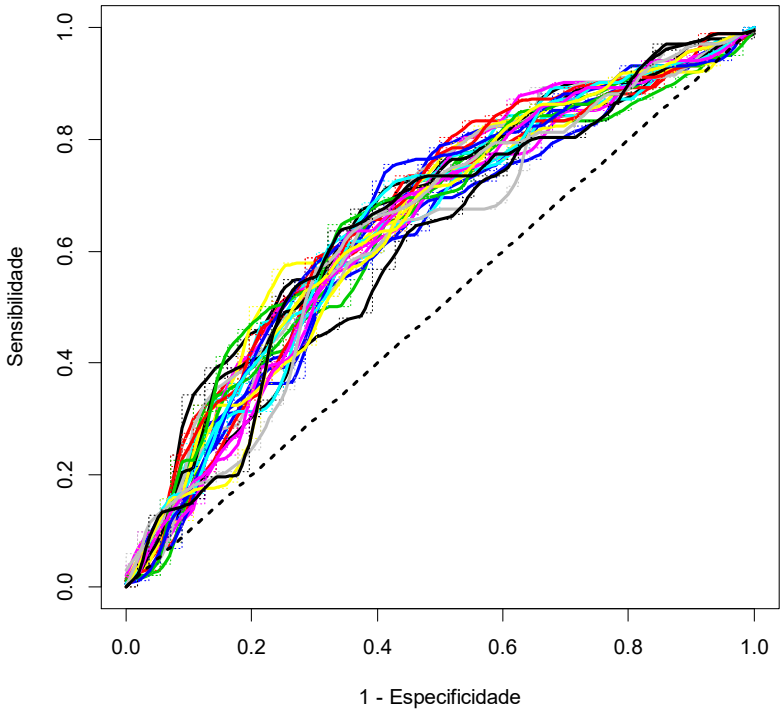
3.


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1 - Especificidade

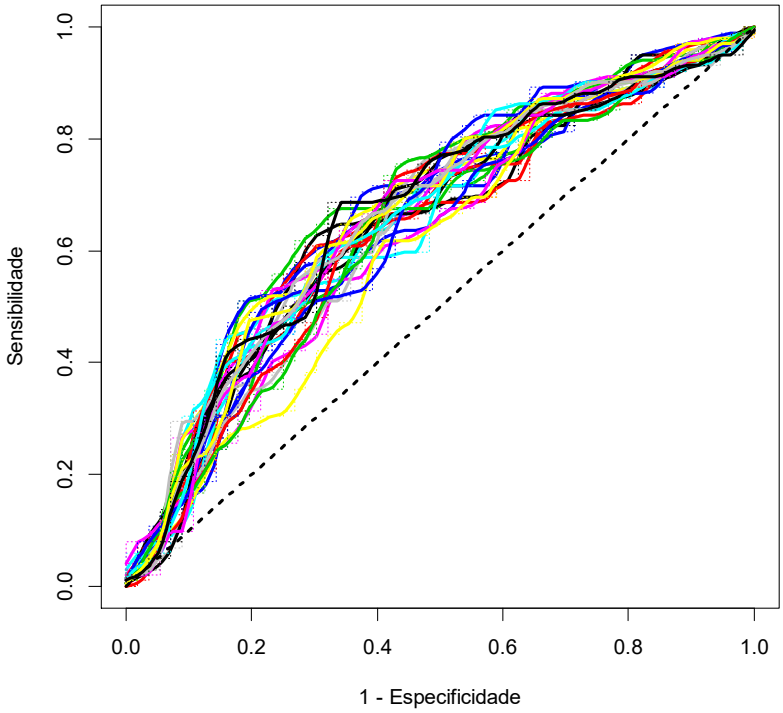
4.


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1 - Especificidade

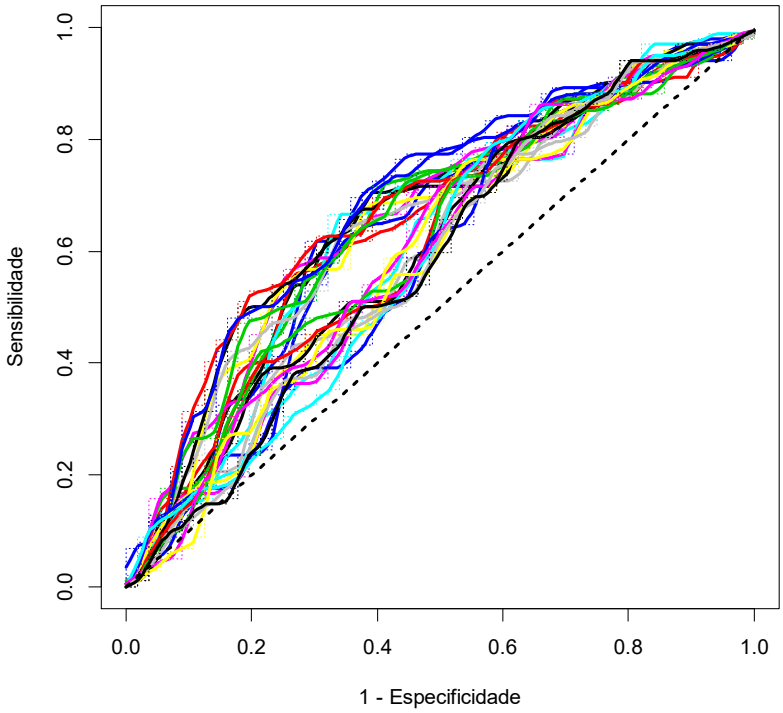
5.


