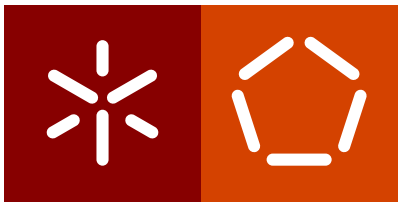


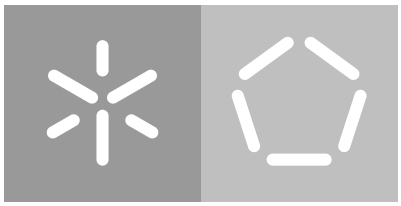
Universidade do Minho
Escola de Engenharia
Departamento de Informática

Augusto Daniel Teixeira Moreira

Desenvolvimento de um programa para comparação de curvas ROC

para amostras independentes e amostras relacionadas

Março 2018



Universidade do Minho
Escola de Engenharia
Departamento de Informática

Augusto Daniel Teixeira Moreira

Desenvolvimento de um programa para comparação de curvas ROC

para amostras independentes e amostras relacionadas

Tese do
Mestrado em Bioinformática

Trabalho efetuado sob a orientação da
Professora Doutora Ana Cristina Braga

Março 2018

AGRADECIMENTOS

Dado este trabalho como concluído, marca assim, o fim da minha formação académica. Desde já, agradeço a todos aqueles que de uma certa maneira, me ajudaram a completar os meus objetivos.

Em primeiro lugar à minha família. Mãe, pai e irmão. Sem este suporte familiar, tudo se tornaria mais difícil, eles sabem que são as pessoas mais importantes da minha vida. Muito obrigado, pelo apoio, confiança e solidariedade, que me deram ao longo da minha formação académica.

À minha orientadora, Professora Ana Cristina Braga, pelo apoio incondicional que me deu na realização desta dissertação. Não só pela ajuda e paciência, mas principalmente pela motivação. Jamais irei esquecer as suas palavras de encorajamento, em momentos difíceis pelos quais passei. São esses pequenos gestos que fazem grandes mudanças, desejando-lhe desde já, muitos sucessos a todos os níveis.

Ao Paulo Ferreira, pela ajuda que me deu na produção do logótipo do meu programa.

Aos meus amigos e colegas de turma, que de uma forma ou de outra, me ajudaram a atingir os meus objetivos. Em especial, ao Pedro Silva e ao João Luís, que estiveram presentes quando mais precisava.

E por último ao meu falecido avô, Joaquim Teixeira, por toda a educação e amor que me deu. Dedico este trabalho a ti, sei que és uma estrelinha no céu e que estarás muito orgulhoso de mim.

Viverás sempre no meu coração.

RESUMO

A análise ROC (Receiver Operating Characteristic) tem vindo a ganhar muita popularidade, principalmente na área da medicina, dado que é uma ferramenta útil para avaliar e especificar problemas no desempenho de um indicador de diagnóstico.

A área abaixo da curva ROC (AUC) é um indicador que pode ser utilizado para comparação de duas ou mais curvas ROC.

Este trabalho, surgiu da necessidade de existência de softwares que permitem o cálculo das medidas necessárias para comparação de sistemas com base nas curvas ROC. Existem vários softwares que efetuam o cálculo de medidas associadas à análise ROC, no entanto apresentam algumas lacunas, nomeadamente no que diz respeito à comparação para amostras independentes com diferentes dimensões e na comparação de duas curvas ROC quando estas se intersetam.

Neste trabalho é apresentado uma nova aplicação que se designa por CERCUS. Esta foi desenvolvida usando a linguagem de programação JAVA e destaca-se pela possibilidade de comparar duas ou mais curvas ROC.

Este programa tem como principal intuito o cálculo de várias estimativas ROC, usando os diferentes métodos sugeridos no desenrolar do trabalho e fazer a comparação de curvas ROC, mesmo que haja interseção, quer para amostras independentes ou amostras emparelhadas. Permite ainda, a representação no plano unitário da curva ROC empírica e a área entre as curvas.

ABSTRACT

Receiver Operating Characteristic (ROC) analysis has gained much popularity, especially in the medical field, as it is a useful tool to assess and specify problems in the performance of a diagnostic indicator.

The area below the ROC curve (AUC) is an indicator that can be used to compare two or more ROC curves.

This work emerged from the need for software to allow the calculation of the necessary measurements to compare systems based on ROC curves.

There are several software that perform the calculation of measures related to ROC analysis, however they present some gaps, particularly as regards the comparison for independent samples with different dimensions and in comparing two ROC curves where they intersect.

In this work a new application is presented that is denominated by CERCUS. This was developed using the programming language JAVA and stands out by the possibility of comparing two or more ROC curves.

The main purpose of this program is the calculation of several ROC estimates, using the different methods suggested along in the dissertation and comparing ROC curves, even if there is an intersection, for independent samples or paired samples. It also allows the representation in the unit plane of the empirical ROC curve and the area between the curves.

CONTEÚDO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE SIGLAS E ACRÓNIMOS

A

AUC Area under Curve.

C

CERCUS Comparison Empirical Roc Curves Cross.

F

FFN Fração de Falsos Negativos.

FFP Fração de Falsos Positivos.

FN Falsos Negativos.

FP Falsos Positivos.

FVN Fração de Verdadeiros Negativos.

FVP Fração de Verdadeiros Positivos.

R

ROC Receiver Operating Characteristic.

S

SE Standard Error.

V

VN Verdadeiros Negativos.

VP Verdadeiros Positivos.

VPP Valores Preditivos Positivos.

1

INTRODUÇÃO

1.1 CONTEXTO E MOTIVAÇÃO

Receiver Operating Characteristic (ROC) surgiu entre 1950 e 1960 e é uma análise que emergiu da teoria da decisão, mais concretamente, na teoria de deteção de sinal (Braga and Oliveira, 2003; Fawcett, 2006). Esta análise apareceu como resultado da necessidade de identificar e diferenciar num operador de radar, um sinal fidedigno (aliados, inimigos) de um ruído (nuvens, aves, etc).

Desde esta altura a análise ROC tem ganhado muita popularidade porque para além de ser uma ferramenta útil para avaliar o desempenho de um indicador, consegue comparar diferentes indicadores e selecionar de uma forma prática um limiar ótimo, que representa a maximização das decisões corretas (Cheam and McNicholas, 2014).

Esta metodologia tem sido aplicada a várias áreas científicas e no campo da medicina tem sido um fator importante em decisões médicas, bem como em áreas da epidemiologia, testes de diagnóstico, radiologia e bioinformática (Hajian-Tilaki, 2013).

O gráfico da curva ROC no plano unitário é uma técnica que pode ser usada para organizar e selecionar classificadores avaliando o seu desempenho. Esta técnica consiste em uma representação gráfica bidimensional que tem como eixo dos "x", "1- especificidade" e no eixo dos "y", "sensibilidade", que variam de 0 a 1 (Fawcett, 2006; Braga, 2000). Em termos de dados em medicina, a sensibilidade corresponde à probabilidade de uma doença estar presente, quando na realidade o individuo está doente e a especificidade corresponde à probabilidade de excluir a doença, quando na realidade esta está ausente (Braga and Oliveira, 2003).

Para comparação de duas curvas ROC, existe um método que obtém, a partir do gráfico da curva ROC, um escalar que representa a sua performance denominado área abaixo da curva (AUC). Como a AUC é uma porção do plano unitário, os seus valores variam entre 0 e 1. Se traçar uma linha na diagonal a partir da origem neste plano, esta representa um valor de AUC de 0,5 , por isso nenhum classificador realista deve ter uma AUC inferior a 0,5 (Fawcett, 2006). Por este motivo, na prática, o valor de AUC varia entre 0,5 e 1,0. Hanley and McNeil (1982) conseguem comparar duas curvas ROC, através de uma estatistica Z,

utilizando o indicador da AUC para os dois sistemas a comparar. O cálculo deste indicador pode ser obtido pela área do trapézio ou pela estatística de Wilcoxon, onde as propriedades estatísticas desta podem ser usadas para prever propriedades da AUC da curva ROC. Para a comparação de curvas ROC que se intersetam, Braga et al. (2005) apresentam uma metodologia que permite esta análise, determinando áreas parciais em diferentes regiões do espaço ROC. Utilizando uma metodologia baseada na comparação de curvas de Pareto em otimização multi objetivo o 'Comp2ROC', pacote desenvolvido em R (Braga et al., 2016), é o resultado desta metodologia.

A análise através das curvas ROC é importante em diferentes campos de aplicabilidade, no entanto existem poucas aplicações disponíveis para sistematizar esta análise, nomeadamente no que diz respeito à representação gráfica e comparação de dois sistemas. A crescente utilização desta metodologia em diferentes áreas, como por exemplo na área da medicina, requer a existência de uma ferramenta única que englobe as mais importantes metodologias do estudo das curvas ROC. O desenvolvimento de um software simples e intuitivo capaz de fazer a analise através das curvas ROC usando as funcionalidades do Comp2ROC é a grande motivação deste trabalho.

1.2 OBJETIVOS

O principal objetivo deste trabalho consiste em desenvolver um programa em JAVA para comparação de dois sistemas com base em curvas ROC que se intersetam ou não. Com base neste objetivo principal, foram delineados os seguintes objetivos específicos:

- permitir analisar amostras independentes e emparelhadas;
- traçar as curvas ROC empíricas no plano ROC unitário assim como os gráficos de diferença entre áreas em função das inclinações das linhas de amostragem e os resultados dos testes de comparação das curvas;
- identificar regiões da curva onde existe melhor desempenho;
- guardar gráficos e resultados em formatos específicos.

1.3 ESTRUTURA DA DISSERTAÇÃO

Esta dissertação está organizada em 6 capítulos.

No **Capítulo 1** da dissertação é apresentada uma introdução ao tema com o respetivo contexto e motivação, os objetivos assim como a estrutura da mesma.

No **Capítulo 2** apresenta-se o estado da arte, passando por uma revisão bibliográfica das metodologias das curvas ROC, onde é descrito também um pouco da sua história e os principais conceitos e definições.

Será também efetuada uma breve revisão dos programas estatísticos para análise ROC.

O **Capítulo 3** retrata as metodologias aplicadas no desenvolvimento do software. É feita uma pequena descrição da linguagem JAVA, da biblioteca Rserve, da base de dados utilizada e dos requisitos que o novo programa deve ter.

No **Capítulo 4** é explicada a abordagem proposta, descrevendo a ferramenta informática que foi desenvolvida. Retrata a sua apresentação e explicação detalhada de todos os ícones e funcionalidades, com uma breve explicação de como o programa compara duas ou mais curvas ROC.

O **Capítulo 5** analisa os resultados obtidos pelo aplicativo, comparando-os com outros softwares disponíveis. Termina com uma breve discussão e conclusão dos resultados obtidos.

No **Capítulo 6** estão expostas as principais conclusões com o desenvolvimento do presente trabalho e uma pequena análise do trabalho desenvolvido. O capítulo termina sugerindo aspetos que podem ser melhorados no programa.

2

METODOLOGIA DAS CURVAS ROC

2.1 HISTÓRIA DAS CURVAS ROC

A análise ROC teve origem no inicio da década de 50 com a combinação da teoria de deteção de sinal e teoria da decisão estatística (Harvey, 2011; Hajian-Tilaki, 2013). Uma das primeiras aplicações foi na deteção de sinais em radar. Neste caso, uma experiência origina dois tipos de eventos: um evento de sinal e um evento de ruído. Dado que os eventos de sinal tendem a produzir uma forte impressão em relação aos eventos de ruído, o observador consegue inferir que tipo de evento está presente (Bamber, 1975; Stanislaw and Todorov, 1999). Para classificar essa decisão em cada acontecimento, Bamber (1975) usa *"sim-não"* no problema de deteção do sinal: *"sim"* na presença do sinal e *"não"* na ausência do sinal (presença do ruído).

Um exemplo muito prático para explicar este processo é a capacidade de identificação e diferenciação num operador de radar, um sinal fidedigno (aliados, inimigos) de um ruído (nuvens, aves, etc) (Collinson, 1998). Na área da psicologia fazem também um aproveitamento desta metodologia na determinação da relação entre os atributos da experiência psicológica com as propriedades dos estímulos físicos (Green and Swets, 1966). Neste caso a decisão dos psicólogos baseia-se na deteção de um fraco sinal causado por algum evento sensorial.

Em meados da década de 60 a análise ROC é utilizada em grande escala devido à necessidade de fazer avaliação do desempenho de testes diagnósticos (Lusted, 1971). Lusted (1971) afirma que este método pode ser adotado para decisão médica colmatando a limitação da escolha de um único par de especificidade e sensibilidade.

A análise ROC tem sido uma ferramenta extremamente útil na avaliação de um teste de diagnóstico, permitindo a comparação de diferentes testes e selecionando de uma forma prática um limiar ótimo (Cheam and McNicholas, 2014).

No campo da medicina, esta metodologia tem sido aplicada a várias áreas científicas, sendo um fator importante em decisões médicas, assim como em áreas da epidemiologia, testes de diagnóstico, radiologia e bioinformática (Hajian-Tilaki, 2013).

2.2 CONCEITOS E DEFINIÇÕES

Um teste de diagnóstico ou classificador, é um mapeamento de certas instâncias (estado da variável) para a previsão de classes/grupos onde o limite de classificação deve ser determinado por um valor (Fawcett, 2006). Sob o ponto de vista processual pode ser algo simples mas também complexo. Por exemplo, um teste pode envolver apenas um passo que resulta em dois resultados possíveis (positivo ou negativo), como pode envolver uma ampla sequência de procedimentos onde o resultado pode ter um vasto leque de possíveis classificações. A implantação de um classificador deve ser pré-condicionada tendo em conta a praticabilidade e o beneficio de um teste para a classificação ou previsão dos resultados (Yanyu, 2010).

A exatidão de um teste diagnóstico, em termos médicos, é a correta classificação de uma doença como estando presente (ou não) numa dada população. Resulta na capacidade de detetar corretamente uma condição quando está verdadeiramente presente e excluir a condição quando está realmente ausente, comparando os resultados com o estado real da condição. A natureza das variáveis dos testes de diagnóstico, pode ser ordinal ou contínua (Martinez et al., 2003). Enquanto que as variáveis contínuas podem tomar qualquer valor numa escala contínua, como por exemplo a idade e a pressão arterial, nas variáveis ordinais os valores são medidos através de uma ordenação entre as categorias, como por exemplo no estágio da dor (ausência, moderada, severa).

Considerando a classificação binária, onde os resultados serão rotulados como positivo (presença da doença) ou negativo (ausência da doença), τ-"classificador" corresponde à verdadeira condição da doença.

$$\tau = \begin{cases} 0 & \textit{não Doente} \\ 1 & \textit{Doente} \end{cases}$$

O resultado do teste diagnóstico é representado pela variável indicador γ-"instância".

$$\gamma = \begin{cases} 0 & \textit{Negativo para doença} \\ 1 & \textit{Positivo para doença} \end{cases}$$

Dado um classificador e um conjunto de instâncias, constrói-se uma tabela de contingência de 2 por 2 (Tabela 1) que também é conhecida como matriz confusão e serve para calcular medidas de desempenho importantes como, por exemplo, sensibilidade e especificidade (Collinson, 1998).

