Universidade do Minho
Escola de Engenharia

Claudia Gouveia Rodrigues

Avaliação de desempenho de modelos de regressão logística multivariada através de curvas ROC num estudo de RN de muito baixo peso

Dissertação de Mestrado

Mestrado em Engenharia de Sistemas

Trabalho realizado sob a orientação da

Professora Doutora Ana Cristina da Silva Braga

outubro de 2020

AGRADECIMENTOS

No decorrer deste trabalho e nestes dois anos de mestrado, tive a oportunidade de poder contar com o apoio de diversas pessoas que, de uma forma direta ou indiretamente, me ajudaram a cumprir os meus objetivos e a concluir mais uma etapa da minha vida.

Assim, os meus sinceros agradecimentos vão, em primeiro lugar, à minha orientadora **Professora Doutora Ana Cristina Braga**, pela orientação e apoio prestado ao longo desta dissertação, pela amizade, pela disponibilidade e partilha de conhecimentos que me deu e pelas sugestões e críticas que me estimularam a querer saber sempre mais e a querer sempre fazer melhorar.

À minha família, nomeadamente ao meu **pai**, **mãe** ao meu **irmão**, pela compreensão e carinho dado ao longo deste percurso.

Ao grupo do **Registo Nacional do Muito Baixo Peso** pela cedência dos dados que possibilitaram a realização deste trabalho.

À **Ana Rita Antunes**, pela ajuda e partilha de conhecimento.

Aos meus **amigos**, pelo encorajamento e pela amizade que me têm dado ao longo destes anos.

E, por fim, ao meu **namorado** pelo incentivo, companheirismo e carinho dado durante todo este percurso.

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

Avaliação de desempenho de modelos de regressão logística multivariada através de curvas ROC num estudo de RN de muito baixo peso

RESUMO

A mortalidade neonatal é uma das principais preocupações da saúde a nível mundial, sendo que, um dos grupos de bebés que mais tem contribuído para tal desfecho diz respeito aos recém-nascidos de muito baixo peso. Deste modo, conhecer as variáveis que contribuem para a mortalidade destes recém-nascidos, assim como, calcular os seus riscos de morte tem-se tornado um assunto importante, pois poderão facilitar a tomada de decisão por parte dos profissionais de saúde no procedimento e tratamento a seguir. Neste contexto, surgiu o tema desta dissertação que tem como objetivo desenvolver um modelo de regressão logística, que funcione como um classificador, que irá prever se um determinado recém-nascido com peso inferior a 1500 gramas irá sobreviver ou falecer. Para isso, fez-se uso das ferramentas R e RStudio, e de uma base de dados, fornecida pelo Registo Nacional de Recém-Nascidos de Muito Baixo Peso, onde se encontram os dados recolhidos pelas unidades de cuidados intensivos neonatais do território português. Para se atingir este objetivo, começou por se fazer uma análise minuciosa à base de dados, seguida da construção de doze modelos diferentes. Para além disso, para facilitar a utilização do modelo final por parte dos profissionais de saúde, criou-se uma aplicação *web,* disponível em https://claudia-rodrigues.shinyapps.io/Previsao_do_risco_de_morte_em_RNMBP/, através da utilização do pacote *shiny* presente no RStudio. Os resultados obtidos revelaram que, o modelo escolhido para funcionar como um classificador é constituído por 9 variáveis, sendo elas: idade gestacional, comprimento ao nascer, corticoides pré-natais, sexo, média dos três índices de apgar (1°,5° e 10° minutos), reanimação insuflador, malformação congénita major, diagnóstico de enterocolite necrotizante (NEC), e administração de Ibuprofeno para tratamento de persistência de ductos arteriosos (PDA). Este modelo apresenta um bom ajuste aos dados e uma boa capacidade preditiva, sendo capaz de prever com 0,926 de certeza o estado de um recém-nascido de muito baixo peso. Relativamente à validação interna e externa, obteve-se valores de área abaixo da curva ROC de 0,891 e 0,797, respetivamente. O facto de apresentar melhores resultados preditivos que os indicadores CRIB e SNAPPE II, fazem dele uma possível ferramenta alternativa a utilizar nas unidades de cuidados intensivos neonatais.

Palavras-Chave: Curva ROC, Recém-nascidos de muito baixo peso, Regressão logística, *shiny*

Performance evaluation of multivariate logistic regression models through ROC curves in a study of very low birth weight newborns

ABSTRACT

Neonatal mortality is a major health concern worldwide, and one of the groups of babies that has contributed most to this outcome are the very low birth weight newborns. Thus, knowing the variables that most contribute to the mortality of these newborns, as well as calculating their risk of death has become an important issue, as they may facilitate decision making by health professionals in procedure and treatment to follow. In this context, the theme of this dissertation appeared, whose main objective is to develop a logistic regression model, which works as a classifier and which will predict whether a particular newborn weighing less than 1500 grams will survive or die. For this, the tools R and RStudio were used, as well as a database, provided by the National Registry of Newborns of Very Low Weight, which include the data collected by neonatal intensive care units in the Portuguese territory.

To achieve this goal, a thorough analysis of the database was started, followed by the construction of twelve models. In addition, to facilitate the use of the final model by health professionals, a web application was created, available in https://claudia-rodrigues.shinyapps.io/Previsao_do_risco_de_morte_em_RNMBP/, using the *shiny* package present in RStudio. The results obtained revealed that the model chosen to work as a classifier is made up of 9 variables, namely: gestational age, length at birth, prenatal corticosteroids, sex, average of the three apgar scores (1st, 5th and 10th minutes), insufflator resuscitation, major congenital malformation, diagnosis of necrotizing enterocolitis (NEC) and administration of Ibuprofen for treatment of persistent ductus arteriosus (PDA). This model presents a good fit to the data and a good predictive capacity, being able to predict with 0.926 certainty the state of a very low birth weight newborn. Regarding internal and external validation, area under the ROC curve of 0.891 and 0.797 were obtained, respectively. In addition, the fact that it presents better predictive results than the CRIB and SNAPPE II indicators, makes it a possible alternative tool to be used in neonatal intensive care units.