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1 - Especificidade

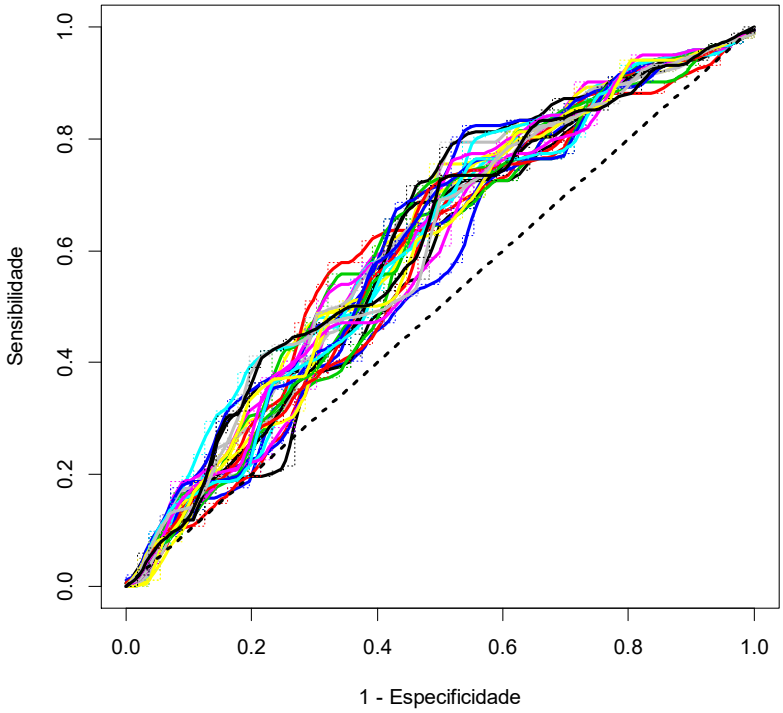
6.


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1 - Especificidade

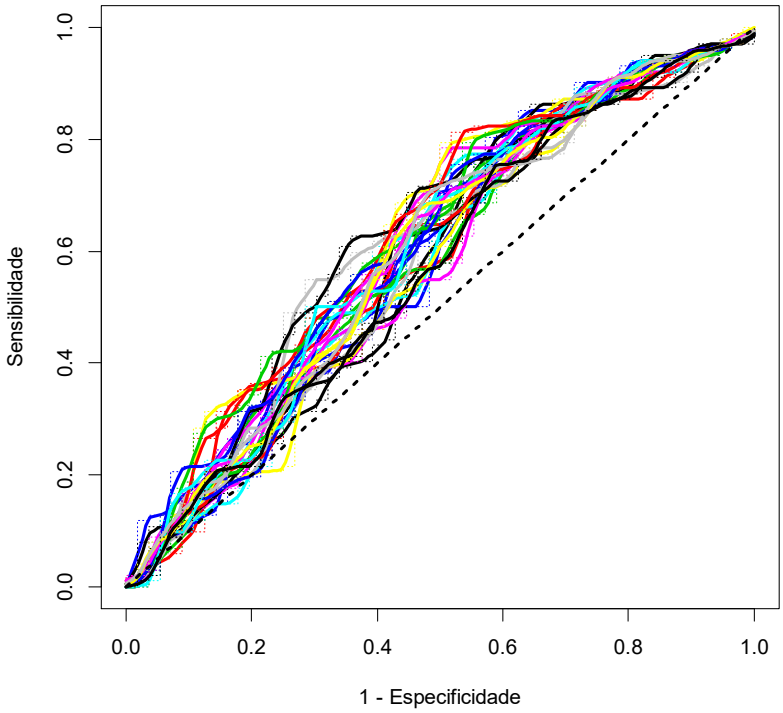
7.


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1 - Especificidade

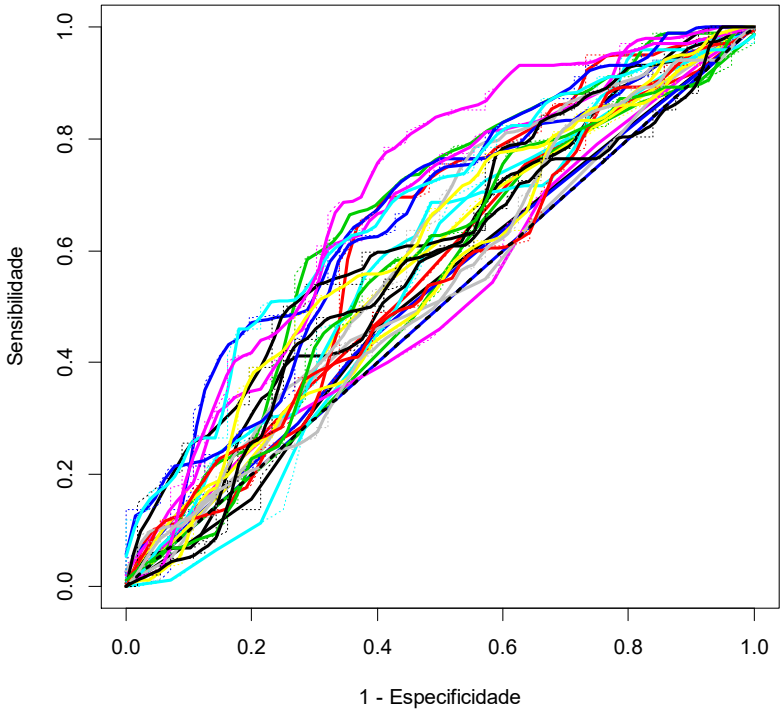
8.