Tabela 1: Matriz Confusão de classificação dos resultados de um teste diagnóstico.

	τ=0	τ=1
γ=0	Verdadeiros Negativos **VN**	Falsos Negativos **FN** *Erro do tipo I*
γ=1	Falsos Positivos **FP** *Erro do tipo II*	Verdadeiros Positivos **VP**

Observando a Tabela 1 com uma perspetiva estatística, verifica-se que para o classificador binário existem quatro resultados possíveis: *Verdadeiros Positivos (VP)*, *Falsos Positivos (FP)*, *Falsos Negativos (FN)* e *Verdadeiros Negativos (VN)*.

- VP, classificação de um individuo doente quando ocorre um teste com resultado positivo para a presença da doença;
- FP, classificação de um individuo não doente quando ocorre um teste com resultado positivo para a presença da doença;
- FN, classificação de um individuo doente quando ocorre um teste com resultado negativo para a presença da doença;
- VN, classificação de um individuo não doente quando ocorre um teste com resultado negativo para a presença da doença.

O resultado de um teste de diagnóstico pode ser induzido em erro derivado a uma má classificação do individuo pela equipa médica.

Como se pode observar pela análise da Tabela 1, existem dois tipos de erro: erro do tipo I e erro do tipo II. Em estatística estes erros correspondem respetivamente à *Fração de Falsos Positivos (FFP)* e à *Fração de Falsos Negativos (FFN)*, sendo o seu valor obtido seguindo as fórmulas a seguir apresentadas:

$$\textbf{FFP} = P\{\gamma = 1 | \tau = 0\} = \frac{P(\gamma = 1 \cap \tau = 0)}{P(\tau = 0)} = \frac{\textbf{Número de Falsos Positivos (FP)}}{\textbf{Total de Negativos (N)}}$$

e na qual

$$\textbf{Total de Negativos (N)} = \textit{Verdadeiros Negativos (VN)} + \textit{Falsos Positivos (FP)} = \textit{Total não Doentes}$$

$$\textbf{FFN} = P\{\gamma = 0 | \tau = 1\} = \frac{P(\gamma = 0 \cap \tau = 1)}{P(\tau = 1)} = \frac{\textbf{Número de Falsos Negativos (FN)}}{\textbf{Total de Positivos (P)}}$$

com

Total de Positivos (P) = *Verdadeiros Positivos (VP)* + *Falsos Negativos (FN)* = *Total de Doentes*

Se um teste de diagnóstico apresenta a FFP igual a zero e a *Fração de Verdadeiros Positivos (FVP)* igual a um, pode-se concluir que este é ideal para a deteção da doença. Por outro lado, se a FFP for igual a FVP, o teste é inconclusivo para a presença da doença. A sensibilidade (FVP) e a especificidade (1-FFP) são duas medidas de precisão comumente usadas na área médica para avaliar o desempenho de um teste de diagnóstico. Estas medidas estão vigorosamente relacionadas com os conceitos de erros tipo I e II (Pepe, 2004).

2.2.1 *Sensibilidade e Especificidade*

Como referido anteriormente, a sensibilidade (FVP) representa a capacidade de classificar corretamente um dado atributo, e a especificidade ou *Fração de Verdadeiros Negativos (FVN)* representa a capacidade de classificar corretamente a inexistência de um dado atributo (Collinson, 1998; Fawcett, 2006; Gönen, 2001; Braga, 2000). Quanto maior o valor da sensibilidade, maior é a probabilidade de um teste classificar corretamente a presença da doença e vice-versa. Estas medidas são, entre si, independentes. No entanto, sensibilidade e a especificidade, dependem respetivamente dos "doentes" e dos "não doentes", como se pode analisar pelas as seguintes expressões:

$$sensibilidade = FVP = \frac{\textbf{Número de Verdadeiros Positivos (VP)}}{\textbf{Total de Positivos (P)}}$$

$$especificidade = FVN = \frac{\textbf{Número de Verdadeiros Negativos (VN)}}{\textbf{Total de Negativos (N)}}$$

2.2.2 *Exatidão e Precisão*

A exatidão de um teste diagnóstico também pode ser quantificada na eficácia que um teste tem para prever o respetivo resultado. Assim, a Exatidão e a Precisão, são também importantes medidas de desempenho de um teste diagnóstico. Enquanto que a Precisão determina se existe concordância nos resultados quando o teste é repetido várias vezes, a Exatidão determina a capacidade do teste fornecer resultados muito próximos do verdadeiro valor do que está a ser medido (Fawcett, 2006; Collinson, 1998).

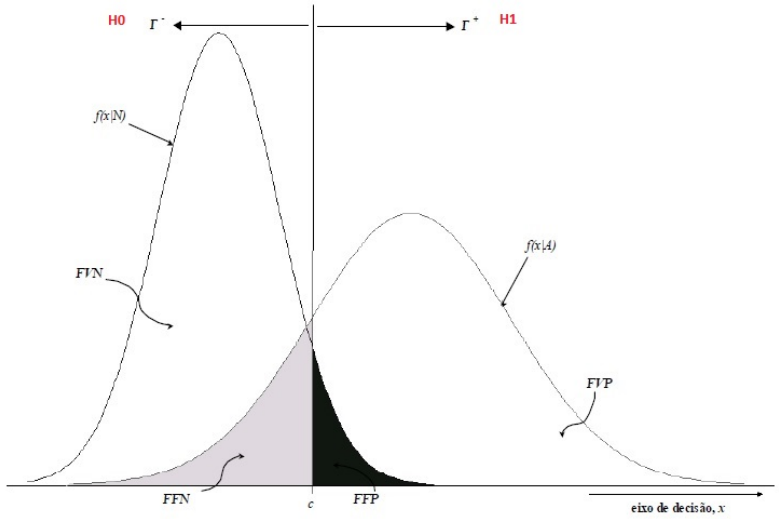


Figura 2.1: Sobreposição de duas distribuições hipotéticas, imagem adaptada de Braga (2000) (Fonte: (Braga, 2000)).

Um teste ideal é aquele que prevê idealmente a condição dos pacientes e para isso a FFP é igual a zero e a FVP é igual a um. Por outro lado, quando FFP=FVP, o teste não é decisivo sobre a presença da doença (Pepe, 2004).

$$Exatidão = \frac{\textbf{Número de Verdadeiros Negativos (VN) + Número de Verdadeiros Positivos (VP)}}{\textbf{Total de Negativos (N) + Total de Positivos (P)}}$$

$$Precisão = \frac{\textbf{Número de Verdadeiros Positivos (VP)}}{\textbf{Total de Negativos (N) + Total de Positivos (P)}}$$

2.2.3 *Testes contínuos de diagnóstico*

Segundo Braga (2000), os testes contínuos de diagnóstico podem ser modelados como um problema de testes de hipóteses em que as variáveis de decisão são relacionadas com as hipóteses nula e alternativa. Em termos médicos, uma dessas hipóteses representa a presença da doença e a outra representa a sua ausência. Este é um método muito utilizado para a previsão de um resultado num teste de diagnóstico onde:

- H_0 = Hipótese nula = ausência da doença
- H_1 = Hipótese alternativa= presença da doença

O valor de corte, c, representado pela linha vertical da Figura 2.1, é um valor importante para a separação dos casos positivos (T^+) dos casos negativos (T^-) (Metz, 1986). Os valo-

res que se apresentam à direita de c, representam a hipótese de o indivíduo ser doente e os valores que se apresentam à esquerda de c, representam a hipótese de o indivíduo ser saudável. Se os valores de c aumentarem, aumenta também a probabilidade de ocorrerem casos positivos (Swets, 1996). Na Figura 2.1 verifica-se que ao diminuir FFP aumenta a FFN.

Associado aos testes de hipóteses, a *sensibilidade*, *especificidade* e os erros do tipo I e II são descritos em função do valor de corte c, como $FVP(c)$, $FVN(c)$, $FFP(c)$ e $FFN(c)$ respetivamente. Considerando um resultado binário τ e uma variável de decisão X, segundo Braga (2000)

$$FVP(c) = sensibilidade = P[X \geq c|\tau = 1]$$

$$FVN(c) = especificidade = P[X \leq c|\tau = 1] = (1 - FFP)$$

$$\textbf{P(Erro do tipo I)} = \alpha = P[rejH_0|H_0] = P(T^+|X_N) = P[x > c|\tau = 0] = FFP$$

$$\textbf{P(Erro do tipo II)} = \beta = P[\overline{rejH_0}|H_1] = P(T^-|X_A) = P[x < c|\tau = 1] = FFN$$

Um valor de corte para além de definir a região de rejeição (define as dimensões dos erros tipo I e II), fixa um par *(sensibilidade, especificidade)*. Estes pares podem ser representados como valores de coordenadas de um gráfico, "y"e "x", dando origem à curva ROC empírica. Resumidamente, em termos de testes de diagnóstico, a representação ROC dá a probabilidade de não rejeitar a Ho.

2.3 ESPAÇO ROC

2.3.1 *Curva ROC*

A representação das curvas ROC pode ser entendida como uma técnica para selecionar classificadores e avaliar o seu desempenho. Nesse sentido, os classificadores produzem um par (FFP, FVP) no espaço ROC, como se pode observar na Figura 2.2. Este par está diretamente relacionado com a variação do valor de corte c, ao longo do eixo de decisão "x"(Fawcett, 2006; Braga, 2000).

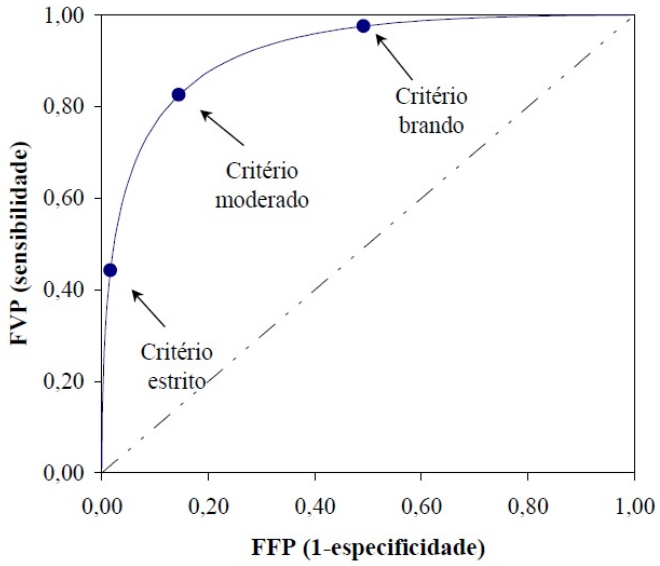


Figura 2.2: Curva ROC e os critérios descritos por Braga (2000) (Fonte: (Braga, 2000))

Uma curva ROC empírica é uma representação gráfica da relação entre a potência de um teste e a probabilidade de se cometer um erro do tipo I, consoante o valor do corte e sob uma perspectiva estatística (Metz, 1986).

Uma importante propriedade da curva ROC é esta ser crescente dado a relação que tem com a sensibilidade e a especificidade. Porém, se a curva for uma diagonal no plano bidimensional será impossível estabelecer qualquer relação entre essas medidas (Metz, 1978).

A importância dos pontos da curva ROC varia entre eles no espaço ROC, onde a sua descriminação é diferente consoante a sua localização. Conforme os critérios que descrevem um ponto na curva ROC, Braga (2000) refere que um *critério "estrito"* é aquele ponto na curva ROC que se situa no canto inferior esquerdo, isto é, aquele que conduz a uma pequena fração de FP com uma pequena fração de VP. Com a progressão dos pontos ao longo do espaço o critério a ser aplicado vai ser diferente, consequente de uma maior fração de FP com uma maior fração de VP. A Figura 2.2 representa graficamente a situação que foi explicada relativa à curva ROC.

Dado que o ponto (0,0) representa uma estratégia que não classifica VP nem FP, o seu oposto (1,1), retrata uma classificação só de VP. Basicamente, os classificadores discretos mais próximos do ponto (0,0) fazem só classificações positivas com uma forte evidencia, onde este comete poucos erros do tipo I. Já os classificadores mais próximos do ponto (1,1) fazem só classificações positivas com pouca evidência, que por sua vez, comete muitos erros do tipo I (Fawcett, 2006).

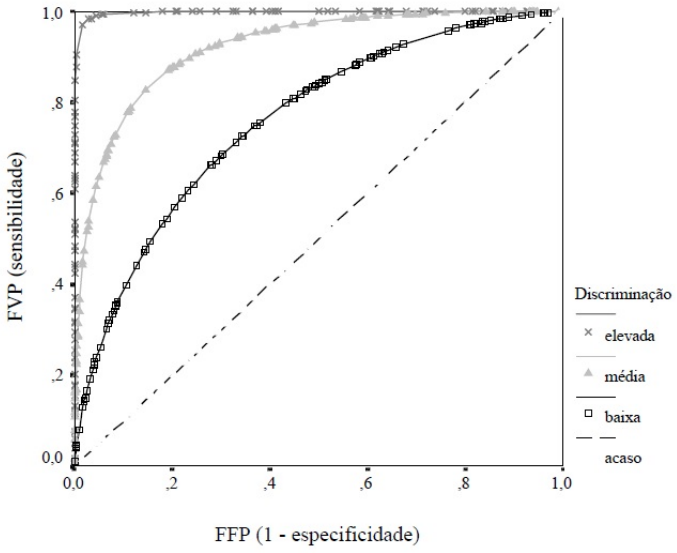


Figura 2.3: Curvas ROC e os três graus de descriminação descritos por Braga (2000) (Fonte: (Braga, 2000))

Por outro lado a linha diagonal y=x, retrata a estratégia de adivinhar aleatoriamente a classe correta, com uma probabilidade associada de 50 % (Fawcett, 2006).

No que diz respeito a uma classificação perfeita no espaço ROC, o ponto (0,1) é o que caracteriza melhor esta situação, com um maior valor de FVP e um menor valor de FFP. Pode-se afirmar então que para dois sistemas de diagnóstico que não se cruzam, o sistema com os valores da curva ROC mais próximos do ponto (0,1) apresenta um maior poder discriminante, porque há um maior valor de FVP do que o valor de FFP. Na Figura 2.3 estão ilustradas três curvas com três graus de descriminação diferentes e a sua diagonal.

2.3.2 *Área abaixo da curva (AUC)*

Para a discriminação de um teste diagnóstico existe um método que obtém a partir de um gráfico ROC, um valor escalar que representa a sua performance denominado a área abaixo da curva, AUC (Fawcett, 2006). Este é um dos índices mais utilizados para retratar a "qualidade" da curva (Hanley and McNeil, 1982; Swets, 1996; Metz, 1986).

Segundo Begg (1991) a AUC é uma medida que pode ser entendida como a probabilidade de um indivíduo doente ter um resultado de maior relevo do que aquele individuo não doente. Resumidamente, este método representa uma propriedade estatística importante, dado que a probabilidade de um classificador escolher, ao acaso, uma instância positiva é maior do que escolher uma instância negativa. Swets (1996) citado por Hanley and

McNeil (1982) refere que a AUC é a probabilidade de detetar se o individuo tem ou não determinado atributo, i.e, a probabilidade de classificar corretamente.