Keywords: Logistic regression, ROC curve, *shiny*, Very low birth weight newborns

ÍNDICE

LISTA DE ABREVIATURAS E SIGLAS

AIC - *Akaike information criterion*

AUC - *Area Under the Curve*

BIC - *Bayesian information criterion*

CART - *Classification and Regression Tree*

CRIB - *Clinical Risk Index for Babies*

CRIB II - *Clinical Risk Index for Babies II*

kNN - *k-Nearest Neighbors*

LR – *Likelihood ratio test*

MAR - *Missing at random*

MCAR - *Missing completely at random*

MFV - Máximo da função de verossimilhança

NEOCOSUR - *Neonatal del Cono Sur*

NICHHD - *National Institute of Child Health and Human Development*

NMAR - *Not missing at random*

NTISS - *National Therapeutic Intervention Scoring System*

RN – Recém-nascidos

RNMBP - Recém-nascidos de muito baixo peso

RNRNMBP – Registo Nacional do Recém-Nascido de Muito Baixo Peso

ROC - *Receiver Operating Characteristic*

SNAP - *Score for Neonatal Acute Physiology*

SNAP II – *Score for Neonatal Acute Physiology II*

SNAPPE - *Score for Neonatal Acute Physiology Perinatal Extension*

SNAPPE II – *Score for Neonatal Acute Physiology Perinatal Extension II*

TISS - *Therapeutic Intervention Scoring System*

LISTA DE FIGURAS

LISTA DE TABELAS

1. INTRODUÇÃO

1.1 Motivação

A mortalidade neonatal é uma das principais preocupações da saúde a nível mundial, sendo que, 45% do total de mortes que tem ocorrido em crianças com menos de cinco anos, tem ocorrido no período neonatal (Nayeri et al., 2019). Além disso, o baixo peso ao nascer tem sido apontado como uma das causas mais comuns de mortalidade em recém-nascidos. Um grupo de bebés que mais tem contribuído de forma significativa para a mortalidade neonatal diz respeito aos recém-nascidos de muito baixo peso (RNMBP), que são aqueles que apresentam um peso abaixo de 1500g e/ou menos de 32 semanas de gestação (M. F. Mourão et al., 2014). Estima-se que, só em Portugal, estes bebés representem cerca de 1% dos nados-vivos, ou seja, o que corresponde 1000 por ano, e que, mundialmente a mortalidade neonatal é 20 vezes mais provável de ocorrer em recém-nascidos com baixo peso do que em recém-nascidos com peso normal (Cunha et al., 2010; Raja et al., 2017). Por outro lado, muitos destes bebés que acabam por sobreviver aquando à alta hospitalar, acabam por apresentar complicações graves que têm impacto na qualidade de vida dos mesmos, o que acaba por exigir uma quantidade significativa de recursos médicos.

Nesse sentido, de forma a se conseguir obter informações cada vez mais rigorosas, os investigadores têm vindo a identificar escalas de gravidade clínica cada vez mais precisas que funcionam como métodos de classificação de estado de qualidade de bebés. Entre os indicadores existentes, o mais utilizado para analisar mais objetivamente a perspetiva de sobrevivência e a qualidade de vida dos recém-nascidos de muito baixo peso é o *Clinical Risk Index for Babies* (CRIB) (Sarquis et al., 2002).

O simples facto de se entender quais poderão ser os fatores que contribuem para a mortalidade dos bebés RNMBP, de entre as variáveis que são utilizadas rotineiramente pelos médicos, poderá ajudar os profissionais de saúde a tomarem as melhores decisões no planeamento da assistência perinatal e durante cada estágio da terapia intensiva neonatal.

Uma abordagem típica para identificar os principais fatores de risco de morte passa pelo uso de modelos preditivos de regressão logística, sendo que, estes modelos permitem igualmente estimar a probabilidade de um determinado recém-nascido vir a sobreviver ou falecer. Para que seja possível realizar esta classificação, será necessário, após a seleção do melhor conjunto de variáveis independentes a incluir no modelo, avaliar a significância estatística dos parâmetros estimados e validar o modelo resultante a

nível interno, quando se utiliza a amostra original do estudo, e a nível externo, quando se utiliza dados de participantes diferentes dos utilizados no desenvolvimento do modelo (A. C. Braga & Carneiro, 2016). Esta validação a nível interno e externo do modelo desenvolvido pode ser realizada através da análise da curva ROC (*Receiver Operating Characteristic*), sendo a área sob a curva ROC uma medida eficaz da validade inerente desse modelo (M. F. Mourão & Braga, 2012).

Por outro lado, muitos dos profissionais de saúde, de modo a obterem uma avaliação mais rápida das suas avaliações recorrem a ferramentas de análise, muitas vezes comerciais, que fornecem métodos básicos para analisar as curvas ROC, contudo acabam por apresentar recursos de análise limitados (Goksuluk et al., 2016). Em contra partida, existem pacotes de *software* livre, como o R, que oferecem recursos de análise avançados, porém mais difíceis de se utilizar pois requerem o uso de uma interface de utilizador baseado em comandos, o que se torna desafiador para utilizadores que não têm conhecimento da linguagem R. Nesse sentido, tem-se desenvolvido aplicações da *web* baseadas na linguagem R no qual não é necessário escrever qualquer linha de código. Isto é possível através da aplicação do pacote *shiny* presente no RStudio.

As principais motivações para o desenvolvimento deste trabalho foram a constante preocupação de se avaliar de uma forma precisa o risco de morte neonatal destes recém-nascidos de muito baixo peso, que até então existe escassez de informação e o facto do modelo de regressão logística ter vindo a ser muito utilizado na construção de classificadores clínicos. Para além disso, a questão da metodologia ROC ter vindo a ser muito utilizada na avaliação e comparação destes testes de diagnóstico e ao crescente interesse no desenvolvimento de ferramentas da *web* de análise rápida, gratuita e de fácil uso, como as aplicações *Shiny*, também foram outros grandes motivos para o desenvolvimento deste trabalho. Assim, com esta dissertação pretende-se desenvolver um modelo de regressão logística que sirva como classificador, de modo a prever se um dado bebé de muito baixo peso irá sobreviver ou falecer.