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1 - Especificidade

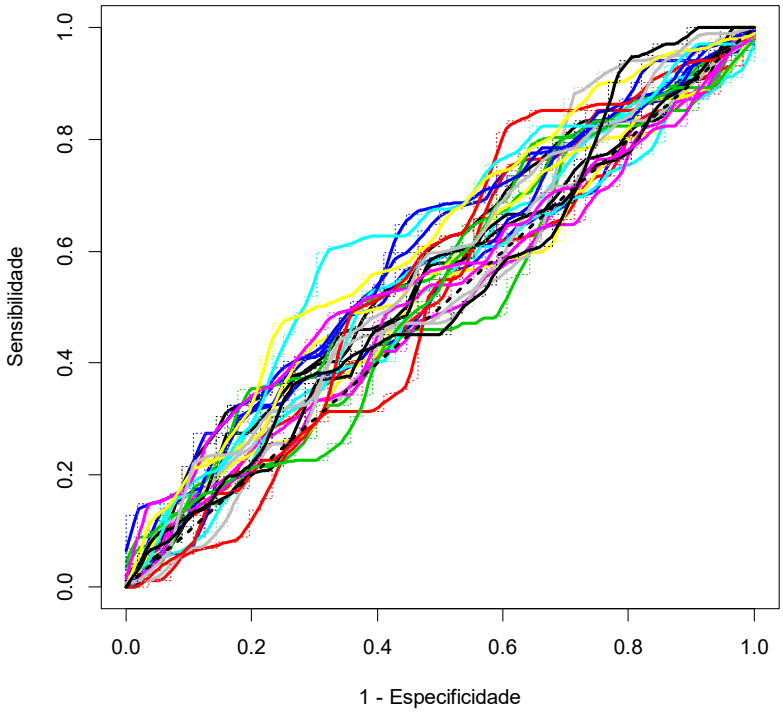
9.


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1 - Especificidade

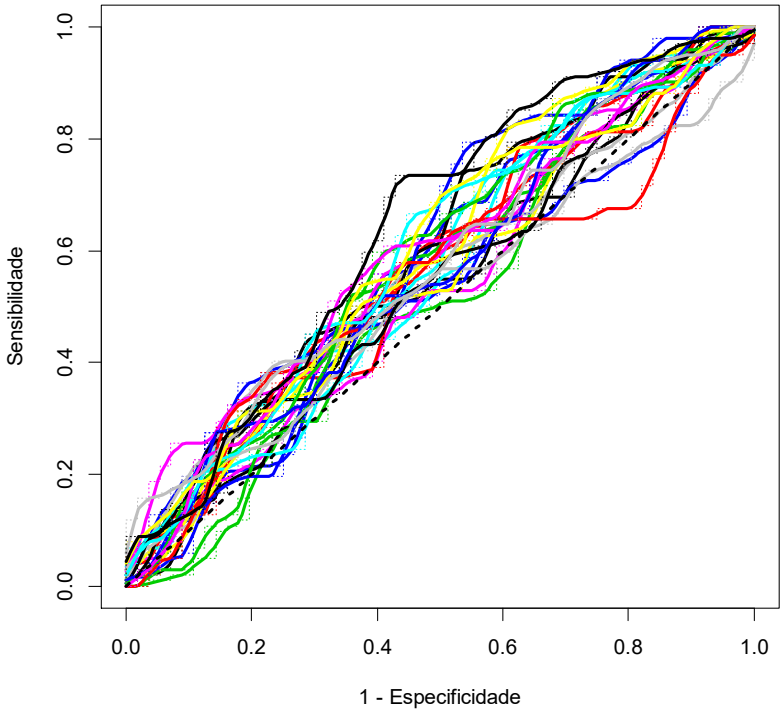
10.