De acordo com Bradley (1997) o desempenho de um classificador num plano bidimensional é dado pela área abaixo da curva, onde produz informações importantes sobre os casos de estudo. Como a AUC é uma porção do plano unitário onde os seus valores variam entre 0 e 1, quando se verifica FFP = FVP, a AUC representa um valor de 0,5 (o classificador não tem poder discriminante). Por isso, nenhum classificador realista deve ter uma AUC inferior a 0,5 (Fawcett, 2006). Por outro lado, quando a AUC atinge o seu valor máximo (AUC = 1), está-se perante um classificador com discriminação perfeita, e a curva está posicionada para o canto superior esquerdo, no qual FVP=1 (Green and Swets, 1966).

Conforme vários autores, entre eles Hanley and McNeil (1982) e Braga (2000), os três métodos mais utilizados para a estimativa do valor escalar da AUC são:

- Estatística de Wilcoxon-Mann-Whitney;
- Regra do Trapézio;
- Área Binormal.

Os métodos atrás descritos são aplicados para valores discretos (Bamber, 1975; Bradley, 1997). Para valores contínuos, a AUC pode ser obtida a partir da função (Bamber, 1975; Pepe, 2004):

$$AUC = \int_0^1 ROC(t)\, dt$$

o que em termos de probabilidade pode ser escrito como

$$AUC = P[X_A > X_N]$$

A estatística de Wilcoxon-Mann-Whitney, W, é usualmente utilizada para testar se os indivíduos que apresentam alguma característica quantitativa x numa população A (doente), tendem a ser maiores do que numa segunda população N (não doente), sem assumir realmente que a característica está distribuída nas duas populações (Hanley and McNeil, 1982). Considerando uma amostra de tamanho n_A a partir de A e uma amostra de tamanho n_N a partir de N o procedimento consiste em fazer todas as comparações possíveis ($n_A \cdot n_N$), entre os valores x_A da amostra n_A e os valores x_N da amostra n_N assinalando cada semelhança

segundo a regra a seguir descrita:

$$S(x_A, x_N) = \begin{cases} 1 & se \quad x_A > x_N \\ \frac{1}{2} & se \quad x_A = x_N \\ 0 & se \quad x_A < x_N \end{cases}$$

A estatística W retrata a média de todos os $S's$ para todas as comparações ($n_A \cdot n_N$):

$$W = \frac{1}{n_A \cdot n_N} \sum_{1}^{n_A} \sum_{1}^{n_N} S(x_A, x_N)$$

que é uma estatística que depende das graduações e não dos valores x, denominada como estatística de Wilcoxon-Mann-Whitney (Hanley and McNeil, 1982).

O resultado de W estará entre 0 e 1 dado a classificação resultante ($S(x_A, x_N)$) estar entre 0, $\frac{1}{2}$ e 1. Conforme descrito anteriormente W será a proporção de x_A maior que x_N (Hanley and McNeil, 1982).

Quando se varia o valor de corte, obtém-se um conjunto de pontos pertencentes à curva ROC e desta forma consegue-se um método simples para calcular a área abaixo da curva ROC: a regra do trapézio (Bradley, 1997). Tendo em conta o gráfico da curva ROC empírica no plano unitário, entre pontos sucessivos da curva, encontram-se representados trapézios. A AUC é obtida, através do somatório da área dos trapézios obtidos pela análise dos pontos FVP e FFP da curva ROC empírica, no espaço unitário, através da expressão:

$$AUC = \sum_{i=1}^{N} \left\{ (FVP_{i-1} \cdot \triangle FFP) + \frac{1}{2}[\triangle FVP \cdot \triangle FFP] \right\}$$

onde:

$$\triangle FVP = FVP_i - FVP_{i-1}$$

$$\triangle FFP = FFP_i - FFP_{i-1}$$

Sendo Φ a função de distribuição da Normal padrão, outra forma de estimar a AUC é através do modelo binormal que é dado por (Braga, 2000)

$$AUC = \Phi\left(\frac{a}{\sqrt{1+b^2}}\right)$$

e na qual

$$a = \frac{\mu_1 - \mu_0}{\sigma_1}$$

$$b = \frac{\sigma_0}{\sigma_1}$$

Nestes dois quocientes:

- σ_0 = desvio padrão da distribuição dos valores de x_N;
- σ_1 = desvio padrão da distribuição dos valores de x_A;
- μ_0 = média da distribuição dos valores de x_N;
- μ_1 = média da distribuição dos valores de x_A.

Uma maneira comum de obter uma estimativa mais adequada da área abaixo da curva ROC é estimar também o erro padrão *Standard Error (SE)* (Jensen et al., 1996). Segundo Hanley and McNeil (1982), esta medida é a mais importante característica dado o interesse de quantificar a variável *W*. Uma estimativa aproximada de *SE(W)* pode ser calculada a partir da AUC da curva ROC:

$$SE(W) = \sqrt{\frac{A(1-A) + (n_A - 1)(Q_1 - A^2) + (n_N - 1)(Q_2 - A^2)}{n_A \cdot n_N}} \tag{1}$$

Na expressão (1)

$$A = AUC$$

$$Q_1 = \frac{A}{(2 - A)}$$

$$Q_2 = \frac{2A^2}{(1 + A)}$$

e, n_A e n_N representam o número de indivíduos doentes e não doentes, respetivamente. A substituição destas expressões na equação (1) conduz ao valor de erro padrão esperado para qualquer valor de A.

2.4 COMPARAÇÃO ATRAVÉS DE CURVAS ROC COM BASE NA AUC

Segundo Pollack and Hsieh (1969) o índice da área abaixo da curva, AUC, é muito importante pois é uma medida não-paramétrica e por consequência não serem necessários pressupostos sobre as distribuições subjacentes aos dados. A visualização dos parâmetros, sensibilidade e especificidade, em gráficos com duas ou mais curvas ROC associadas a diferentes testes diagnósticos contínuos, permitem uma imediata comparação dos seus desempenhos (Martinez et al., 2003), mas se duas ou mais curvas são construidas com base em diferentes testes de desempenho para o mesmo conjunto de dados, é necessário efetuar uma análise estatística das curvas ROC, de forma a obter o teste com um melhor desempenho (Braga, 2000; Hanley and McNeil, 1983; Delong et al., 1988). Para esse fim, serão retratadas as diferentes abordagens não paramétricas para amostras independentes e amostras emparelhadas.

2.4.1 *Em amostras independentes*

Para verificar se são significativas as diferenças entre duas áreas abaixo da curva ROC resultantes de duas amostras independentes, aplica-se a razão critica Z definida por Hanley e McNeil (Hanley and McNeil, 1983)

$$Z = \frac{A_1 - A_2}{\sqrt{SE_1^2 + SE_2^2}} \sim N(0,1) \tag{2}$$

onde as áreas abaixo das curvas ROC para cada uma das modalidades a comparar estão representados por A_1 e A_2, e os erros padrão respetivos por SE_1 e SE_2. Para obter o valor das áreas abaixo da curva é usada a estatística Wilcoxon-Mann-Whitney, atrás descrita, e se o valor destas for superior a 0,5, os erros padrão associados às áreas são obtidos pela equação 1.

2.4.2 *Em amostras emparelhadas*

A razão critica Z descrita anteriormente para amostras independentes, aplica-se da mesma forma para amostras emparelhadas, com a introdução do termo $2rSE_1SE_2$ na raíz do denominador. A introdução deste termo é devida aos dados estarem correlacionados, porque foram obtidos da mesma amostra. A sua ausência iria causar um denominador de maior valor e, consequentemente, o valor de Z mais pequeno o que, provavelmente, reduziria a probabilidade de detetar diferenças significativas entre as duas modalidades (Hanley and

McNeil, 1983):

$$Z = \frac{A_1 - A_2}{\sqrt{SE_1^2 + SE_2^2 - 2rSE_1SE_2}} \tag{3}$$

O parâmetro r, coeficiente de correlação entre áreas, retrata a correlação estimada entre A_1 e A_2.

Cálculo do coeficiente de correlação entre áreas

O procedimento descrito nesta secção é baseado no estudo de Braga (2000). O método sugerido por Hanley and McNeil (1983) usa uma tabela para a resolução do coeficiente de correlação r entre as áreas A_1 e A_2, através do cálculo de dois coeficientes de correlação intermédios r_N, para as classificações dadas para pacientes normais (não doentes) e r_A para as classificações dadas para pacientes anormais (doentes). Existem duas maneiras tradicionais para o cálculo destes coeficientes: o método de cálculo do produto dos momentos para a correlação de Pearson e o método tau de Kendall. Como as variáveis em medicina são usualmente obtidas numa escala ordinal, utiliza-se o tau de Kendall para calcular r_N e r_A.

As entradas que vão constituir a tabela construída pelos autores Hanley and McNeil (1983), da qual se retira o valor de r são:

- o coeficiente de correlação médio $\Rightarrow \frac{r_N + r_A}{2}$,
- a área média $\Rightarrow \frac{A_1 + A_2}{2}$.

Por outro lado, os coeficientes de correlação entre áreas podem também ser determinados através do método sugerido por Delong et al. (1988), que se passa a descrever de acordo com Braga (2000).

Admitindo que se tem m indivíduos que apresentam a doença e n indivíduos que não apresentam a doença, a matriz de covariâncias estimada para o vetor estatístico de parâmetros $\hat{\theta} = (\hat{\theta^1}, \hat{\theta^2}, ..., \hat{\theta^k})$ que representa valores de AUC e na qual k representa o número de modalidades a comparar, é tal que:

$$\mathbf{S} = \frac{1}{m}\mathbf{S}_{10} + \frac{1}{n}\mathbf{S}_{01}$$

Sejam $\{X_i^r\}$, $\{Y_j^r\}$ (i= 1,2,..., m; j= 1,2,..., n; $1 \leq r \leq k$) os valores das variáveis onde o teste de diagnóstico é baseado. As matrizes $\mathbf{S}_{10}$ e $\mathbf{S}_{01}$ com dimensões $k \times k$ são definidas para o elemento de ordem *(r,s)* pelas seguintes expressões:

$$s_{10}^{r,s} = \frac{1}{m-1}\sum_{i=1}^{m}[V_{10}^r(X_i) - \hat{\theta^r}][V_{10}^s(X_i) - \hat{\theta^s}]$$

$$s_{01}^{r,s} = \frac{1}{n-1}\sum_{j=1}^{n}[V_{01}^{r}(Y_j) - \hat{\theta}^r][V_{01}^{s}(Y_j) - \hat{\theta}^s]$$

Para a r-ésima estatística $\hat{\theta}^r$, V_{10}^r e V_{01}^r representam as componentes em X e Y, representadas por:

$$V_{10}^{r}(X_i) = \frac{1}{n}\sum_{j=1}^{n}\psi(X_i^r, Y_j^r) \ (i = 1,2,...,m)$$

$$V_{01}^{r}(Y_j) = \frac{1}{m}\sum_{i=1}^{m}\psi(X_i^r, Y_j^r) \ (j = 1,2,...,n)$$

Com $\psi(X,Y)$ definida através da expressão da equação:

$$\psi(X,Y) = \begin{cases} 0 & Y < X \\ \frac{1}{2} & Y = X \\ 1 & Y > X \end{cases}$$

Como referido anteriormente, a média desta função conduz à estimativa da estatística de Wilcoxon-Mann-Whitney, correspondendo a um estimador $\hat{\theta}$ da área abaixo da curva ROC (Braga, 2000).

2.4.3 *Método alternativo para comparação de duas Curvas ROC*

Braga et al. (2005) apresentam uma metodologia que permite a comparação de curvas ROC que se intersetam, através da determinação de áreas parciais. Este é baseado em otimização multi-objetivo onde existe um conjunto de soluções que definem a frente de soluções ótimas de Pareto (Costa and Fernandes, 2003). Segundo Knowles and Corne (2000) uma curva de aproximação é construída de tal forma que divide o espaço em duas regiões distintas. Assim, para um dado conjunto de soluções, uma das regiões conterá todas as soluções que as dominam e a outra terá todas as soluções que são dominadas por elas. Tendo em conta que as retas de amostragem partem do mesmo ponto de referência, as distâncias deste ponto até aos pontos de interseção permitem comparar as curvas em diferentes regiões do espaço. Assim, esta metodologia permite determinar e identificar a região do espaço em que uma curva é melhor que a outra.

A metodologia descrita nesta secção tem por base o trabalho realizado por Braga et al. (2005). Considerando os pontos de coordenadas (x_i,x_j), com $i = 0, \ldots, n$ e $j = 0, \ldots, n$, em que (x_0,x_0)= (0,0) e (x_n,x_n)= (1,1), então os n segmentos de reta s_j, com j=0,...,n, são obtidos por:

$$s_j = y_j + m_j(x - x_j)$$

onde,

$$m_j = \frac{y_j - y_{j-1}}{x_j - x_{j-1}}$$

A amostragem das curvas ROC é feita utilizando K retas de amostragem com declive variável e que partem de um ponto de referência (x_R,y_R):

$$l_k = y_R - m_k(x - x_R)$$

onde,

$$m_k = \tan\left(\frac{(K+1-k)\pi}{2(K+1)}\right) \tag{4}$$

em que k = 1,..., K.

De seguida, calculam-se as coordenadas do ponto de interseção da reta de amostragem k com o segmento j da curva ROC, que são dadas por:

$$(x_k, y_k) = \left(\frac{y_R - y_i + m_j x_j + m_k x_k}{m_k + m_j}, y_R - m_k(x - x_R)\right)$$

As distâncias euclidianas são obtidas a partir dos pontos de interseção das retas de amostragem com as curvas ao ponto de referência:

$$d_k = \sqrt{(x_k - x_R)^2 + (y_k - y_R)^2}$$

Esta distância d_k permite comparar o desempenho das curvas ROC, sendo o desempenho superior onde a distância da curva é maior.

Por fim, calcula-se a área de cada triângulo definido pelas linhas de amostragem:

$$A_k = \frac{1}{2} d_k d_{k-1} \sin\left(\frac{\pi}{2(K+1)}\right)$$

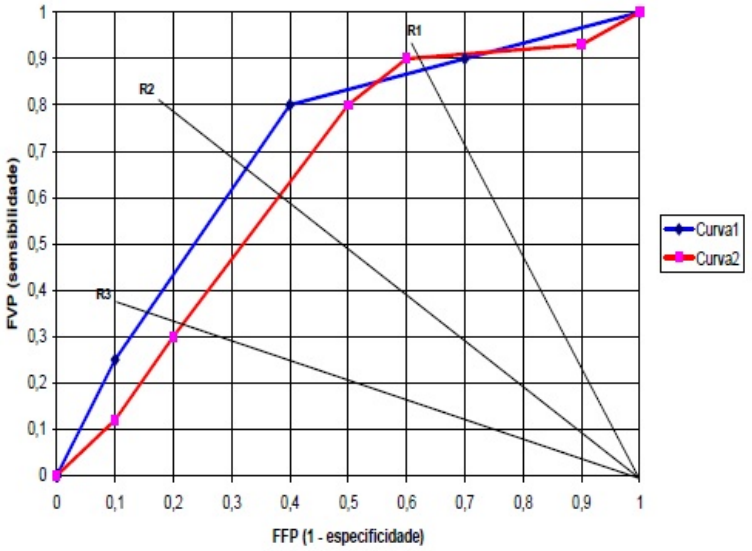


Figura 2.4: Comparação de duas curvas ROC (Fonte: (Braga et al., 2005))

Uma estimativa da área abaixo da curva ROC pode ser obtida pelo somatório da área de cada triângulo A_k.