1.2 Objetivos

O objetivo principal desta dissertação passa por responder à seguinte questão: "Será possível avaliar se existem classificadores que possam ser utilizados para prever o risco de morte ao nascer de recém-nascidos de muito baixo peso, tendo em conta a base de dados fornecida?".

Para se responder a esta questão, procedeu-se à construção de modelos de regressão logística, tendo para isso, utilizado dados de teste e de treino provenientes da base de dados que foi cedida pelo Registo Nacional do Recém-Nascido de Muito Baixo Peso (RNRNMBP). Para além disso, também se utilizou o

software R, no qual foi feito todo o processo de desenvolvimento do modelo de regressão logística, e o pacote *shiny* presente no R, de forma a criar uma aplicação *web* na qual se implementou o modelo desenvolvido.

Assim, para que se conseguisse atingir os objetivos deste trabalho, foi necessário, numa primeira fase, responder a algumas perguntas, tais como:

- ✓ Como organizar a base de dados?
- ✓ Como construir um modelo explicativo para a sobrevivência/morte de um recém-nascido de muito baixo peso?
- ✓ Quais as variáveis a selecionar para melhorar o processo preditivo?

1.3 Estrutura da Dissertação

Esta dissertação desenvolve-se ao longo de 7 capítulos.

No **Capítulo 1** é apresentado uma introdução ao tema com a respetiva motivação, os objetivos, assim como, a estrutura geral da dissertação.

No **Capítulo 2** apresenta-se o estado de arte, no qual é revista a bibliografia sobre a importância do estudo de recém-nascidos de muito baixo peso e o que existe atualmente para os avaliar a nível de risco de morte, a utilização da metodologia ROC, assim como, o surgimento do interesse na construção de aplicações *web* através do pacote *shiny* presente no RStudio.

No **Capítulo 3**, é apresentado a fundamentação teórica dos conceitos relativos aos modelos de regressão logística. Começa-se por fazer uma breve introdução ao tema, explicando em que consiste uma regressão, em que áreas do saber a regressão logística é aplicada, como esta se distingue da regressão linear e qual a principal diferença entre uma regressão logística univariada e multivariada. De seguida, explorou-se os vários tipos de tratamentos de valores ausentes que se possam realizar numa base de dados a utilizar para a construção de um modelo de regressão logístico. Posto isto, analisou-se mais a fundo os modelos multivariados, seguido de formas de estimação dos parâmetros do modelo, como testar a significância de coeficientes e métodos de otimização que se possam utilizar para selecionar as variáveis a incluir nos modelos. Por fim, investigou-se como é feita uma interpretação de um modelo de regressão logística tendo em conta o tipo de variável em estudo.

No **Capítulo 4**, é apresentado a fundamentação teórica relativa à classificação e previsão em modelos de regressão logística, começando com uma breve contextualização em que consiste um classificador, qual a sua aplicabilidade, como se constroem e como se analisam os seus desempenhos. De seguida, são

apresentadas algumas medidas de diagnóstico utilizadas em modelos de previsão, de forma a avaliar as suas capacidades preditivas e os seus ajustes aos dados em estudo. Por fim, investigou-se como é realizado um diagnóstico de pontos influentes e de *outliers* nesse tipo de modelos.

No **Capítulo 5**, é apresentado a fundamentação teórica relativa ao desenvolvimento de uma aplicação *web*, utilizando para isso o pacote *shiny* presente no *software* R. Começa-se por fazer uma descrição no que consiste a ferramenta R, assim como, alguns pacotes que este *software* oferece. Por fim, faz-se um pequeno enquadramento de estudos onde se construíram aplicações *Shiny*, como se estrutura esses tipos de aplicações e algumas características que apresentam.

No **Capítulo 6**, é analisada em pormenor a amostra em estudo, quer a nível geográfico, quer a nível estatístico, e também é apresentado alguns exemplos de escalas de gravidade clínica utilizadas para classificar o estado de qualidade dos bebés recém-nascidos de muito baixo peso. Neste capítulo também se encontra explicado os passos tomados na realização do tratamento da base de dados e dos valores omissos, a construção do modelo que funcionará como classificador, a identificação de possíveis pontos influentes, a sua interpretação ao nível dos coeficientes estimados, o cálculo de algumas previsões com o modelo desenvolvido e a comparação do classificador construído com outros índices de gravidade. Por fim, explica-se em que consiste e como se encontra estruturada a aplicação *web* desenvolvida, acompanhado de alguns exemplos de predições realizadas.

Por fim, no **Capítulo 7**, apresenta-se a conclusão do trabalho desenvolvido, assim como, as propostas para trabalho futuro.

2. ESTADO DE ARTE

Apesar dos recentes avanços na medicina intensiva neonatal, a mortalidade infantil continua a ser um dos principais problemas da saúde pública dos países mais desenvolvidos. Nas populações neonatais, os recém-nascidos prematuros e com baixo peso são considerados os grupos mais vulneráveis, sendo que, os recém-nascidos de muito baixo peso (RNMBP) são os que mais contribuem para altas taxas de mortalidade no primeiro ano de vida (Carneiro et al., 2012; M. F. Mourão & Braga, 2012). Considera-se um recém-nascido de muito baixo peso quando este tiver menos de 32 semanas de gestação e/ou menos de 1500g de peso à nascença (M. F. Mourão et al., 2014).

Embora os RNMBP representam cerca de 1% de todos os nascimentos, esta população é responsável por 50% das mortes neonatais (Jafrasteh et al., 2017). Já em Portugal, cerca de 1000 recém-nascidos de muito baixo peso morrem por ano (Cunha et al., 2010).

Para além disso, segundo Cutland et al. (2017), os recém nascidos que apresentem um baixo peso, estão sujeitos a apresentar diversas morbidades, como incapacidade neurológica, aumento de risco de doenças crónicas como doenças cardiovasculares e diabetes, maior probabilidade de ocorrência de hemorragia intracraniana, dificuldade respiratória, cegueira, distúrbios gastrointestinais, entre outros. Essas condições deixam estes bebés mais vulneráveis ao óbito.