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1 - Especificidade

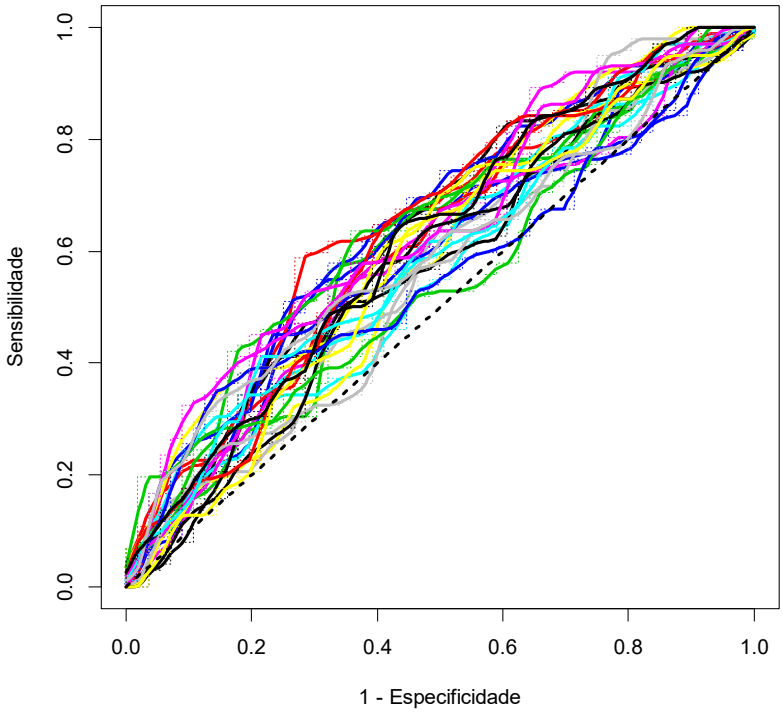
11.


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1 - Especificidade

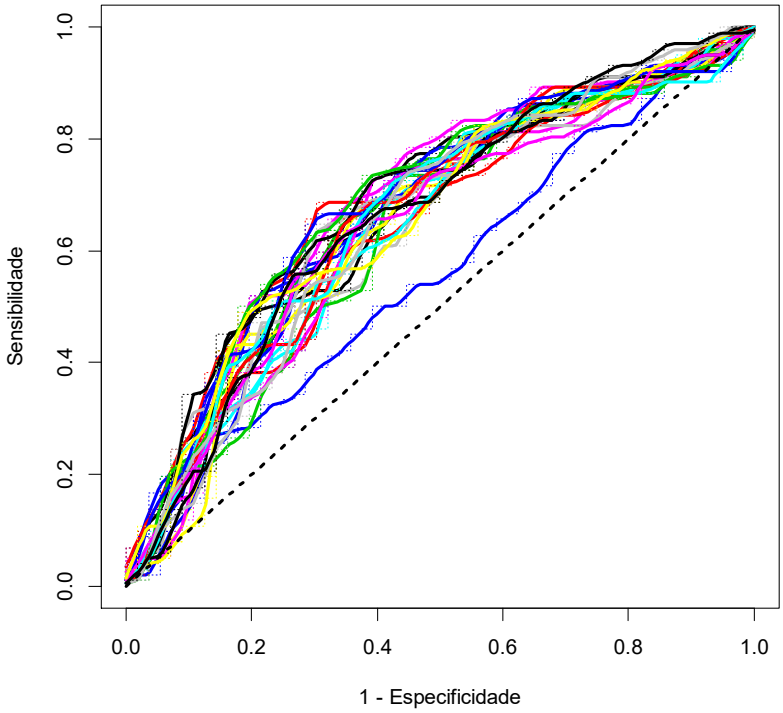
12.


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1 - Especificidade

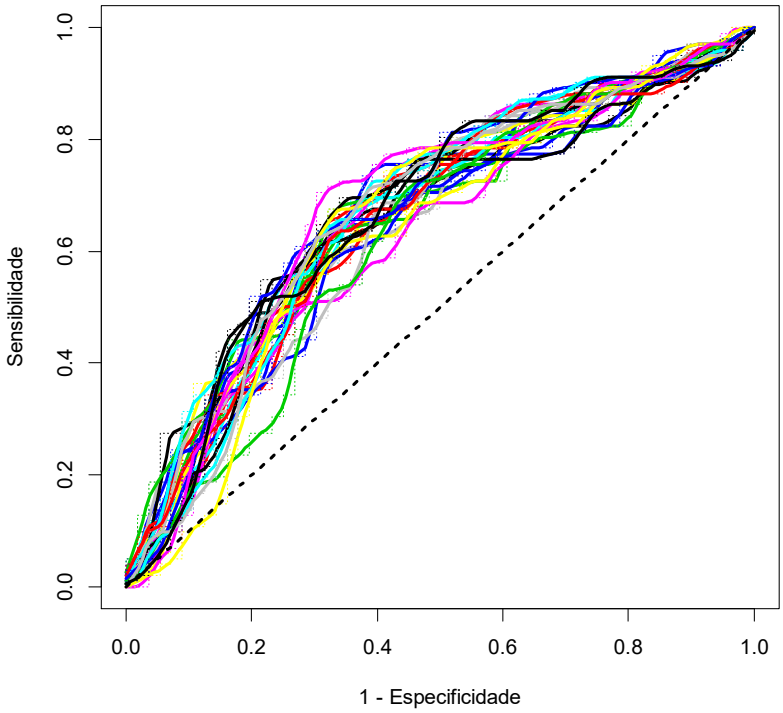
13.