Considerando uma amostragem baseada em três retas de amostragem, a Figura 2.4 exemplifica o método alternativo descrito por Braga et al. (2005). Com base na observação da Figura 2.4 verifica-se que a reta R1 interseta a Curva1 num ponto mais próximo que a Curva2 (desempenho da Curva2 superior ao da Curva1), e as retas R2 e R3 intersetam a Curva2 num ponto mais próximo que a Curva1 (desempenho da Curva1 superior ao da Curva2). Dado que para as três retas de amostragem o desempenho da Curva1 foi superior em duas, verifica-se que a Curva1 tem melhor desempenho do que a Curva2 em termos globais no espaço unitário (Braga et al., 2005).

2.5 programas estatísticos para análise roc/ revisão de literatura

Segundo Braga (2000), Dorfman et al. (1973) apresentam um algoritmo chamado *RSCORE* para a obtenção das estimativas de máxima verosimilhança dos parâmetros da teoria de deteção de sinal. Este usa uma variante do método de Newton-Raphson, designado por *método de scoring*. As estimativas iniciais são calculadas usando o método dos mínimos quadrados. Em 1973, Dorfman et al. (1973) comparam a eficiência do algoritmo com outras

sub-rotinas alternativas. Este foi um dos primeiros programas desenvolvidos e foi criado para *MsDOS*.

Rifkin et al. (1990) comparam duas técnicas de imagem na deteção do cancro da próstata em variados estágios da doença. Nesta comparação é utilizado o programa *CORROC2* especifico para análise ROC para duas amostras correlacionadas e desenvolvidas por Metz (Braga, 2000).

Jiang Y et al. (1996) utilizam a metodologia ROC, usando o programa *LABROC4*, para classificação e comparação de uma técnica computorizada de deteção de micro calcificações benignas ou malignas.

Metz et al. (1998) propuseram um novo método generalizado para o ajuste da curva ROC, que permite aos pesquisadores utilizar todos os dados recolhidos para comparação de duas modalidades de diagnóstico, mesmo quando os pacientes não tenham sido estudados para ambas as modalidades. O algoritmo *ROCKIT* é o resultado desta nova metodologia que também calcula diferenças estatísticas significativas entre curvas ROC. Este dispõem da vantagem de poder representar várias curvas num só gráfico mas em contrapartida apresenta desvantagens como por exemplo: o programa bloqueia e não tem um botão de ajuda,... (Braga, 2000).

O programa *GraphROC* usa o método de Hanley and McNeil (1982, 1983) para calcular as curvas ROC. Kairisto and Poola (1995) e Stephan et al. (2003) referem que este é o único programa que consegue comparar curvas com certos valores de sensibilidade e especificidade.

Um algoritmo que também faz uso do método Hanley and McNeil (1982, 1983) é o *Analyse-It*. Este programa tem a vantagem de ter o Excel integrado e de ser de simples utilização pelo outros programas. A desvantagem deste é que não consegue comparar AUC's se alguma AUC for inferior a 0,7 (Braga, 2000).

Embora o *SPSS* seja um programa estatístico muito usado, na comparação das curvas ROC ainda apresenta algumas lacunas. Este consegue obter a representação bidimensional da curva ROC no plano unitário, assim como os valores da AUC, sensibilidade, especificidade e respetivos intervalos de confiança. É também possível calcular os coeficientes de correlação das áreas com recurso à metodologia tau de kendall, mas ainda não consegue comparar curvas ROC integrando as diferentes metodologias.

No programa estatístico *R* existem vários pacotes relativos ao cálculo da curva ROC como por exemplo (Da Cunha and Braga, 2017):

- *Pacote caTools* contém enumeras funções, entre elas a função *colAUC* que permite o cálculo da AUC pelo método não paramétrico, estatística de Wilcoxon-Mann-Whitney, e a visualização das curvas ROC;
- *Pacote ROCR* é uma ferramenta útil e essencial para a criação bidimensional da curva ROC. Utiliza vários métodos estatísticos e várias medidas de desempenho;
- *Pacote Comp2ROC* (Braga et al., 2016), é uma ferramenta utilizada para a comparação de duas curvas ROC que se cruzam.

O programa *ROCNPA*, desenvolvido em JAVA, foi elaborado por Braga (2000), com o intuito de criar uma plataforma que fosse versátil para a análise ROC. Este permite determinar um ajuste para a curva ROC, avaliar o desempenho do teste de diagnóstico através de um índice de determinação simples e comparar mais que três diagnósticos quer para dados independentes quer para correlacionados e pode ser descarregado através de `http://pessoais.dps.uminho.pt/acb/englacb/feedback.htm`.

3

METODOLOGIAS APLICADAS NO DESENVOLVIMENTO DE UM SOFTWARE

A metodologia das curvas ROC é utilizada para a avaliação de desempenho de sistemas e comparação dos mesmos, para amostras independentes e emparelhadas. Para o desenvolvimento do aplicativo recorre-se à técnica de algoritmia (programação por objetos), fazendo uso de bibliotecas já desenvolvidas e disponíveis em JAVA.

3.1 PROGRAMAÇÃO EM JAVA

JAVA é uma linguagem de programação de computador orientada a objetos que foi originalmente lançada em 1995 pela Sun Microsystems (que foi adquirida pela Oracle Corporation). O código é compilado para *bytecode* que pode ser executado em qualquer máquina virtual JAVA (Java Virtual Machine, JVM), independentemente do sistema operacional (Martins, 2009).

Diferente das outras linguagens de programação, JAVA não é apenas uma linguagem que consiste somente em programação por objetos. Esta tem como base um ambiente atrativo e apropriado de programação e desenvolvimento de aplicações, especialmente a partir do sistema JDK (Java Development Kit) (Martins, 2009).

A principal característica da linguagem JAVA é que inclui um idioma simples que pode ser programado sem treino extensivo do programador, onde os principais conceitos são apreendidos rapidamente.

A robustez e segurança deste tipo de linguagem consiste em possuir uma extensa verificação de tempo de compilação, seguida de um segundo nível de verificação de tempo de execução. Isto é, no desenvolvimento de código JAVA o sistema irá encontrar erros rapidamente, onde os principais problemas não serão suspensos até que exista uma atualização do código. Por outro lado, o JAVA permite incluir chaves criptográficas no próprio código, possibilitando deste modo a identificação da origem do mesmo (Martins, 2009).

Basicamente o desenvolvimento de aplicações usando este tipo de linguagem origina um software de alta segurança e desempenho que inclui múltiplas arquiteturas, sistemas operacionais e interface gráficas.

Para além disso, os programadores têm acesso a bibliotecas já existentes de objetos testados que fornecem funcionalidades complementares ao novo programa.

3.2 BIBLIOTECA RSERVE

As linguagens de programação, como por exemplo JAVA, são muito utilizadas para o desenvolvimento de aplicações, mas não são muito eficientes quando se trata de modelação estatística e matemática. Para compensar essa lacuna, temos linguagens como R, que possui um rico conjunto de bibliotecas de aprendizagem e estatística. Integrando essas duas tecnologias, podemos criar aplicações baseados em modelação estatística de alta qualidade.

Rserve é uma biblioteca disponível em JAVA, que permite estabelecer comunicação entre JAVA e R, tornando possível a obtenção de resultados estatísticos usando funções e bibliotecas disponíveis em R.

A interpolação da aplicação com o Rserve é realizada através da incorporação do programa R no projeto. Com esta operação, na aplicação é possível abrir o R, executar o algoritmo e posteriormente fechá-lo.

Na Figura 3.1 está ilustrado um esquema, explicando superficialmente o funcionamento do Rserve com o JAVA.

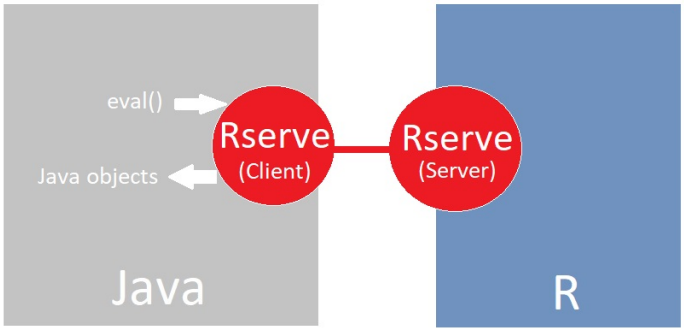


Figura 3.1: Esquema Rserve interligando Java com R (Fonte: próprio).

3.3 BASE DE DADOS

Para exemplificar a operacionalidade da aplicação desenvolvida, foram utilizadas bases de dados referentes a indicadores de gravidade clínica neonatal para recém-nascidos de muito baixo peso (peso inferior a 1500g). Estes indicadores de risco, são escalas ordinais e fazem parte do estudo incluído em Braga (2000).

3.3.1 *Índices de Gravidade Clínica Neonatal para amostras emparelhadas*

A medida mais importante de risco neonatal inicial, devido à facilidade de avaliação, foi sem dúvida durante décadas, o peso do recém-nascido. As taxas de mortalidade neonatal com base neste estudo, são um dos indicadores mais importantes para a avaliação do desempenho dos cuidados de saúde e de como se encontra o desenvolvimento da própria sociedade a este nível (Gagliardi et al., 2004; Marshall et al., 2005; Parry et al., 2003).

No entanto, para estas avaliações serem mais precisas, mais fiáveis, começaram a ser necessárias comparações entre os próprios serviços, regiões e países. Constatou-se que os recém-nascidos de muito baixo peso, menos de 1500 gramas ao nascer, contribuem em larga escala para as taxas de mortalidade e assim, foram desenvolvidas escalas de gravidade clínica específicas, para este grupo (Marshall et al., 2005). Dessas escalas, salientam-se:

- *CRIB (Clinical Risk Index for Babies)*,
- *NTISS (Neonatal Therapeutical Intervention Score System)*,
- *SNAP (Score for Neonatal Acute Physiology)*,
- *SNAP-PE (Score for Neonatal Acute Physiology - Perinatal Extension)*.

De referir que estes diferentes sistemas de pontuação ordinal implicam a recolha de variáveis ao longo de um determinado tempo, que varia entre 6 (*CRIB*), 26 (*SNAP*), 29 (*SNAP-PE*) e 48 (*NTISS*).

A sobrevivência de recém-nascidos de muito baixo peso depende do tempo de gestação e do peso à nascença. As amostras são recolhidas nas primeiras 24 horas de vida, sendo que para o CRIB, o período é reduzido para as 12 horas pós-parto, tornando-se num índice de maior facilidade utilitária, em termos de tempo (Pollack et al., 2000; Parry et al., 2003).

A amostra utilizada neste estudo é proveniente de um hospital em Portugal recolhida durante o período de três anos (1992 a 1995). Dos 169 recém-nascidos de muito baixo peso, 133 sobreviveram, tendo-se observado 36 óbitos.

3.3.2 *Índices de Gravidade Clínica Neonatal para amostras independentes*

Tendo por base a taxa de mortalidade nas unidades dos cuidados intensivos neonatais, há que ter em conta os métodos aplicados para ajustar as diferenças existentes no risco inicial dos pacientes e comparar o desempenho destas unidades onde nascem os bebés, com as condições em que são recebidos os recém-nascidos encaminhados de outras unidades, podendo estes apresentar um risco inicial de sobrevivência muito elevado.

Sendo o peso à nascença uma medida importante na determinação do risco neonatal, não houve necessidade de desenvolver novos sistemas de classificação para estes cuidados, no entanto há que prever outras diferenças no risco, tais como o grau (risco) inicial da doença (Network, 1993).

Tendo por base o *CRIB*, como medida de risco neonatal inicial, este pode ser utilizado para comparar os cuidados oferecidos por unidades de cuidados intensivos neonatais de diversos hospitais. A amostra utilizada para esta ilustração é constituída por 234 recém-nascidos de muito baixo peso (inferior a 1500 g) provenientes de 4 hospitais em Portugal durante o ano de 1995, com a designação:

- Hospital 1 - H1,
- Hospital 2 - H2,
- Hospital 3 - H3,
- Hospital 4 - H4.

3.4 REQUISITOS

Embora existam alguns programas que realizam análise ROC, não existe nenhum que consiga contemplar a apresentação gráfica com a comparação de dois ou mais sistemas ROC. Para facilitar o processo de informação relativas a estimativas ROC, uma aplicação que contemple as várias metodologias ROC pode ser criada.

Tendo em conta os objetivos delineados para este trabalho, a nova aplicação deve seguir os seguintes requisitos:

1. O utilizador deve poder criar, abrir e guardar ficheiros de dados;
2. A ferramenta deve permitir ao utilizador editar ficheiros de dados;
3. Deve ser possível importar/exportar ficheiros EXCEL (.xls);
4. A ferramenta deverá ter comandos básicos como copiar, cortar e colar;

5. Precisará de apresentar os resultados das estimativas ROC de uma forma simples e intuitiva;

6. Deverá ser capaz de fazer uma representação gráfica, que o utilizador poderá gravar em ficheiro de imagem (.jpeg);

7. Precisará de ter um botão de ajuda, para facilitar a utilização do novo programa.

3.5 ABORDAGEM

A abordagem escolhida foi o desenvolvimento de um software em JAVA que implemente as metodologias ROC descritas no capitulo 2.

A Figura 3.2 representa um esquema simplificado do novo algoritmo. A classe *interface* é a principal responsável pela estrutura do programa e para obter resultados ou gráficos terá de usar classes como *DataFrame* e *Table*. Estas classes permitem recolher valores que o utilizador fornece ao programa e fazer os cálculos das várias estimativas ROC, usando métodos específicos presentes dentro dessas. Caso haja interseção das curvas ROC, utiliza-se uma outra classe (*Comp2Roc*), que faz uso da biblioteca Rserve e do método alternativo atrás descrito para cálculo das estimativas da curva ROC.

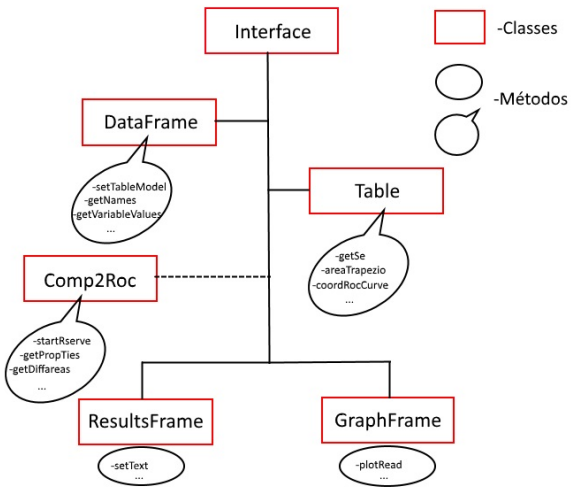


Figura 3.2: Esquema do algoritmo.