Representando assim, uma porção da população que suscita uma notável preocupação para a saúde pública, tem sido necessário encontrar uma solução adequada de modo a reduzir a taxa de mortalidade desta população. Para isso, tem sido essencial a recolha de estatísticas atuais e confiáveis de modo a se conseguir entender quais os fatores que poderão estar associados à mortalidade desses bebés.

Entender quais poderão ser as causas da morte nos recém-nascidos de muito baixo peso ao nascer, têm sido apontados como fatores importantes para que os profissionais de saúde consigam tomar as melhores decisões no planeamento da assistência perinatal e durante cada estágio da terapia intensiva neonatal, como por exemplo, quando devem intervir na cesariana, se devem tentar ressuscitar e iniciar a ventilação mecânica e quando se deve interromper um tratamento (Kardum et al., 2019; Medlock et al., 2011). Assim, ter um prognóstico preciso pode ajudar na tomada dessas decisões difíceis. Consequentemente a previsão da mortalidade destes recém-nascidos é de extrema importância.

Neste sentido, nas últimas décadas, a avaliação do risco de morte, utilizando variáveis que aparentam interferir nas taxas de mortalidade, tem sido cada vez mais objeto de estudo em diversas unidades neonatais (Brito et al., 2003). Apesar de no início somente o peso ao nascer e a idade gestacional terem sido consideradas como as únicas variáveis significativas de mortalidade neonatal, estudos mais recentes

têm vindo a considerar outras variáveis como possíveis fatores de risco de morte, como o sexo masculino, malformações congénitas, o não uso de esteroides pré-natais, tipo de parto, entre outros (Ballot et al., 2010; Vincer et al., 2015). Neste seguimento, escalas de gravidade, que agregam parâmetros fisiológicos que refletem o estado clínico inicial de recém-nascidos, tem sido desenvolvido de forma a avaliar o risco de morte destes indivíduos (Brito et al., 2003). Entre as diversas escalas de gravidade existentes, algumas são mais simples, apresentando um número de variáveis mais reduzido e cuja aplicabilidade é mais rápida e outras mais complexas, pois são constituídas por um maior número de variáveis e a sua aplicabilidade é mais demorada.

Segundo a literatura, sistemas de pontuação que têm vindo a ser mais estudadas e utilizadas em recém-nascidos de muito baixo peso são o CRIB (*Clinical Risk Index for Babies)* e o SNAPPE (*Score for neonatal acute physiology—perinatal extension*) (Gagliardi et al., 2004). Estas escalas têm vindo a ser validadas e reaplicadas em estudos distintos em diferentes países, sendo capazes de permitir comparações justas de resultados em diferentes unidades hospitalares. Porém, estas escalas, que foram desenvolvidas numa década onde não se fazia uso generalizado de surfactantes e esteroides pré-natais, apresentam algumas limitações, o que levou com que as suas adequações têm vindo a ser questionadas. Por exemplo, um dos parâmetros postos em causa no CRIB têm sido a fração de oxigénio inspirada, pois acredita-se que esta não é uma medida fisiológica verdadeira pelo facto de ser determinada pela equipa de assistência (Jašić et al., 2016). Outra questão posta em causa relaciona-se com os dados que são recolhidos até às 12 horas após a admissão por aparentemente introduzirem enviesamento no tratamento precoce.

Por outro lado, atualmente é cada vez mais frequentemente que os profissionais de saúde, assim como, os pais destes bebés acabem por exigir estimativas para as chances de sobrevivência destes recém-nascidos, principalmente quando estes se deparam com a decisão de iniciar ou suspender um certo tratamento (Jeschke et al., 2016). Assim, daqui tem surgido o interesse na construção de classificadores que acabam por estimar a probabilidade de ocorrer um certo acontecimento, neste caso em particular, se falece ou sobrevive.

Apesar de ainda não existir muita informação na literatura, vários tipos de modelos de predição de mortalidade têm sido desenvolvidos para estimar o risco de morte de recém-nascidos de muito baixo peso. Medlock et al. (2011) na sua revisão examinou publicações existentes cujos temas se relacionavam com modelos de predição de mortalidade em bebés de muito baixo peso. Com isto, pretendeu identificar variáveis que são frequentemente significativas em modelos multivariados e avaliar a qualidade desses modelos de forma a poder fazer recomendações para pesquisas futuras. A maioria dos estudos examinados faziam uso de modelos de regressão logística, sendo que uma minoria fazia uso de outros

tipos de modelos de predição como modelos de rede neuronal e modelos CART (*Classification and Regression Tree*).

Vincer et al. (2015) também fizeram uso da regressão logística na construção de um modelo preditivo para a mortalidade neonatal, tendo para isso, utilizado informações disponíveis no período pré-natal dos RNMBP. Neste estudo, conclui-se que de facto, a previsão da probabilidade de mortalidade neonatal é influenciada por fatores maternos e fetais, tais como um baixo valor de idade gestacional, distúrbios psiquiátricos maternos, antibioticoterapia pré-natal e gémeos monocorionicos. Já no estudo liderado por De Castro et al. (2016) chegaram à conclusão que a alta taxa de mortalidade nos RNMBP no primeiro dia de vida está relacionada com algumas variáveis biológicas, tais como, o peso e o sexo do bebé, assim como, uma baixa vitalidade ao nascer e piores infraestruturas do hospital onde nasceram.

Contudo, antes que um classificador seja adotado nos cuidados de saúde, é necessário determinar a sua capacidade preditiva, de forma a evitar que este realize uma classificação errada de um determinado recém-nascido. Posto isto, é usual demonstrar o desempenho de um classificador clínico pela análise da curva ROC (*Receiver Operating Characteristic*).