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1 - Especificidade

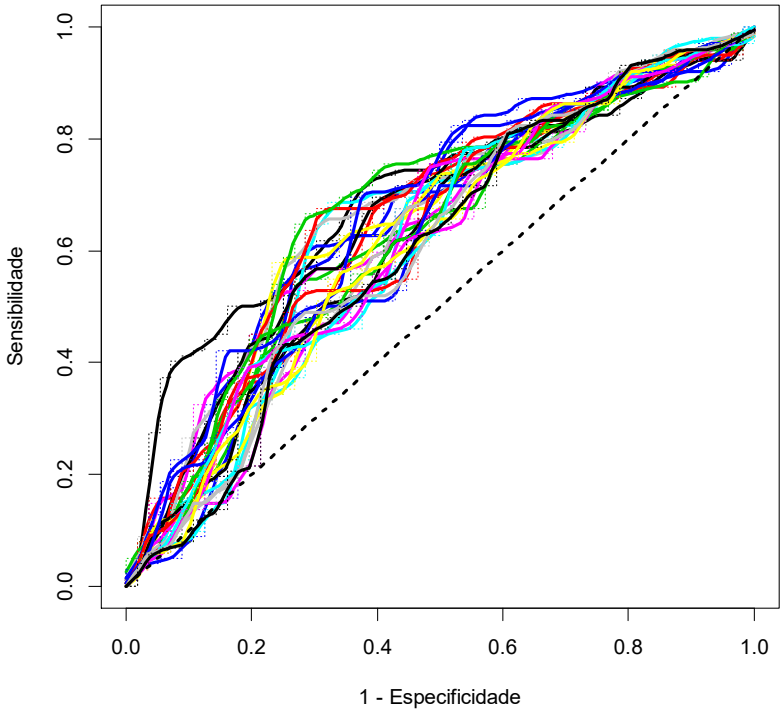
14.


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1 - Especificidade

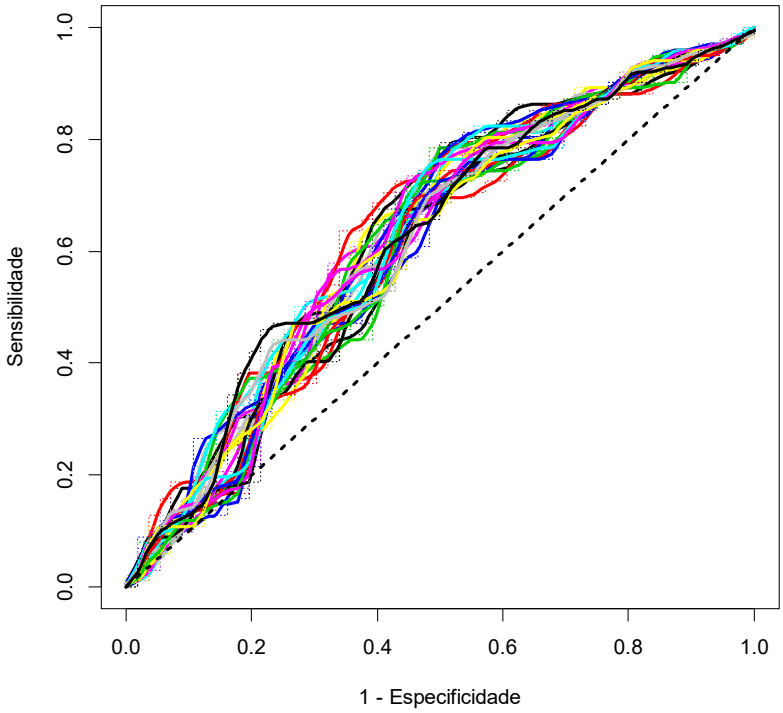
15.


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1 - Especificidade

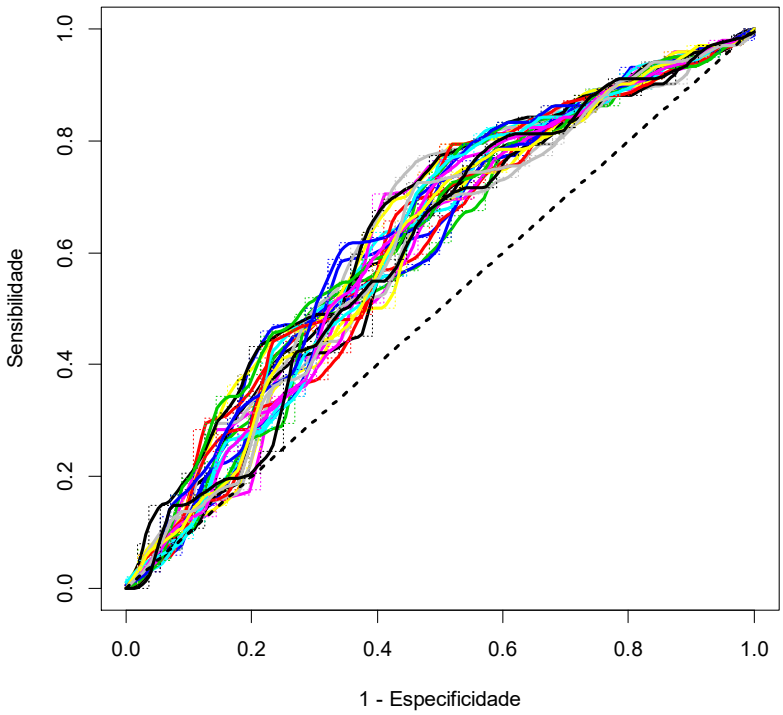
16.