O aplicativo deve seguir o esquema e todos os requisitos listados anteriormente.

4

CERCUS

Comparison Empirical Roc Curves Cross (CERCUS) é uma aplicação (software) desenvolvida em JAVA que facilita a análise através das curvas ROC, fornecendo os resultados das curvas e os respetivos gráficos.

O nome CERCUS foi obtido usando palavras chaves (”Comparison”, ”Empirical”, ”Roc”, ”Curves”e ”Cross”), num gerador de acrónimos disponível em: `http://acronymcreator.net/ace.py`. O logótipo é original e foi inspirado na representação das curvas ROC no espaço ROC unitário.

A aplicação possibilita a incorporação e edição de dados, sendo possível a comparação de duas ou mais curvas ROC. O software é dividido em três setores, como observado na Figura 4.1.

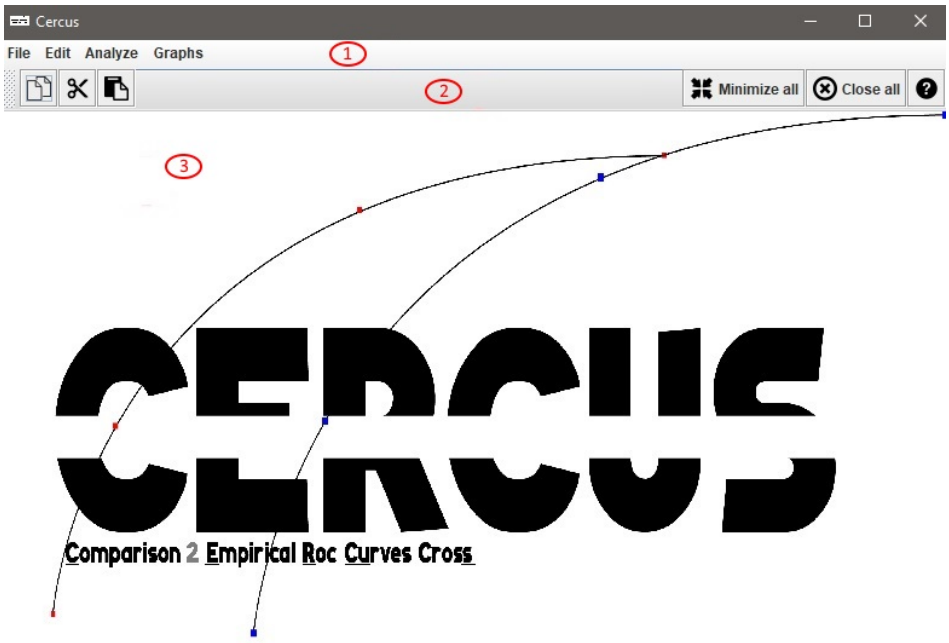


Figura 4.1: CERCUS apresentado em 3 setores distintos.

1. Barra de menus
2. Barra de ferramentas
3. Painel de fundo

4.1 BARRA DE MENUS

A Barra de Menus foi dividida em quatro grupos:

- "File" é um menu próprio para abrir, criar e guardar dados.
- "Edit" é um menu que consiste na edição dos dados e janelas. Esta encontra-se representada na barra de ferramentas.
- "Analyze" é um menu que se baseia no cálculo das estimativas ROC de suporte à metodologia abordada no capitulo 2.
- "Graphs" é um menu que refere a ilustração dos respetivos gráficos.

4.2 BARRA DE FERRAMENTAS

Figura 4.2: imagem da Barra de Ferramentas.

A barra de ferramentas foi divida em dois setores, sendo estas o setor de edição de dados (situado mais a esquerda da Figura 4.2) e o setor de edição de janelas (situado mais a direita da Figura 4.2), com exceção do botão ajuda "help". Este botão quando premido faz surgir um documento em formato pdf, que explica o funcionamento do programa e irá estar disponível em 2 idiomas, português e inglês.

No grupo de edição de dados, os três botões servem para copiar, cortar e colar valores na janela de dados. Estes só funcionarão após a introdução e posterior seleção da janela de dados.

No grupo de edição de janelas, estão disponíveis dois botões que serve para minimizar "Minimize all" e fechar "Close all" todas as janelas disponíveis no painel de fundo da aplicação.

4.3 INTRODUÇÃO DE DADOS

A introdução de dados no programa pode ser feita de três formas distintas:

- criação de um novo ficheiro de dados que pode ser guardado para edição;
- a partir de um ficheiro previamente guardado na aplicação;
- a partir de um ficheiro EXCEL (.xls).

A Figura 4.3 serve para melhor entender o menu "File" apresentada no CERCUS.

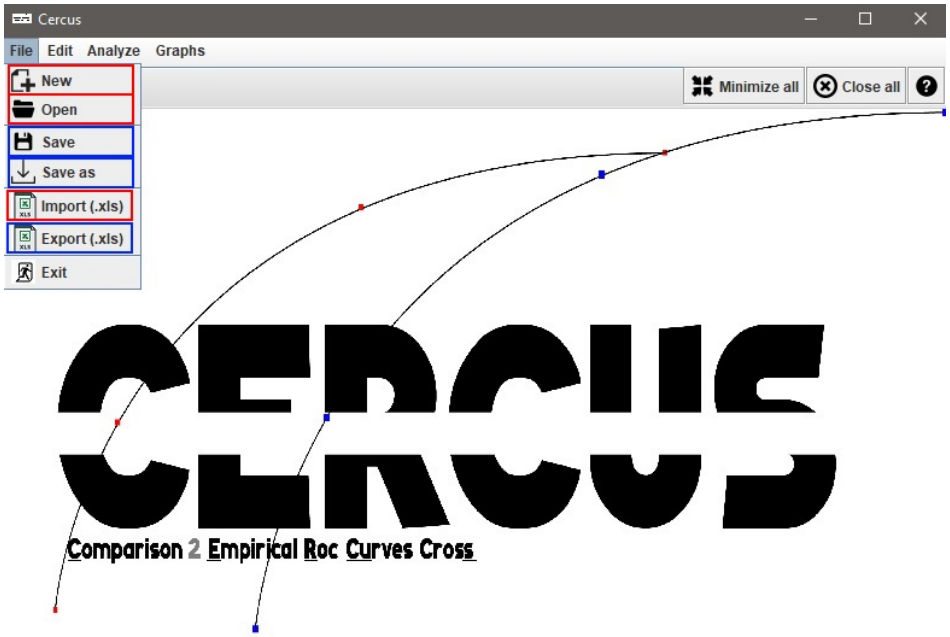


Figura 4.3: Janela de Menu "File"do CERCUS.

4.3.1 *Criação de um novo ficheiro*

Para a criação de um novo ficheiro de dados é necessário premir o botão "New" do Menu "File" onde o programa apresentará três janelas de diálogo de definição de variáveis. A primeira serve para caracterizar a amostra, isto é, questiona quantas variáveis estão em estudo e identifica se trata de dados provenientes de amostras emparelhadas ou independentes, como exemplificado na Figura 4.4.

Este menu permitirá definir a estrutura de dados a exibir.

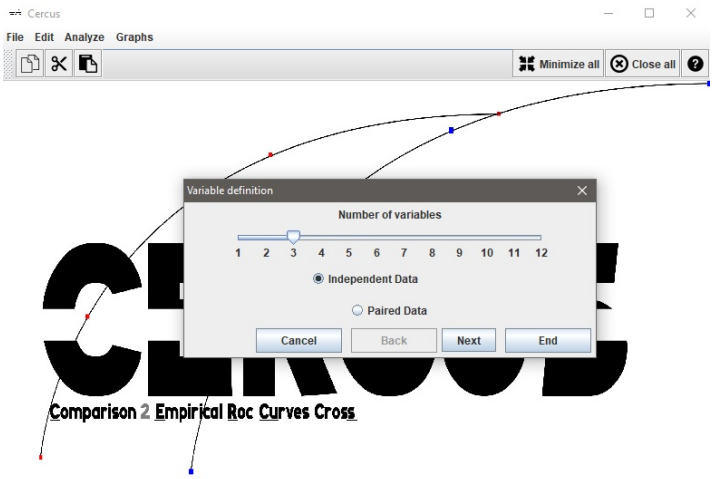


Figura 4.4: Primeira janela de diálogo para caracterização da amostra.

Depois de pressionado o botão "Next", a segunda janela será apresentada. Esta terá como intuito definir o nome das variáveis tal como ilustrado na Figura 4.5.

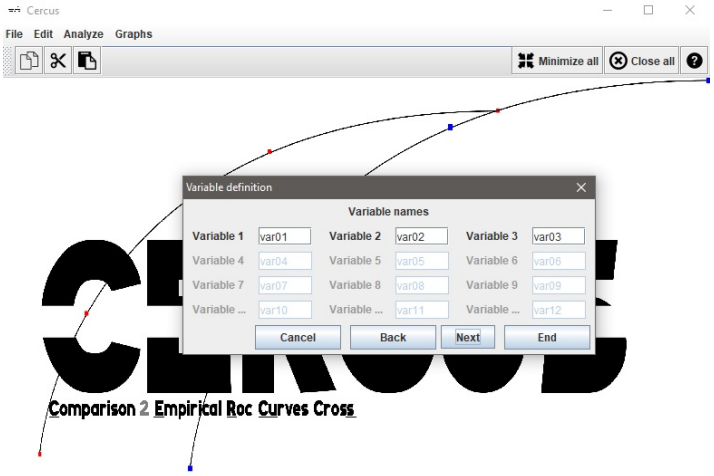


Figura 4.5: Segunda janela de diálogo para a definição de nomes para as variáveis.

Após a definição de nomes para as variáveis, uma última janela de diálogo será apresentada. Esta permite completar a definição da amostra, isto é, qual o valor da escala que corresponde ao teste positivo (se são os valores maiores ou menores que correspondem ao teste positivo), demonstrado na Figura 4.6.

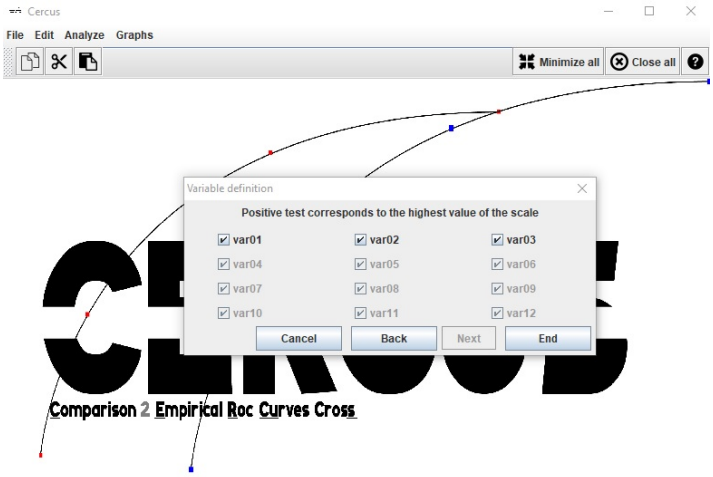


Figura 4.6: Terceira janela de diálogo para a definição do valor da escala.

4.3.2 *Seleção de um ficheiro*

Para abrir um ficheiro já existente é necessário que o utilizador clique no botão "Open" ilustrado na Figura 4.3. A janela da Figura 4.7 será exibida, fornecendo opção de escolha e procura. O CERCUS permite o acesso a um ficheiro de leitura cuja extensão, própria do software, é ".cer", como se pode verificar na Figura 4.3.

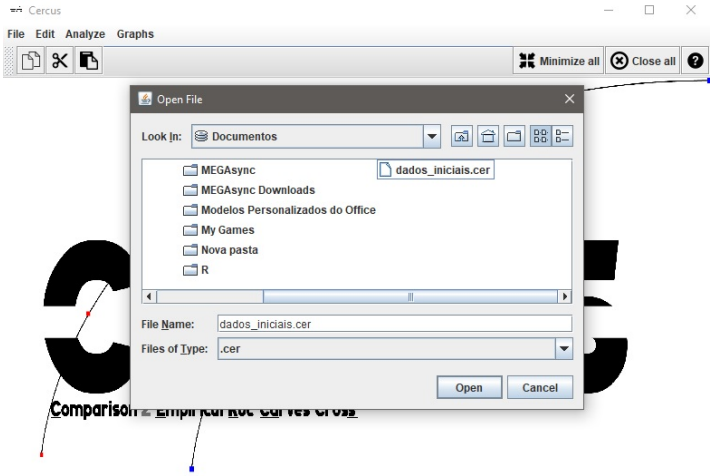


Figura 4.7: Janela de diálogo para abrir um projeto.

4.3.3 *Importação de ficheiros .xls*

Para importar um ficheiro EXCEL é necessário clicar no botão "Import (.xls)" localizado no menu "File", exemplificado na Figura 4.3. A janela da Figura 4.8 será exibida de forma a que o botão "Open File" fará a seleção do ficheiro, neste caso limitado pela a extensão ".xls". Ainda nesta janela é fundamental a seleção do tipo de dados. Caso o utilizador prima o botão "Next" sem a caracterização da amostra, uma mensagem de erro irá aparecer.

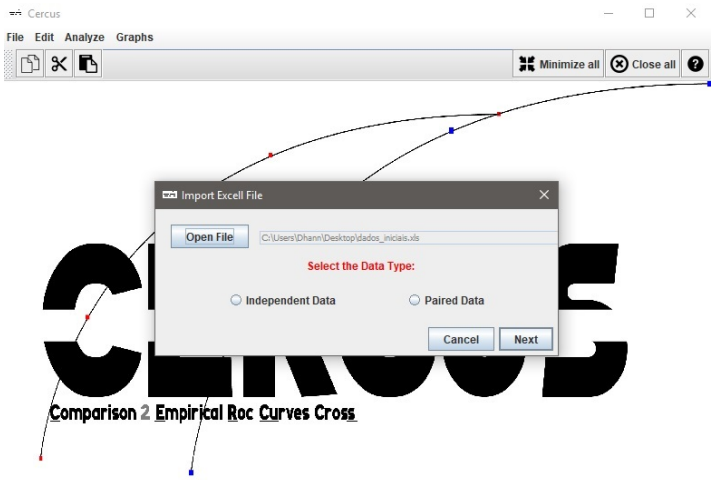
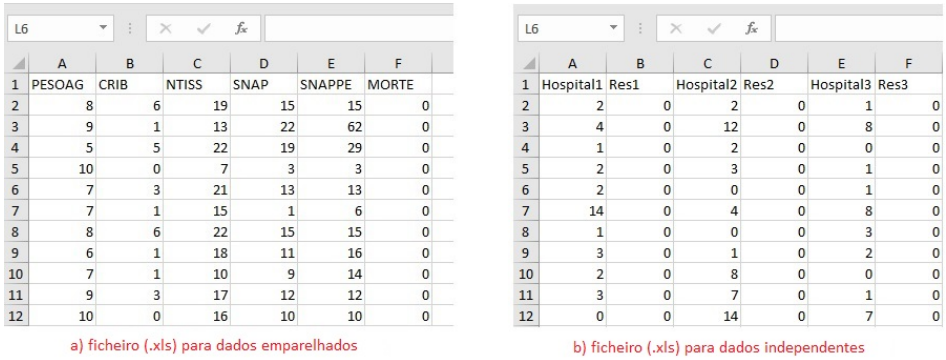


Figura 4.8: Janela de diálogo para importar um ficheiro de EXCELL (.xls).