A análise ROC foi originalmente desenvolvida durante a Segunda Guerra Mundial de modo a detetar se um determinado sinal na tela do radar representava um objeto ou um ruído (Goksuluk et al., 2016). Posteriormente, essa metodologia começou por ser amplamente utilizada em outros campos da ciência como na medicina, radiologia, bioinformática e em várias aplicações de *machine learning* e *data mining*. Na medicina, foi utilizada pela primeira vez por Lusted, mais precisamente na área da radiologia, nos finais da década de 60 (Swets, 1996). Por outro lado, na área da neonatologia também tem surgido vários estudos em que se tem feito uso da análise da curva ROC. Por exemplo, Brito et al. (2003) pretenderam no seu estudo avaliar a taxa de mortalidade em RNMBP de acordo com as variações nos *scores* CRIB, peso ao nascer e idade gestacional. Para isso, fizeram uso da metodologia ROC de forma a avaliar a capacidade preditiva de cada um e verificar qual deles seria capaz de fazer uma melhor previsão do risco de morte destes recém-nascidos. De facto, devido à sua capacidade discriminativa, a metodologia ROC é uma ferramenta frequentemente utilizada em processos decisórios médicos (A. C. da S. Braga, 2000).

Entende-se por curva ROC como sendo uma abordagem gráfica no qual os seus eixos são representados pelas frações de verdadeiros positivos de um classificador (sensibilidade), no eixo das ordenadas, e pelas frações de falsos positivos (1-especificidade), no eixo das abcissas, em que cada ponto é gerado por um valor limite diferente (Goksuluk et al., 2016). Assim, pode-se considerar que, a maior vantagem da curva ROC reside na sua simplicidade, visto que se trata de uma representação visual direta do desempenho

de um classificador. Para além disso, a análise ROC também permite a visualização de vários classificadores, de modo a selecionar o melhor modelo preditivo com base nas suas medidas de desempenho.

Um dos indicadores mais utilizados nesta metodologia, que permite avaliar o desempenho dos modelos, é a área abaixo da curva ROC, também conhecido por AUC (*Area Under the Curve*) (Fawcett, 2006). Segundo Fawcett (2006), a AUC de um classificador é equivalente à probabilidade de o classificador ter uma maior habilidade para, aleatoriamente, escolher um caso positivo corretamente classificado do que um caso negativo. Para que um modelo de classificação apresente um perfeito poder discriminativo, este deve apresentar um valor de AUC de 1 (Zou et al., 2007).

Posto isto, pode-se dizer que a utilização de modelos preditivos na hora da tomada de decisão por parte dos profissionais de saúde, apresenta diversas vantagens, visto que, um modelo devidamente treinado e validado, consegue ter em conta muito mais fatores que um simples ser humano. Para além disso, a utilização destes modelos evita a situação de que, perante um mesmo diagnóstico, os médicos tomem diferentes decisões, principalmente se se tratar de um médico inexperiente.

Contudo, o facto de muitos dos profissionais não se sentirem à vontade em utilizar esse tipo de modelos preditivos ou até mesmo por sentirem que se trata de um processo mais demorado para obter uma avaliação mais rápida das suas avaliações, muitos desses profissionais recorrem a ferramentas de análise comerciais. No entanto, muitos desses *softwares* oferecem uma interface de utilizador baseada em comandos, tornando o processo de análise mais complicado para utilizadores que não têm conhecimento de programação. Consequentemente, começou-se por desenvolver novas ferramentas da *web* de fácil uso, sem ser necessário escrever linhas de código. Exemplos de ferramentas são aplicações criadas por meio da linguagem R e do pacote *shiny* presente no RStudio (Goksuluk et al., 2016).

Por exemplo, Goksuluk et al. (2016) desenvolveram uma ferramenta *web* conhecida por "easyROC", que é de fácil uso e é construída sob a linguagem R, sendo que teve como prepósito, apoiar os pesquisadores nas suas decisões sem necessitarem de escrever qualquer código em R. Esta ferramenta fornece estatísticas relacionadas com a análise ROC e ferramentas gráficas e permite o cálculo de pontos de corte ideais, assim como, a comparação entre vários indicadores. Já no campo da genética, Criscuolo & Angelini (2020) construíram a ferramenta "StructuRly" que permite realizar análises genéticas populacionais pelo agrupamento de população, avaliando vários índices genéticos e comparando os resultados. Além disso, "StructuRly" apresenta representações gráficas interativas que podem ser igualmente personalizadas. Por sua vez, Peng et al. (2020) desenvolveram uma aplicação *Shiny* no qual implementaram um classificador, construído segundo o algoritmo *Random Forest*, que faz uma triagem

neonatal para distúrbios metabólicos congénitos, melhorando a previsão de verdadeiros e falsos positivos.

Neste sentido, como em Portugal os estudos que abordam esta problemática são ainda reduzidos e se deseja encontrar a melhor forma de se conseguir diminuir o número de taxa de mortalidade dessa amostra de recém-nascidos no território nacional, pretende-se com este trabalho desenvolver um classificador baseado em dados que caracterizam a realidade portuguesa. Para além disso, verificada a ampla utilização da metodologia ROC na avaliação e comparação de classificadores, esta metodologia será aplicada na elaboração desta dissertação de modo a determinar a capacidade de desempenho do modelo preditivo a desenvolver, assim como, comparar o classificador desenvolvido com dois indicadores de gravidade clínica, CRIB (*Critical Risk Index for Babies*) e SNAPPE-II (*Score for neonatal acute physiology—perinatal extension*), que são os sistemas mais utilizados e foram relatados como sendo índices com um bom desempenho na avaliação de risco de morte para bebés com muito baixo peso (Medvedev et al., 2020; M. F. Mourão et al., 2014). Por fim, dada a fácil utilização de aplicações *Shiny* e de forma a facilitar o uso do classificador por parte dos profissionais da saúde, pretende-se criar uma aplicação *web*, utilizando para isso o pacote *shiny* presente no *software* R. Esta aplicação irá permitir avaliar, em tempo real, a probabilidade de um dado recém-nascido com muito baixo peso vir a falecer, tendo em conta alguns fatores de risco de mortalidade determinados significativos para o modelo encontrado.