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1 - Especificidade

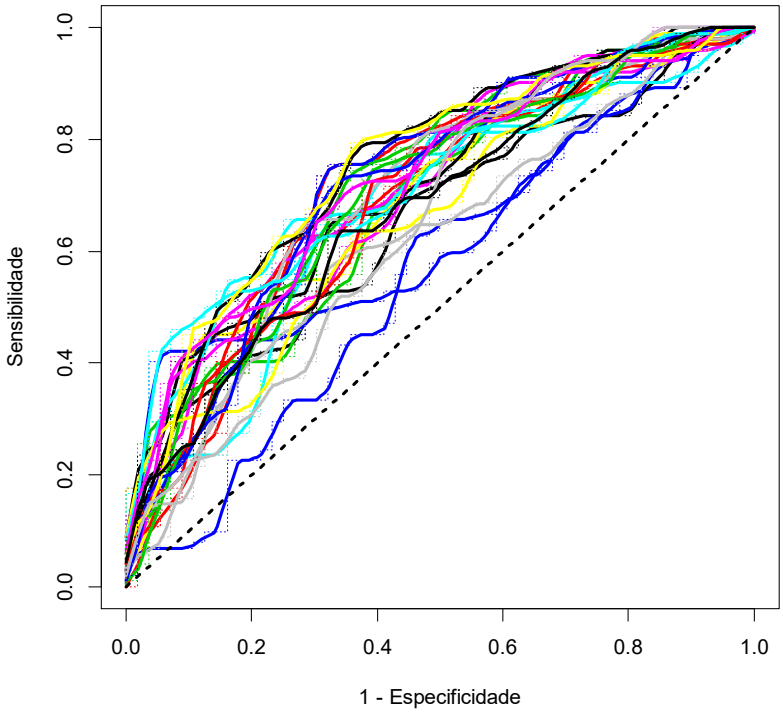
17.


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1 - Especificidade

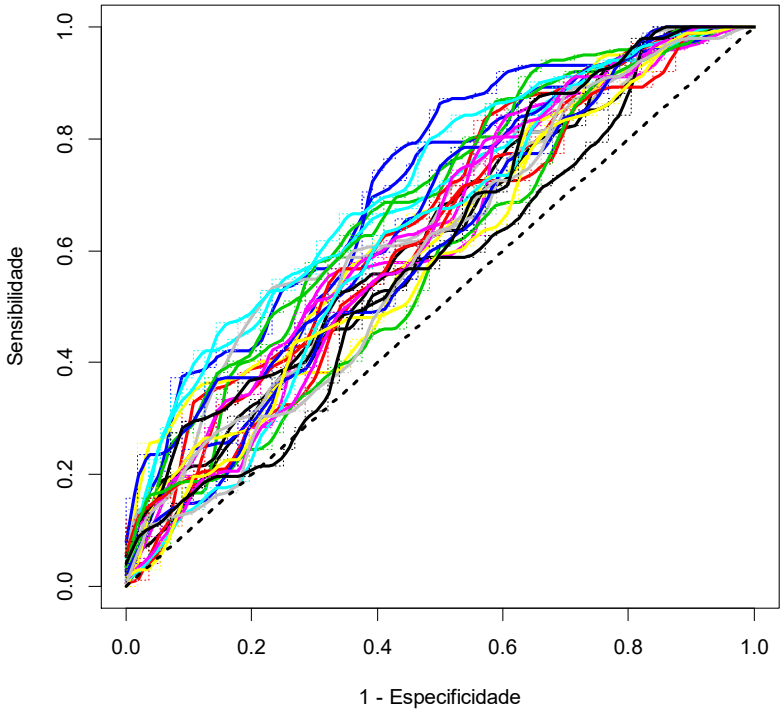
18.


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1 - Especificidade

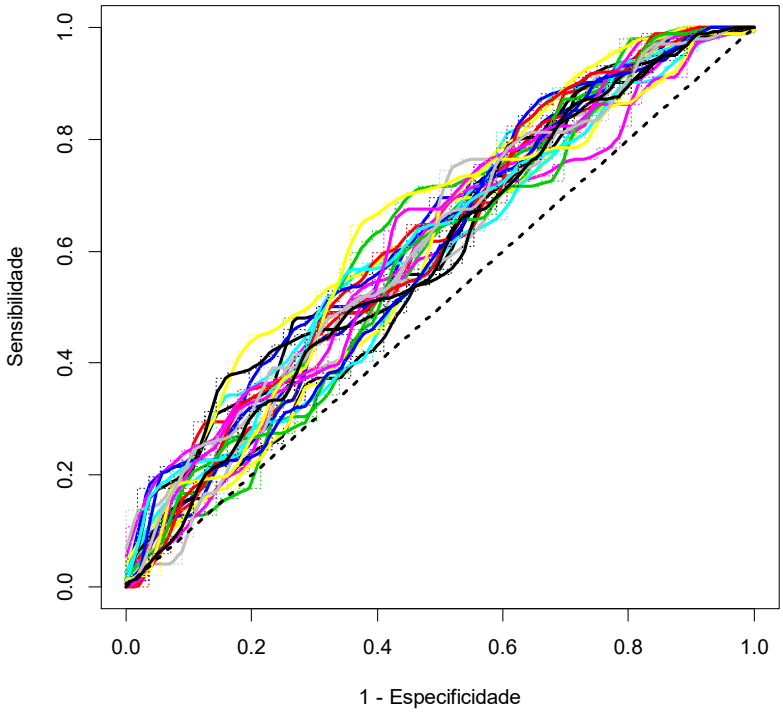
19.