Uma janela de diálogo similar à da Figura 4.6 é mostrada, com exceção do nome das variáveis, caso não ocorra nenhum problema com a importação do ficheiro EXCEL, para identificar as variáveis que correspondem ao teste positivo.

Para o ficheiro de EXCEL ser importado corretamente, este deve estar em formato .xls (Livro de Excell 97-2003) e também ser preenchido desde a primeira linha e coluna. Isto é, não pode haver espaços em branco entre colunas na primeira linha. Caso isso aconteça a leitura do ficheiro não será corretamente procedida. Também é preciso ter atenção em que os nomes devem estar unicamente na primeira linha. Caso o ficheiro encontre caracteres não numéricos após a primeira linha, a importação do ficheiro irá ser impossível. Por outro lado, nos ficheiros para amostras emparelhadas, a última coluna deve estar destinada à variável de resposta (0 ou 1) e nos ficheiros para amostras independentes as variáveis e a sua resposta (0 ou 1) devem estar intercaladas.

A Figura 4.9 ilustra um exemplo de como os dados devem estar distribuídos no EXCEL, podendo estes não estar ordenados.



L6

	A	B	C	D	E	F
1	PESOAG	CRIB	NTISS	SNAP	SNAPPE	MORTE
2	8	6	19	15	15	0
3	9	1	13	22	62	0
4	5	5	22	19	29	0
5	10	0	7	3	3	0
6	7	3	21	13	13	0
7	7	1	15	1	6	0
8	8	6	22	15	15	0
9	6	1	18	11	16	0
10	7	1	10	9	14	0
11	9	3	17	12	12	0
12	10	0	16	10	10	0

a) ficheiro (.xls) para dados emparelhados

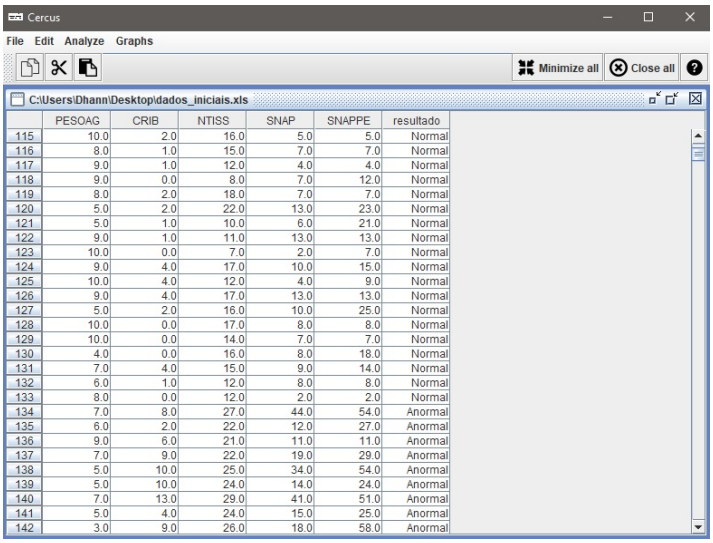
L6

	A	B	C	D	E	F
1	Hospital1	Res1	Hospital2	Res2	Hospital3	Res3
2	2	0	2	0	1	0
3	4	0	12	0	8	0
4	1	0	2	0	0	0
5	2	0	3	0	1	0
6	2	0	0	0	1	0
7	14	0	4	0	8	0
8	1	0	0	0	3	0
9	3	0	1	0	2	0
10	2	0	8	0	0	0
11	3	0	7	0	1	0
12	0	0	14	0	7	0

b) ficheiro (.xls) para dados independentes

Figura 4.9: Exemplo de como os dados devem estar representados no EXCEL.

Na Figura 4.10, apresenta-se o aspeto da janela de dados para quatro amostras emparelhadas, cujo maior valor da escala corresponde ao teste positivo para todas as variáveis com exceção do "PESOAG". Isto é, o resultado negativo significa sobrevivência (Normal) e o positivo significa falecimento (Anormal) enquanto no caso do "PESOAG", os valores menores da escala indicam valor positivo (falecimento).



Cercus

File Edit Analyze Graphs

Minimize all Close all

C:\Users\Dhann\Desktop\dados_iniciais.xls

	PESOAG	CRIB	NTISS	SNAP	SNAPPE	resultado
115	10.0	2.0	16.0	5.0	5.0	Normal
116	8.0	1.0	15.0	7.0	7.0	Normal
117	9.0	1.0	12.0	4.0	4.0	Normal
118	9.0	0.0	8.0	7.0	12.0	Normal
119	8.0	2.0	18.0	7.0	7.0	Normal
120	5.0	2.0	22.0	13.0	23.0	Normal
121	5.0	1.0	10.0	6.0	21.0	Normal
122	9.0	1.0	11.0	13.0	13.0	Normal
123	10.0	0.0	7.0	2.0	7.0	Normal
124	9.0	4.0	17.0	10.0	15.0	Normal
125	10.0	4.0	12.0	4.0	9.0	Normal
126	9.0	4.0	17.0	13.0	13.0	Normal
127	5.0	2.0	16.0	10.0	25.0	Normal
128	10.0	0.0	17.0	8.0	8.0	Normal
129	10.0	0.0	14.0	7.0	7.0	Normal
130	4.0	0.0	16.0	8.0	18.0	Normal
131	7.0	4.0	15.0	9.0	14.0	Normal
132	6.0	1.0	12.0	8.0	8.0	Normal
133	8.0	0.0	12.0	2.0	2.0	Normal
134	7.0	8.0	27.0	44.0	54.0	Anormal
135	6.0	2.0	22.0	12.0	27.0	Anormal
136	9.0	6.0	21.0	11.0	11.0	Anormal
137	7.0	9.0	22.0	19.0	29.0	Anormal
138	5.0	10.0	25.0	34.0	54.0	Anormal
139	5.0	10.0	24.0	14.0	24.0	Anormal
140	7.0	13.0	29.0	41.0	51.0	Anormal
141	5.0	4.0	24.0	15.0	25.0	Anormal
142	3.0	9.0	26.0	18.0	58.0	Anormal

Figura 4.10: Janela de dados no CERCUS para um conjunto de quatro variáveis emparelhadas.

4.4 GUARDANDO E EXPORTANDO UM PROJETO

Depois de criado o ficheiro de dados, o utilizador pode guardá-lo ou exportá-lo para posterior utilização. Para guardar um projeto de dados o utilizador deve clicar no botão "Save" ou "Save as" mostrado na Figura 4.3. A janela da Figura 4.11 será exibida solicitando que seja informada a pasta e o nome do arquivo. Se o projeto se encontrar já guardado o botão "Save" só faz atualização dos dados para o mesmo nome do arquivo. Caso o utilizador esteja a trabalhar num projeto previamente guardado pelo programa e queira mudar de nome do arquivo, necessita de clicar no botão "Save as" onde a janela da Figura 4.11 será mostrada novamente.

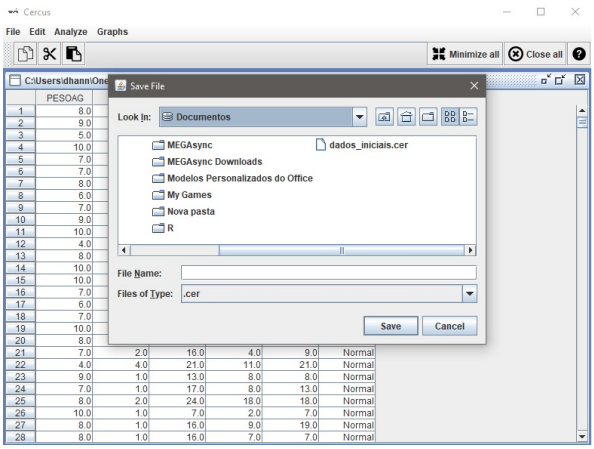


Figura 4.11: Janela de dialogo para guardar/exportar um projeto.

Para exportação do projeto o processo é muito semelhante ao guardar, onde a única diferença é a própria extensão do arquivo. No CERCUS o utilizador pode trabalhar em projetos com a extensão .cer, optar por fazer a devida exportação (.xls) e trabalhar em ficheiro importados (.xls) e posteriormente guardar em formato próprio do programa .cer.

4.5 COMPARAÇÃO DE DUAS OU MAIS CURVAS ROC

O CERCUS permite a comparação de duas ou mais curvas ROC. Serão apresentados vários resultados relativos as informações presentes na janela de dados, quer se trate de dados provenientes de amostras independentes ou emparelhadas. O menu "Analyze" está destinado para efetuar essa comparação fornecendo ao utilizador duas opções:

- Teste de comparação múltipla tradicional ("Traditional Multiple Comparison Test"), através do procedimento descrito da estatística Z (equações 2 e 3)

- Resultado da amostragem ROC ("Roc Sampling Results"), através do procedimento descrito quando duas curvas se cruzam (Braga et al., 2005).

Estas opções só serão possíveis após a introdução de dados e subsequentemente a seleção da janela de dados.

4.5.1 *Teste de comparação múltipla tradicional*

O teste de comparação múltipla tradicional fornece ao utilizador uma série de estimativas ROC que são usadas quando as curvas ROC não apresentam cruzamentos entre si.

Internamente, a operação está dividida em três etapas, sendo estas, o armazenamento da informação presente na janela de dados, o calculo de estimativas relativas à curva ROC e apresentação dos resultados. Começando por localizar a janela de dados, a informação vai ser atribuída a uma variável interna para posteriormente ser usada como base para cálculo das estimativas relativas à curva ROC.

Estas serão:

- o índice da área abaixo da curva ROC (AUC), que é determinado pela a aproximação não paramétrica à estatística de Wilcoxon-Mann-Whitney;
- os valores dos erros padrão (SE), que são determinados pela rotina sugerida por Hanley and McNeil (1982);
- os valores da razão critica (Z), definida por Hanley and McNeil (1983);
- os valores de (p-value), obtidos a partir da distribuição Normal da razão crítica Z.

Caso a janela de dados seja para amostras emparelhadas é necessário determinar os coeficientes de correlação, método aplicado e sugerido por Delong et al. (1988), definido no capitulo 2.4.2.

Por fim é aberta uma nova janela na qual serão apresentadas as estimativas ROC, tal como exemplificado na Figura 4.12.

A comparação é efetuada através da AUC da curva ROC e por comparações múltiplas dois a dois, usando a estatistica de teste Z e valores p correspondentes, definida por Hanley and McNeil (1983) e referida nas expressões (2 e 3) do capitulo 2.4.

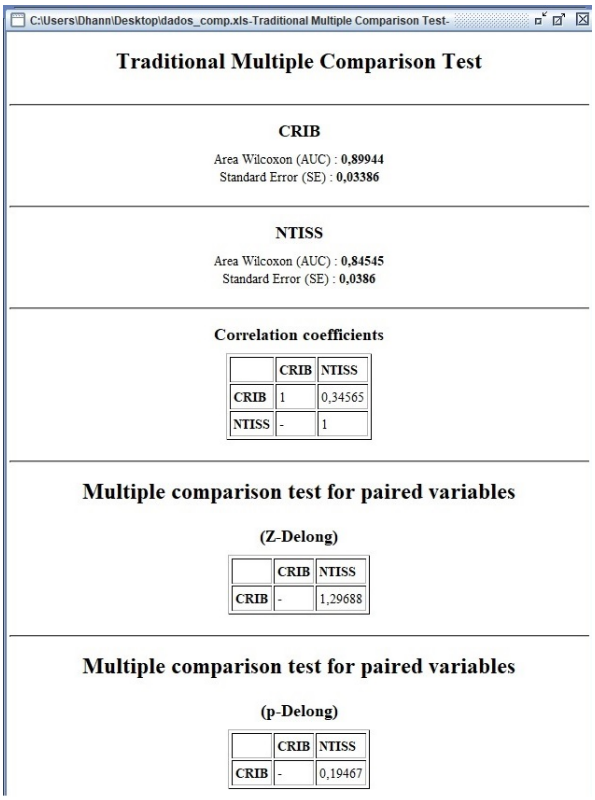


Figura 4.12: Exemplo de representação de dados do Teste de comparação múltipla tradicional, neste caso só para duas variáveis.

4.5.2 *Resultado da amostragem ROC*

O resultado da amostragem ROC disponibiliza ao utilizador os resultados analíticos da comparação de duas curvas ROC pelo método proposto por Braga et al. (2005). Esta opção é usada para quando as curvas ROC apresentam cruzamentos entre si resultando na apresentação de estimativas ROC mais detalhadas.

Intrínseca ao programa CERCUS, a biblioteca Rserve é utilizada nesta opção para calcular os respetivos resultados da comparação ROC, usando a livraria "Comp2ROC" (Braga et al., 2016).

Basicamente, após selecionada a opção "Roc Sampling Results" a janela de diálogo, exemplificada na Figura 4.13, permite a seleção das variáveis que o utilizador quer comparar.

Internamente após a seleção da janela de dados, esta retira a informação relativa às duas variáveis selecionadas na janela de diálogo (ver Figura 4.13), calcula os resultados usando a ligação com R (Rserve) e consequentemente apresenta os resultados numa nova janela tal como se apresenta na Figura 4.14.

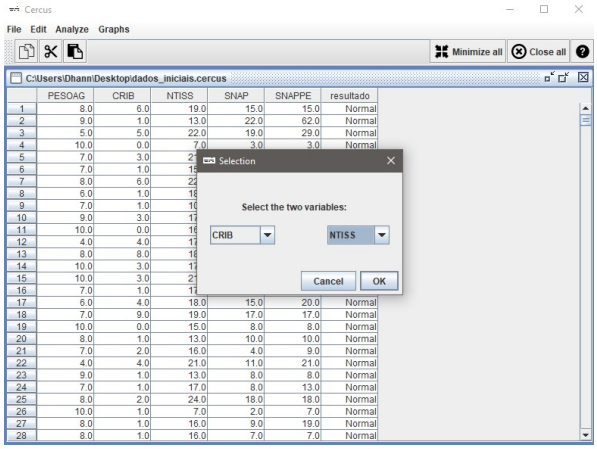


Figura 4.13: Janela de diálogo de seleção das variáveis.

Nesta opção a comparação pode ser feita não só pelo valor de AUC e os teste de diferenças mas também pelo cálculo da proporção que uma curva ganha à outra como definido no trabalho desenvolvido por Braga et al. (2005). Isto é, quanto maior o valor da proporção, melhor vai ser o seu desempenho em relação a outra curva, no espaço ROC unitário.