3. REGRESSÃO LOGÍSTICA BINÁRIA

Cada vez mais os modelos de previsão multivariáveis são desenvolvidos, validados e implementados em diferentes áreas do conhecimento, tendo como objetivo auxiliar os profissionais nas suas tomadas de decisão. Tendo esta ideia em mente, pretende-se com este trabalho desenvolver e validar um modelo de regressão logística que prediz o risco de mortalidade de bebés de muito baixo peso. Assim, numa primeira fase, será apresentado ao longo deste capítulo uma revisão da literatura relativa ao modelo de regressão logística, mais precisamente ao modelo de regressão logística multivariado. Será abordado algumas técnicas que se possam utilizar para tratar uma base de dados com valores omissos, que será utilizada para a construção de um modelo de regressão logística, como estimar os parâmetros do modelo e como testar a significância de coeficientes, assim como, métodos automáticos para selecionar as variáveis a incluir nos modelos. Por fim, será abordado como interpretar os modelos desenvolvidos de acordo com o tipo de variáveis independentes que fazem parte.

3.1 Introdução

Ao se utilizar recursos matemáticos e estatísticos que são fornecidos pela análise de regressão, é possível encontrar-se funções matemáticas que são capazes de estimar o comportamento de um conjunto de dados (da Silva, 2011).

O método de regressão, que tem vindo a ser um componente integral de qualquer análise de dados, é um modelo estatístico que descreve a relação entre a varável dependente/resposta com uma ou mais variáveis independentes/exploratórias (David W. Hosmer & Lemeshow, 2000).

Existe uma grande variedade de modelos estatísticos utilizados em problemas de previsão, desde modelos mais simples, como os modelos lineares, até modelos mistos, mais complicados. Ao longo das últimas décadas, o modelo de regressão logística tem vindo a ser um dos métodos de análise mais utilizados nesta situação, sendo usado principalmente em estudos de diagnóstico e de curto prazo (A. C. Braga & Carneiro, 2016). No entanto, a aplicação da regressão logística não é somente feita numa perspetiva preditiva, mas também numa perspetiva de análise de relações entre variáveis.

A regressão logística foi originalmente aplicada na área epidemiologia, sendo hoje em dia empregue igualmente em outros ramos como as ciências biomédicas, finanças, criminologia, ecologia, engenharia, línguas, biologia, entre outros (David W. Hosmer & Lemeshow, 2000). Por exemplo, no ramo da economia, Zaghdoudi (2013) tentou desenvolver um modelo preditivo de falência bancária da Tunísia

com a contribuição do método de regressão logística binária tendo em consideração indicadores microeconómicos de falências bancárias. Através do modelo desenvolvido, Zaghdoudi verificou que a capacidade do banco de pagar sua dívida, o coeficiente de operações bancárias, a rentabilidade do banco por funcionário e alavancar o índice financeiro têm um impacto negativo na probabilidade de falha. Por outro lado, Considine & Zappalà (2002) apresentam um modelo de regressão logístico para avaliar a influência da desvantagem social e económica no desenvolvimento académico de estudantes da Austrália. Na medicina, Asoglu et al. (2020) recorreram a um modelo de regressão logística de modo a conseguirem responder à questão se o tipo de parto realizado a bebés, que apresentaram uma gestação complicada por defeitos cardíacos congénitos fetais, tem um impacto significativo no resultado neonatal. Assim, com o modelo de regressão logística pretenderam avaliar os fatores associados a cada tipo de parto e o seu impacto nos resultados neonatais. Também na área da saúde, Sathar et al. (2018) aplicaram uma análise de regressão logística para avaliar quais seriam os fatores de risco associados à doença retinopatia em recém-nascidos de muito baixo peso à nascença, tendo concluído que o baixo peso ao nascer, idade gestacional, apneia, transfusão sanguínea, doença da membrana hialina, fototerapia, suporte ventilatório e uso de oxigénio por mais de sete dias seriam os fatores de risco significativos. Por outro lado, Basu et al. (2008) construíram um modelo de regressão logística de modo a determinar que fatores estariam associados à mortalidade de recém-nascidos de muito baixo peso na Índia.

Segundo Hosmer e Lemeshow (2000), enquanto que nos modelos de regressão linear a variável resposta é uma variável aleatória de natureza contínua, sendo esta em alguns casos qualitativa e expressa em função de duas ou mais variáveis de natureza categórica, ou seja, admite dois ou mais valores, no modelo de regressão logística a variável resposta é normalmente binária ou dicotómica. Isto exige que o resultado da análise possibilite associações a certas categorias, como por exemplo, positivo ou negativo, morrer ou sobreviver, doente ou não doente, entre outros. Esta diferença que existe entre a regressão logística e a linear deve-se à escolha do modelo paramétrico e nas hipóteses a serem consideradas. Todavia, as técnicas utilizadas em ambos os métodos são muito semelhantes.

Apesar de um modelo de regressão logística possibilitar a classificação de fenómenos ou indivíduos em categorias específicas, ela também permite estimar a possibilidade de ocorrência de um determinado acontecimento ou evento em termos de probabilidade, o que neste caso, a saída deve circunscrever todos os resultados que se possam atribuir á variável dependente ao intervalo compreendido entre zero e um.

Considera-se um modelo de regressão logística univariado quando existe apenas uma variável independente, e multivariado quando existe duas ou mais. É de salientar que será dado mais ênfase à regressão logística multivariada, uma vez que, neste trabalho será esse o método a ser aplicado à base de dados em estudo.

3.2 Tratamento de Valores Omissos

Quando se pretende construir um modelo preditivo, uma das primeiras etapas a considerar diz respeito ao tratamento e preparação da base de dados a utilizar. Nesta fase, é necessário explorar cuidadosamente a amostra de dados de modo a tratar variáveis mal codificadas, reconhecer observações que serão irrelevantes para o estudo, assim como, identificar valores ausentes. Esta etapa é considerada como uma das partes mais importantes e mais dispendiosas, em termos de tempo, do processo de construção de um modelo.

Em todos os tipos de investigação, mesmo em estudos bem projetados e controlados, um dos problemas que é mais frequente ocorrer, durante o processo de tratamento de uma base de dados, diz respeito ao surgimento de valores omissos.