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1 - Especificidade

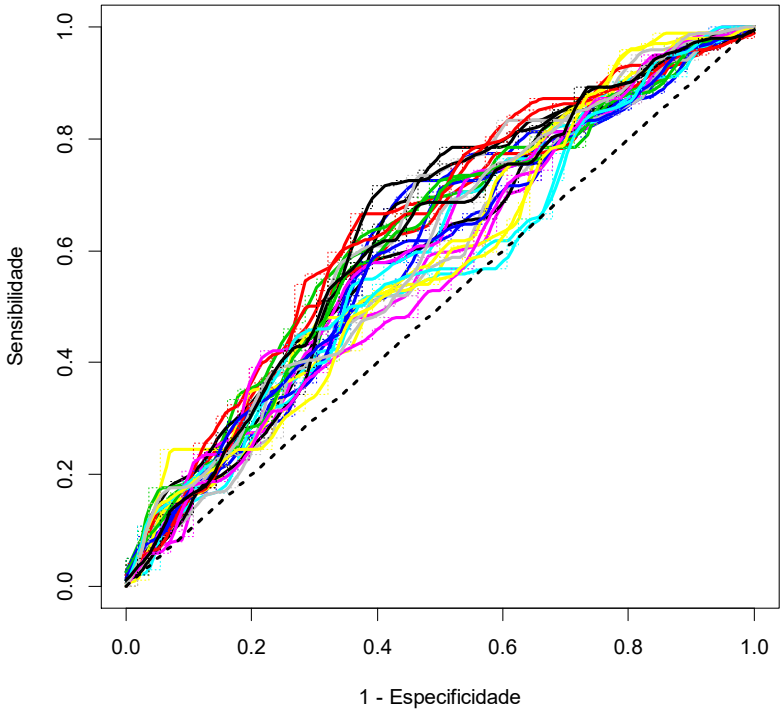
20.


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1 - Especificidade

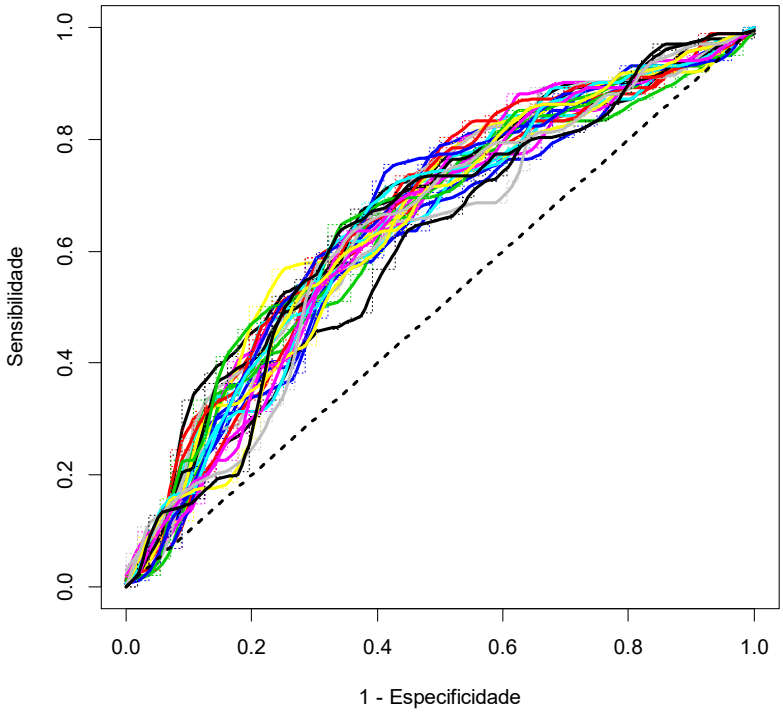
21.