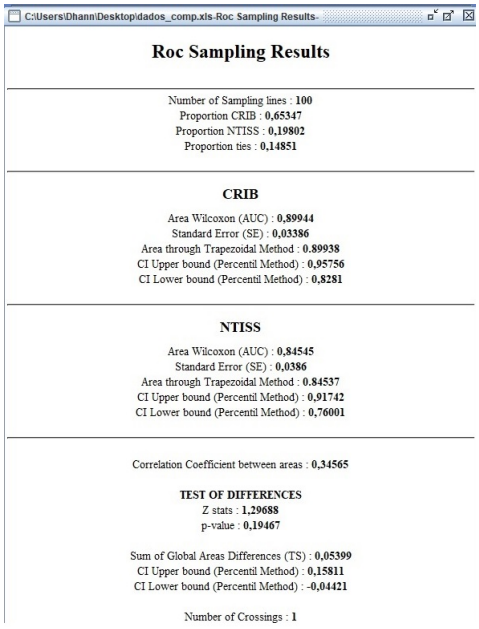


Figura 4.14: Exemplo de representação de dados no Resultado da amostragem ROC.

4.5.3 *Representação dos gráficos*

Por fim, o menu "Graphs" está destinado para a representação dos gráficos. Esta representação está dividida em três partes podendo o utilizador posteriormente gravar os gráficos em ficheiro (.jpeg):

- Curvas ROC empiricas "Empirical ROC curve(s)"
- Curvas ROC empiricas (2 a 2) "Empirical ROC curves (2 by 2)"
- Área entre curvas ROC "Area Between ROC curves"

A Figura 4.15 serve para melhor entender o menu "Graphs" apresentado no CERCUS.

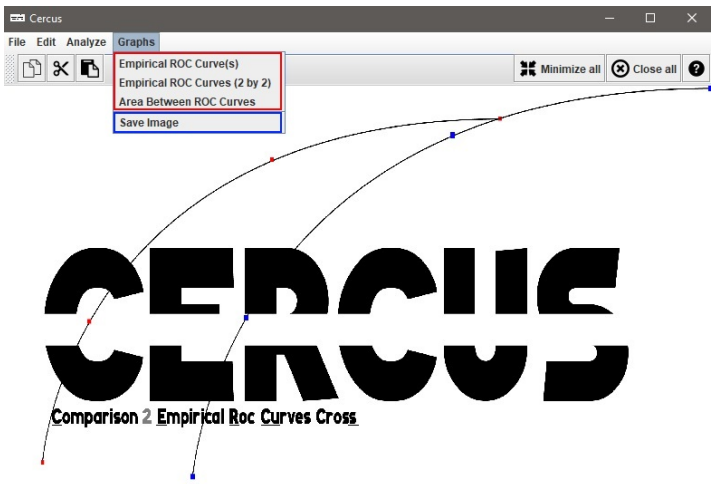


Figura 4.15: Janela de Menu "Graphs" do Cercus.

As curvas ROC empíricas são produzidas pela união dos pontos coordenados, que correspondem aos pares (1 - especificidade, sensibilidade), calculados para cada caso. Para a opção "Empirical ROC curves (2 by 2)" primeiramente será solicitado a seleção das variáveis (ver Figura 4.13) e só depois a união dos pontos coordenados.

Para as áreas entre curvas ROC, após a seleção das variaveis, a aplicação usa o *Rserve* para obter os valores de "Lower Bound", "Upper Bound" e "Degrees". Usando o método proposto por Braga et al. (2005) a aplicação calcula a diferença de áreas entre as curvas e procede a união dos pontos como referencia a variável "Degrees" (corresponde aos declives das linhas de amostragem definidas, com um valor fixo igual a 100).

Caso o utilizador queira guardar o respetivo gráfico, tem de selecionar a janela em que o gráfico esta presente e clicar no botão "Save Image" (ver Figura 4.15). Uma janela de diálogo irá aparecer, similar à Figura 4.11, onde a única diferença é a sua extensão (.jpeg).

Na Figura 4.16 encontra-se exemplificado o conjunto de janelas de resultados produzidos pela introdução de cinco amostras emparelhadas (Figura 4.9).

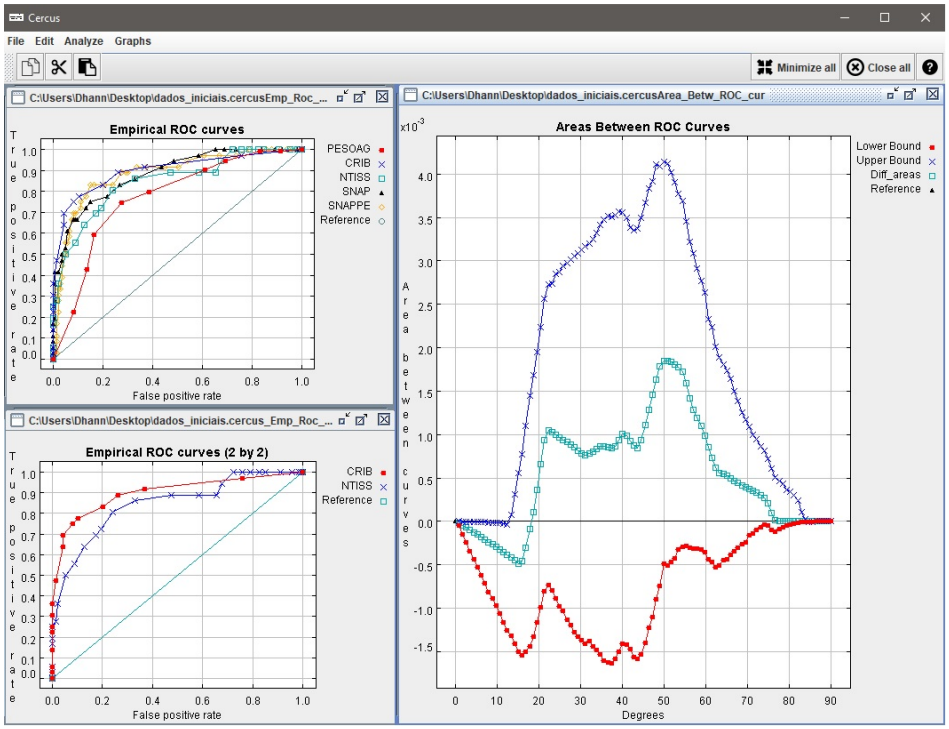


Figura 4.16: Janela de gráficos produzidos por um conjunto de dados emparelhados.

5

ANÁLISE DOS RESULTADOS

Para validar os resultados calculados pelo programa CERCUS, utilizou-se as bases de dados referidas no capitulo 3. Estas irão ser utilizadas para análise de dados através das curvas ROC utilizando alguns programas disponíveis.

Para análise das métricas referentes à analise ROC, optou-se por utilizar o SPSS 22.0, o Comp2ROC e o ROCNPA devido a utilização das mesmas metodologias para o respetivo cálculo, enquanto para análise dos resultados gráficos, optou-se por utilizar o Comp2ROC, sendo este, o único possível para comparação de duas curvas ROC que se intersetam.

A introdução de dados no SPSS é relativamente fácil e eficaz, dado que este encontra-se preparado para importação de dados em formato EXCEL. Por outro lado, na obtenção de resultados analíticos para um conjunto de dados independentes o SPSS não está preparado para fazer análise num conjunto de duas ou mais variáveis, tendo esta sido realizada uma a uma. No caso do Comp2ROC foi necessário fazer um pequeno script em R, para obtenção dos respetivos resultados. Teve que ser criado um ficheiro formato .csv para fornecer os respetivos dados ao Comp2ROC.

Efetuando a análise no programa ROCNPA foi preciso introduzir os dados diretamente através do teclado, onde o processo foi fastidioso e moroso.

5.1 ANÁLISE DE DOIS CONJUNTOS DE DADOS EMPARELHADOS

Para analisar dois conjuntos de dados emparelhados, numa perspetiva ROC, utilizou-se os dados referentes às variáveis CRIB (Clinical Risk Index for Babies) e NTISS (Neonatal Therapeutical Intervention Score System), referidas no capitulo 3.3.1.

Trata-se de indicadores que variam em uma escala ordinal entre (0 a 16) para o CRIB e (6 a 33) para o NTISS na qual o resultado irá diferenciar entre 0 e 1. Se a variável resultado representar o valor zero esta retrata que o recém-nascido não irá falecer (teste negativo - "Normal") e se a variável resultado apresentar o valor 1 esta demonstra que o recém-nascido irá falecer (teste positivo - "Anormal"). Dos dados relativos aos 169 recém-nascidos de muito baixo peso (menos de 1500g) em estudo, 133 sobreviveram, tendo sido registados 36 óbitos.

Quando se introduz o respetivo ficheiro de dados, há que ter em conta que se trata de variáveis cujo maior valor da escala corresponde ao teste positivo.

Em termos de análise de resultados, a Tabela 2 apresenta o resumo dos valores obtidos em cada um dos programas testados, para a AUC, os erros padrão e os respetivos testes de diferença.

Tabela 2: Resumo dos valores obtidos para dois conjuntos de dados emparelhados.

	SPSS	**Comp2ROC**	**ROCNPA**	**CERCUS**
CRIB	Área = 0.899 SE = 0.034	Área = 0.899436 SE = 0.033864	Área = 0.899436 SE = 0.033864	Área = 0.89944 SE = 0.03386
NTISS	Área = 0.845 SE = 0.038	Área = 0.845447 SE = 0.038599	Área = 0.845447 SE = 0.038599	Área = 0.84545 SE = 0.0386
Testes de Diferenças Z		Z = 1.296885 p = 0.194671	Z = 1.300461 p = 0.193601	Z = 1.29688 p = 0.19467
Testes de Diferenças (TS) Limite inferior Limite superior número cruzamentos		 0.053989 -0.0366568 0.155206 1		 0.05399 -0.04560 0.15122 1

A análise destes valores permite concluir que os resultados da comparação conduzem ao mesmo tipo de decisão independentemente do teste utilizado.

Em termos de análise gráfica as Figuras 5.1 e 5.2, traduzem os gráficos obtidos no Comp2ROC e CERCUS para as curvas ROC empíricas e a área entre as curvas ROC em função das linhas de amostragem.

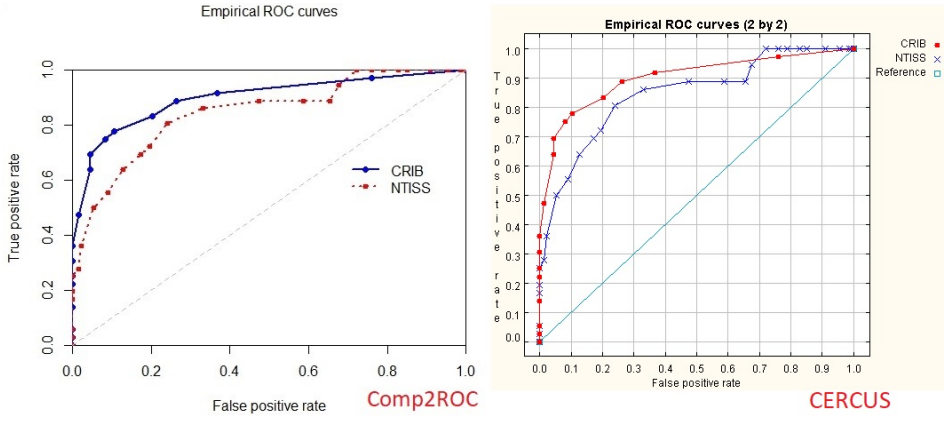


Figura 5.1: Curvas ROC empíricas obtidas pelo Comp2ROC e o CERCUS para dados emparelhados.

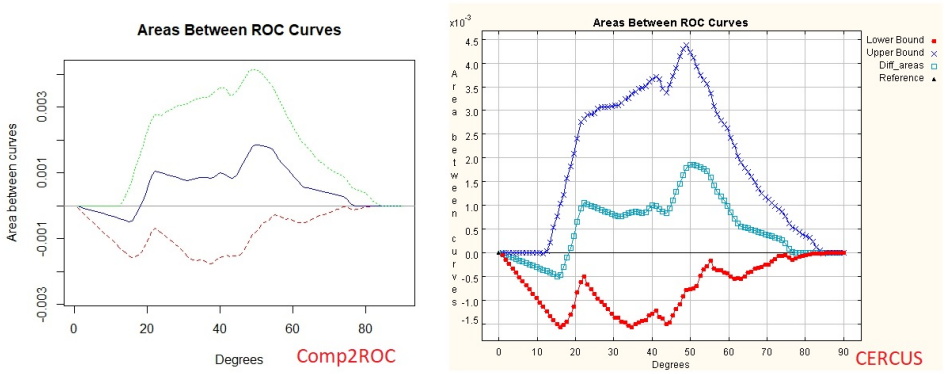


Figura 5.2: Área entre as curvas ROC, obtidas pelo Comp2ROC e o CERCUS, para dados emparelhados.

A análise destes gráficos permite concluir que não existem mudanças significativas, para a representação das curvas ROC empíricas e a área entre as curvas ROC.

5.2 Análise de dois conjuntos de dados independentes

Para ilustrar a análise de dois conjuntos de dados independentes, numa perspetiva ROC, utilizou-se os dados referentes à variável CRIB (Clinical Risk Index for Babies), para os recém-nascidos de muito baixo peso, relativo a dois hospitais (Hospital 1- H1 e Hospital 2- H2), referidas no capitulo 3.3.2. Esta permite a comparação de desempenho em termos de cuidados prestados para os dois hospitais, sendo estas variáveis independentes.

Trata-se de variáveis que variam em uma escala ordinal entre (0 a 20) na qual o resultado irá descriminar entre 0 e 1. Se a variável resultado representar o valor 0 esta retrata que o recém-nascido não irá falecer (teste negativo - "Normal") e se a variável resultado apresentar o valor 1 esta demonstra que o recém-nascido irá falecer (teste positivo - "Anormal"). Dos dados relativos aos 111 recém-nascidos de muito baixo peso em estudo, 90 sobreviveram, tendo sido registado 21 óbitos.

Quando se introduz o respetivo ficheiro de dados, há que ter em conta que se trata de variáveis cujo maior valor da escala corresponde ao teste positivo.

Para obtenção dos resultados para amostras independentes no programa SPSS, teve-se que ter atenção à seleção individual das variáveis com o respetivo resultado, dado que neste programa não é possível a representação conjunta das respetivas estimativas ROC para diferentes indicadores.

Em termos de análise de resultados, a Tabela 3 apresenta o resumo dos valores obtidos em cada um dos programas testados, para a AUC, os erros padrão e os respetivos testes de diferenças.

Tabela 3: Resumo dos valores obtidos para dois conjuntos de dados independentes.