Esta falta de dados pode ocorrer por diversas razões, como por exemplo, a pessoa que fez o registo dos dados esqueceu-se de responder ou passou à frente uma pergunta, existiu falta de informação necessária para responder a determinada questão, ou então, por exemplo em contexto médico, ocorrência de morte de um indivíduo que fazia parte de um determinado estudo.

Dados ausentes podem reduzir o poder estatístico de um determinado estudo e consequentemente levar a conclusões inválidas. Isso também é valido para estudos que envolvem o desenvolvimento de classificadores, sendo que, a presença de valores omissos em um conjunto de dados a utilizar como amostra de treino, pode afetar o desempenho do mesmo (Acuña & Rodriguez, 2004).

Segundo Acuña & Rodriguez (2004), quando se fala em ausência de dados durante uma análise estatística, estas podem ser consideradas triviais quando representam menos de 1%, valores ausentes gerenciáveis quando correspondem a 1-5%, valores omissos que exigem métodos sofisticados para se conseguir lidar com eles quando correspondem a cerca de 5-15% e quando se tratam de valores omissos que podem afetar seriamente qualquer tipo de interpretação quando representam mais de 15%.

Para se conseguir lidar da melhor forma possível, com a presença de valores ausentes, existem algumas etapas a ter em consideração (Salgado et al., 2016):

1. Identificar padrões e razões para a existência de valores ausentes;

2. Analisar a proporção de dados omissos;
3. Escolher o melhor método de imputação.

Relativamente aos diferentes tipos de dados ausentes, Rubin (1976) descreveu e dividiu os dados omissos em três principais grupos, de acordo com o motivo:

- MCAR (*Missing completely at random*): O mecanismo que gera a não resposta é completamente aleatória, sendo que, a causa da omissão de dados não se encontra associado a nenhuma escolha estratégica do sujeito respondente (Curley et al., 2019). Além disso, o padrão de valores omissos observado não está relacionado a nenhum outro dado presente ou ausente. Um exemplo deste tipo de dado pode ser quando um indivíduo acidentalmente se esquece de responder a uma pergunta. No MCAR, pode-se considerar os casos com dados ausentes como uma subamostra aleatória simples de todos os casos presente no conjunto de dados, no qual, quando se remove uma subamostra aleatória simples da amostra total, a amostra restante ainda representa a população quanto a amostra original (van Ginkel et al., 2019).

  Uma das formas para verificar se os dados são MCAR pode passar por se fazer a pergunta "Existe algum motivo teórico para que o respondente possa querer evitar de responder a essa pergunta?" ou então, recorrer ao teste MCAR de Little (1988), no qual se calcula um teste qui-quadrado para examinar padrões em falta para um número de variáveis especificas. No teste qui-quadrado, a suposição nula é que os dados são MCAR, sendo que não rejeita a hipótese nula quando o valor é maior que 0,05.
- MAR (*Missing at random*): Trata-se de dados que podem ser previstos através de variáveis observadas, ou seja, o valor dos dados ausentes depende do valor das respostas observadas mas não dos valores que estão em falta (Curley et al., 2019). No MAR, é provável que seja mais frequente encontrar em alguns subgrupos os dados ausentes do que em outros, porém as informações que definem os subgrupos são observadas em todos os indivíduos (van Ginkel et al., 2019).
- NMAR (*Not missing at random*): Neste tipo de dados, não é possível aproximar os valores ausentes, pois também não são conhecidos os valores de outras variáveis relevantes (Curley et al., 2019). Trata-se de dados NMAR quando a probabilidade da ausência dos dados estiver relacionada com o valor desconhecido da variável (Henry et al., 2013).

Esta taxonomia torna-se útil, no sentido que, fornece informações sobre qual ferramenta será a mais adequada para lidar com cada tipo de dados ausentes. Vários métodos foram propostos para tratar de

valores omissos, sendo que, entre os métodos mais utilizados, podemos enquadrá-los em três principais categorias (Salgado et al., 2016):

- Métodos de exclusão: *listwise deletion, pairwise deletion*;
- Métodos de imputação simples: substituição pela média ou mediana, interpolação linear, *Hot deck* e *Cold deck*;
- Métodos baseados em modelo: regressão, imputação múltipla, k-vizinhos mais próximos.

Provavelmente, a abordagem usada com maior frequência é a *listwise deletion,* que simplesmente exclui os casos que contêm pelo menos um valor ausente e analisa os restantes dados (Kang, 2013). Devido à sua frequente utilização, acabou por se tornar a opção padrão para análise adotado pela maioria dos pacotes de *software* estatísticos. Contudo, alguns autores afirmam que esta abordagem poderá introduzir uma distorção na estimativa dos parâmetros.

A aplicação da *listwise deletion* tem apresentado resultados satisfatórios quando se lida com dados ausentes que são MCAR e quando a amostra permanece grande após a exclusão de observações individuais (Curley et al., 2019). A exclusão de dados omissos do tipo MCAR é menos consequente, pois se a falta de valor for completamente aleatória, os dados excluídos também serão aleatórios e não causarão a perda de variação importante. Normalmente, é considerado seguro optar por recorrer a esse método quando a quantidade de valores omissos corresponder aproximadamente, menos de 5% do número total de casos.

Técnicas como, eliminar uma variável, geralmente são consideradas como uma boa medida quando nos deparamos com um número excessivo de valores omissos, como por exemplo, mais de 50%, contudo, não é um procedimento isento de riscos (Salgado et al., 2016).

Outras abordagens bastante utilizadas para lidar com valores ausentes envolvem imputação. A ideia principal deste termo é que, se um dado importante estiver ausente para uma variável específica, é possível estimá-lo a partir de dados existentes (Salgado et al., 2016). Os métodos mais simples de imputação passam por substituir os valores em falta pela média, moda ou pela mediana da variável em questão, sendo mais eficaz o uso da mediana na presença de valores observados que sejam discrepantes, ou seja, de *outliers*. Apesar de a imputação pela média reduzir a variação da amostra, esta técnica poderá ser apropriada quando se está perante uma situação em que o grau de falta de dados é pequeno e o tamanho da amostra é grande (Curley et al., 2019). Se estiver perante um caso em que o valor em falta se insere numa variável categórica, a melhor opção será recorrer ao uso da imputação simples pela moda (Acuña & Rodriguez, 2004). Contudo, é preciso ter em atenção que, a utilização desse tipo de imputação na presença de variáveis com uma alta correlação entre si, isto é, que

apresentam informações semelhantes, ou então, que apresentam um poder de predição semelhantes, pode-se tornar inútil.