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1 - Especificidade

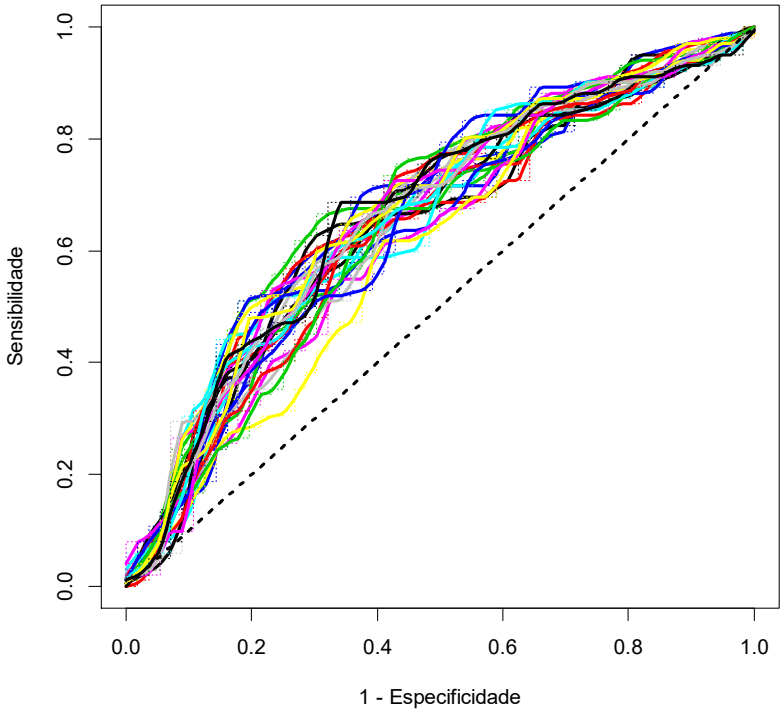
22.


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1 - Especificidade

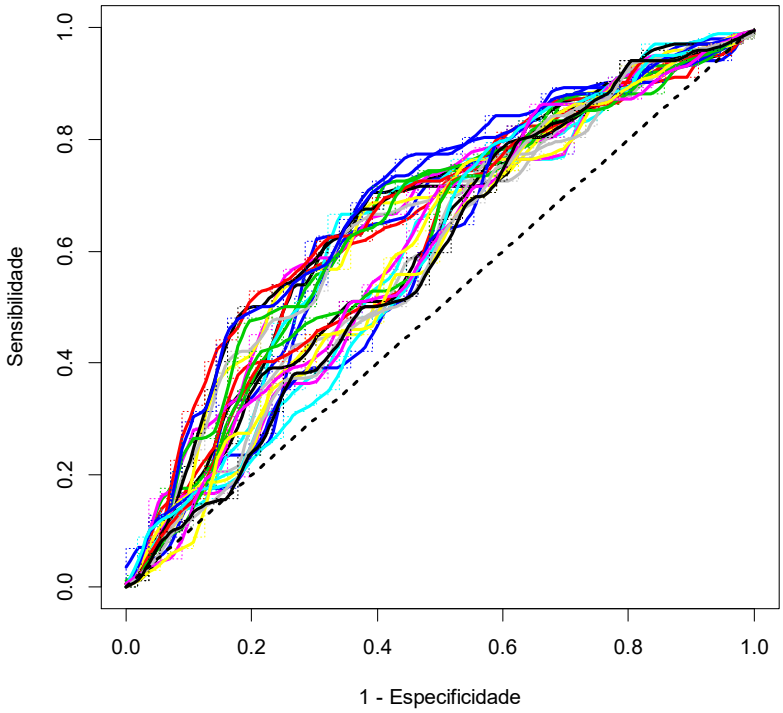
23.


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1 - Especificidade

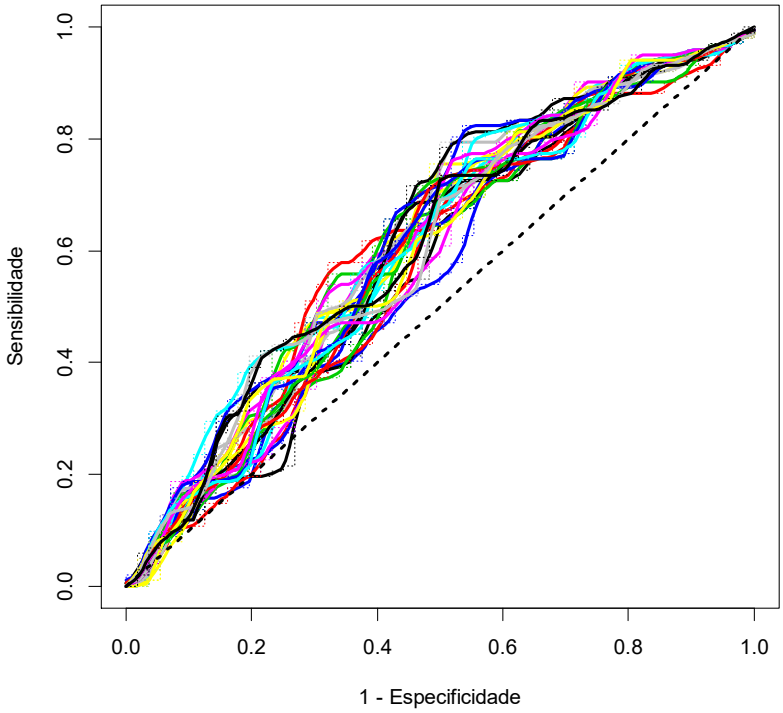
24.


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1 - Especificidade

25.


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1 - Especificidade

26.


Sensibilidade
0.0
0.2
0.4
0.6
0.8
1.0
1 - Especificidade

27.