	SPSS	**Comp2ROC**	**ROCNPA**	**CERCUS**
Hospital 1	Área = 0.592 SE = 0.105	Área = 0.523077 SE = 0.108035	Área = 0.592308 SE = 0.108035	Área = 0.59231 SE = 0.10496
Hospital 2	Área = 0.7925 SE = 0.076	Área = 0.7925 SE = 0.076257	Área = 0.7925 SE = 0.076257	Área = 0.7925 SE = 0.076
Teste de Diferenças		Z = -1.513886 p = 0.130055	Z = -1.54485 p = 0.131043	Z = -1.54485 p = 0.12238
Testes de Diferenças (TS) Limite inferior Limite superior numero cruzamentos		 -0.200192 -0.503814 0.135768 1		 -0.20019 -0.51093 0.10297 1

A análise destes valores permite concluir que os resultados da comparação conduzem ao mesmo tipo de decisão independentemente do teste utilizado, com algumas diferenças em termos numéricos no valor de AUC e o respetivo erro padrão para o Hospital 1 e no valor de Z.

Em termos de análise gráfica as Figuras 5.3 e 5.4, traduzem os gráficos obtidos no Comp2ROC e CERCUS para as curvas ROC empíricas e a área entre as curvas.

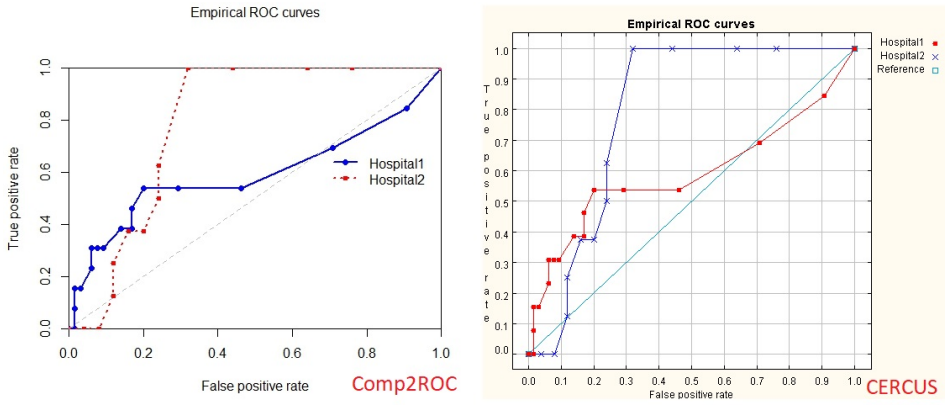


Figura 5.3: Curvas ROC empíricas obtidas pelo Comp2ROC e o CERCUS para dados independentes.

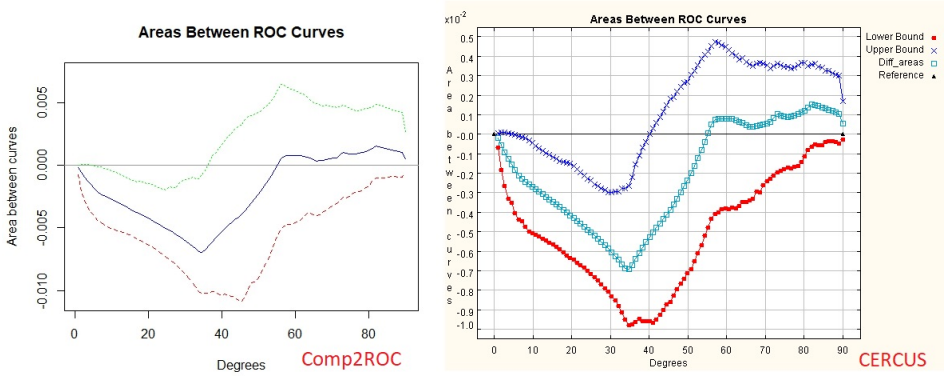


Figura 5.4: Área entre as curvas ROC, obtidas pelo Comp2ROC e o CERCUS, para dados independentes.

A análise destes gráficos permite concluir que não existe mudanças significativas, para a representação das curvas ROC empíricas e a área entre as curvas ROC na globabilidade, no entanto consegue-se demonstrar que na região do espaço entre as linhas de amostragem 10^{o} e 40^{o}, o H2 apresenta um desempenho significativamente superior quando comparado com o H1 (LB e UB abaixo do valor zero).

5.3 Discussão e Conclusão

Como pode ser verificado através dos resultados obtidos, o CERCUS apresenta praticamente os mesmos resultados analíticos e gráficos que os restantes softwares, pois tal fato é devido à utilização das mesmas metodologias. Nos softwares testados estes usam a aproximação à estatística de Wilcoxon-Mann-Whitney para o cálculo da AUC e o respetivo erro padrão. No cálculo da razão critica Z e p, o Comp2ROC e o ROCNPA, usam a mesma metodologia definida por Hanley and McNeil (1983). Em termos gráficos, apesar de haver pequenas diferenças em termos visuais, as coordenadas são obtidas usando as mesmas metodologias.

Quanto à capacidade para efetuar uma análise ROC, verificou-se que o CERCUS apresenta maior versatilidade, quer para a introdução de dados quer na obtenção de resultados para amostras emparelhadas e/ou independentes. Isto é, os resultados alcançados são facilmente transportadas para qualquer processador de texto, dado que os gráficos podem ser guardados num formato de imagem .jpeg, e a folha de resultados analíticos copiada para um bloco de notas.

Dos programas testados há que referir que o CERCUS é uma aplicação gratuita, dado que o SPSS 22.0 apresenta várias licenças nas quais maioritariamente são pagas. A licença utilizada para obtenção destes resultados foi a de utilização num prazo de 15 dias.

6

CONCLUSÕES E TRABALHO FUTURO

O foco principal deste trabalho foi o desenvolvimento de uma aplicação para computadores pessoais que consiga integrar as várias metodologias ROC, fazendo a comparação de dois sistemas com base em curvas ROC que se intersetam ou não, tendo o referido desenvolvimento sido concluído com sucesso. Apesar de identificar visualmente as regiões da curva onde existe melhor desempenho de um sistema em relação ao outro, não foi possível implementar um algoritmo de conversão de métricas para permitir identificar no espaço ROC unitário quais os pares (1-especificidade e sensibilidade) correspondentes com essa região.

Este trabalho vem dar resposta à inexistência de um software capaz de sistematizar a análise através das curvas ROC nomeadamente na representação gráfica e comparação de dois sistemas quer para dados independentes ou para dados emparelhados.

A elaboração do algoritmo teve por base a estrutura do programa ROCNPA tentando, dentro do possível, simplificar ao máximo as funcionalidades do programa. Apesar de ainda haver muitos aspetos a melhorar no interface do CERCUS, este apresenta uma versatilidade e robustez para análise de amostras de qualquer tipo.

No decorrer do trabalho foram encontradas muitas dificuldades, nomeadamente na estrutura de código, na pesquisa de livrarias e na obtenção dos resultados. A inexistência de bibliotecas em JAVA capazes de realizarem *bootstraping*, previsões de resultados analíticos, levou a uma complicação na realização desta dissertação, que foi colmatada com o estudo e compreensão da implementação da biblioteca Rserve.

Muitas idas e vindas, bastante código descartado, pode-se afirmar que os principais requisitos de implementação e abordagem deste trabalho foram cumpridos, deixando no entanto uma porta aberta, para sugestões de trabalhos futuros que podem ser traduzidos em melhorias no programa.

6.1 TRABALHO FUTURO

O trabalho desenvolvido pode ser melhorado e complementado. A implementação de um botão que consiga traduzir os resultados em um ficheiro texto, ajudará ao utilizador a fazer comparação mais detalhada das curvas ROC.

Dado que as janelas não estão disponibilizadas de uma forma intuitiva, a criação de um menu que disponibiliza as janelas abertas, assistirá o utilizador a fazer a devida seleção.

Para criar viabilidade do software desenvolvido é necessário a elaboração de pequenas funções que restringe o utilizador, como por exemplo, não aceder aos menus "Analyze" e "Graphs" enquanto a janela de dados não estiver selecionada.

Dentro da introdução das variáveis, deve ser possível a criação de um algoritmo que consiga determinar o tipo de dados que está presente. Isto leva, a uma menor ocorrência de erros dentro do programa.

Por fim, a implementação de novas metodologias de análise ROC, como o ajuste da curva e a apresentação dos intervalos de confiança, ajudará no desenvolvimento futuro do CERCUS.

BIBLIOGRAFIA

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, 12(4):387–415.

Begg, C. B. (1991). Advances in statistical methodology for diagnostic medecine ni the 1980's. *Statistics in Medicine*, 10(12):1887–1895.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.

Braga, A. C. (2000). *Curvas ROC: Aspectos Funcionais e Aplicações*. PhD thesis, Universidade do Minho.

Braga, A. C., Costa, L. A., and Oliveira, P. N. (2005). Metodologia não paramétrica para a comparação global e parcial de curvas ROC.

Braga, A. C., Frade, H., Carvalho, S., and Santiago, A. M. (2016). Package ' Comp2ROC '.

Braga, A. C. and Oliveira, P. (2003). Diagnostic analysis based on ROC curves: theory and applications in medicine. *International Journal of Health Care Quality Assurance*, 16(4):191–198.

Cheam, A. and McNicholas, P. D. (2014). Modelling Receiver Operating Characteristic Curves Using Gaussian Mixtures. pages 1–15.

Collinson, P. (1998). Of bombers, radiologists, and cardiologists: time to ROC. *Heart*, 80(3):215–217.

Costa, L. and Fernandes, A. A. (2003). Algoritmos Evolucionários em Optimização Uni e Multi-objectivo. page 237.

Da Cunha, D. F. and Braga, A. C. (2017). Receiver operating characteristic (ROC) packages comparison in R. In Gervasi, O., Murgante, B., Misra, S., Borruso, G., Torre, C. M., Rocha, A. M. A. C., Taniar, D., Apduhan, B. O., Stankova, E., and Cuzzocrea, A., editors, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 10405, pages 545–559, Cham. Springer International Publishing.

Delong, E. R., Delong, D. M., Clarke-pearson, D. L., and Carolina, N. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves : A Nonparametric Approach Published by : International Biometric Society Stable URL : http://www.jstor.org/stable/2531595 REFERENCES Linked references are available. 44(3):837–845.

Dorfman, D. D., Beavers, L. L., and Saslow, C. (1973). Estimation of signal detection theory parameters from rating-method data : A comparison of the method of scoring and direct search *. 1(3):207–208.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

Gagliardi, L., Cavazza, A., Brunelli, A., Battaglioli, M., Merazzi, D., Tandoi, F., Cella, D., Perotti, G. F., Pelti, M., Stucchi, I., Frisone, F., Avanzini, A., and Bellù, R. (2004). Assessing mortality risk in very low birthweight infants: a comparison of CRIB, CRIB-II, and SNAPPE-II. *Archives of Disease in Childhood - Fetal and Neonatal Edition*, 89(5):F419 LP – F422.

Gönen, M. (2001). Receiver Operating Characteristic (ROC) Curves. *Sugi 31*, pages 1–18.

Green, D. and Swets, J. (1966). Signal detection theory and psychophysics. *First ed. New York: John Wiley & Sons*.

Hajian-Tilaki, K. (2013). Receiver operating characteristic (ROC) curve analysis for medical diagnostic test evaluation.

Hanley, A. and McNeil, J. (1982). The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology*, 143:29–36.

Hanley, J. A. and McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 148(3):839–43.

Harvey, L. O. J. (2011). Detection Theory: - coursework. *Psych-Www.Colorado.Edu*, 4165-100.

Jensen, A. L., Th\ofner, M. T., and Iverasen, L. (1996). Application of receiver-operating-characteristic (ROC) curves to veterinary clinical pathology. *Comparative Haematology International*, 6(3):176–181.

Jiang Y, N., de Kort, G. A. P., Beijerinck, D., and Deurenberg, J. J. M. (1996). Malignant and Benign Clustered Microcalcifications: Automated Feature Analysis and Classification. *Radiology*, 201(2):581.

Kairisto, V. and Poola, A. (1995). Software for illustrative presentation of basic clinical characteristics of laboratory tests - GraphROC for Windows. *Scandinavian Journal of Clinical and Laboratory Investigation*, 55(sup222):43–60.

Knowles, J. D. and Corne, D. W. (2000). Approximating the nondominated front using the Pareto Archived Evolution Strategy. *Evolutionary computation*, 8(2):149–172.

Lusted, L. B. (1971). Signal Detectability and Medical Decision-Making. *Science*, 171(3977):1217–1219.

Marshall, G., Tapia, J. L., D'Apremont, I., Grandi, C., Barros, C., Alegria, A., Standen, J., Panizza, R., Roldan, L., Musante, G., Bancalari, A., Bambaren, E., Lacarruba, J., Hubner, M. E., Fabres, J., Decaro, M., Mariani, G., Kurlat, I., and Gonzalez, A. (2005). A new score for predicting neonatal very low birth weight mortality risk in the NEOCOSUR South American Network. *Journal of Perinatology*, 25(9):577–582.

Martinez, E. Z., Louzada-Neto, F., and Pereira, B. D. B. (2003). A Curva ROC para Testes Diagnósticos.

Martins, F. (2009). *Programação Orientada Aos Objectos Em JAVA*.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, 8(4):283–298.

Metz, C. E. (1986). *Statistical Analysis of ROC Data in Evaluating Diagnostic Performance*.

Metz, C. E., Herman, B. A., and Roe, C. A. (1998). Statistical Comparison of Two ROC-curve Estimates Obtained from Partially-paired Datasets. *Medical Decision Making*, 18(1):110–121.

Network, T. I. N. (1993). The CRIB (clinical risk index for babies) score: a tool for assessing initial neonatal risk and comparing performance of neonatal intensive care units. *The Lancet*, 342(8865):193–198.

Parry, G., Tucker, J., and Tarnow-Mordi, W. (2003). CRIB II: an update of the clinical risk index for babies score. *The Lancet*, 361(9371):1789–1791.

Pepe, M. S. (2004). *The satistical evaluation of medic tests for classification and prediction.* Oxford University Press.

Pollack, I. and Hsieh, R. (1969). Sampling variability of the area under the ROC-curve and of d'e. *Psychological Bulletin*, 71(3):161–173.

Pollack, M. M., Koch, M. A., Bartel, D. A., Rapoport, I., Dhanireddy, R., El-Mohandes, A. A. E., Harkavy, K., and Subramanian, K. N. S. (2000). A Comparison of Neonatal Mortality Risk Prediction Models in Very Low Birth Weight Infants. *Pediatrics*, 105(5):1051–1057.

Rifkin, M. D., Zerhouni, E. A., Gatsonis, C. A., Quint, L. E., Paushter, D. M., Epstein, J. I., Hamper, U., Walsh, P. C., and McNeil, B. J. (1990). Comparison of Magnetic Resonance Imaging and Ultrasonography in Staging Early Prostate Cancer. *New England Journal of Medicine*, 323(10):621–626.

Stanislaw, H. and Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, {&} Computers*, 31(1):137–149.

Stephan, C., Wesseling, S., Schink, T., and Jung, K. (2003). Comparison of eight computer programs for receiver-operating characteristic analysis. *Clinical Chemistry*, 49(3):433–439.

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics: collected papers*. Scientific psychology series. Lawrence Erlbaum Associates.

Yanyu, Z. (2010). *ROC analysis in diagnostic medicine*. PhD thesis, Jiangxi Normal University.