Para variáveis onde é percetível haver séries temporais, o método mais adequado será o de interpolação linear. Nesta abordagem, um valor ausente é calculado interpolando os valores das medições disponíveis anteriormente e seguintes de um dado indivíduo (Salgado et al., 2016).

Outros tipos de imputação simples são o *Hot deck* e *Cold deck*, que se caracterizam por substituir o valor em falta por valores observados nos indivíduos com o mesmo atributo, tendo em conta o próprio conjunto de dados ou uma fonte de dados externa, respetivamente (Salgado et al., 2016). Estes tipos de imputação normalmente são implementados em duas fases. Na primeira fase, os dados são repartidos em *clusters* e na segunda fase, cada variável com dados omissos é associada a um dos *clusters*. Os registos que se encontram completos em um *cluster*, são utilizados para preencher os dados ausentes. Para isso, pode-se recorrer ao cálculo da média ou à moda do atributo dentro de um *cluster*.

Na imputação baseada em modelo, é criado um modelo preditivo que irá estimar os valores que irão substituir os valores ausentes (Salgado et al., 2016). Neste caso, os dados da base de dados irão ser divididos em dois subconjuntos, um que não apresenta valores ausentes para a variável em avaliação, que funcionará como dados de treino, e o outro que só contem valores ausentes, que se pretendem estimar. Para tal, vários métodos de modelação podem ser utilizados, tais como, a regressão logística, redes neuronais, entre outros. Uma das principais desvantagens deste tipo de abordagem é que o modelo preditivo irá estimar geralmente valores com melhores comportamentos do que se se tratasse de valores reais.

Relativamente à regressão, uma das abordagens mais utilizadas é a regressão linear, em que, todas as variáveis de uma base de dados são utilizadas para criar um modelo de regressão linear e as observações da variável de interesse funcionam como *output* (Salgado et al., 2016). A principal vantagem desta técnica é que tem em consideração a relação entre as variáveis. Contudo, apresenta como desvantagem a questão de que subestima o ajuste do modelo e a correlação entre as variáveis, pois não tem em consideração a incerteza nos dados ausentes. Neste sentido, surgiu um novo método, a regressão linear estocástica, que já tem em consideração a incerteza nos dados omissos.

Uma outra poderosa técnica estatística, que nos últimos anos tem ganho muita popularidade, é a imputação múltipla**.** Este método, que é uma técnica Monte Carlo, surgiu como uma alternativa à imputação simples, também conhecida por imputação única, de forma a corrigir suas desvantagens. A sua principal vantagem é que ela tem em consideração a variabilidade entre as imputações nos resultados (Noghrehchi et al., 2020).

Dentro da imputação múltipla, existem duas abordagens principais, a abordagem de modelagem conjunta (*joint modeling approach*), que não é muito utilizada, e a abordagem de especificação totalmente condicional (*fully conditional specification*) (van Ginkel et al., 2019). Dentro desta última abordagem, é possível ainda fazer uma distinção entre a abordagem de regressão e a correspondência média preditiva.

De uma forma generalizada, pode-se dizer que, na imputação múltipla, cada valor ausente é substituído por um conjunto de valores, ou seja, é imputado m vezes, gerando assim, m diferentes bases de dados que serão, numa primeira fase, analisados separadamente e, posteriormente combinados num só resultado (Noghrehchi et al., 2020).

Relativamente ao número de imputações que serão necessárias para se conseguir produzir resultados que incorporam variação suficiente no processo de previsão, a fórmula de Rubin (1978) sugere que será necessário de 10 imputações (Curley et al., 2019). Por outro lado, existem autores que alegam que o número de imputações deve ser semelhante à percentagem de valores perdidos (Bodner, 2008; Royston & White, 2011). Graham et al. (2007) sugeriram que, para uma percentagem de 0,10,0,30,0,50,0,70 e 0,90 de *missing value*, o número ideal de imputações será de 20,20,40,100 e acima de 100, respetivamente.

Geralmente, a imputação múltipla funciona bem quando nos deparamos com dados ausentes que são do tipo MCAR ou MAR (Curley et al., 2019). Para além disso, esta técnica funciona melhor que outros métodos em um conjunto de dados alargado (Acock, 2005) e tem uma boa performance em estudos que apresentem mais de 25% de dados ausentes (Scheffer, 2002).

Na técnica *k-Nearest Neighbors* (k-NN), os valores ausentes de uma dada variável são preenchidos, tendo em consideração um determinado número de variáveis,k, que são semelhantes à variável de interesse (Acuña & Rodriguez, 2004). A semelhança entre duas variáveis é determinada através da minimização da função de distância, que pode ser Euclidiana, Pearson, entre outras (Acuña & Rodriguez, 2004). Depois de se encontrar os k vizinhos mais próximos, um valor de substituição deve ser estimado para substituir o valor ausente da variável. Uma das vantagens do k-NN, é que pode prever atributos qualitativos, através do valor mais frequente entre os k vizinhos mais próximos, e atributos quantitativos, através da média entre os k vizinhos mais próximos (Tsai & Chang, 2016). Para além disso, não precisa de criar um modelo preditivo para cada variável com dados omissos, como acontece para os métodos de imputação baseados em modelo. Porém, a escolha do valor k pode ser muito crítico, pois se o seu valor atribuído for muito elevado, estar-se-á a incluir atributos que são significativamente diferentes da observação destino, e se este for um valor muito baixo, poderá estar em falta atributos significativos

(Salgado et al., 2016). Segundo Batista & Monard (2003), o valor ideal para k, quando se está perante um grande conjunto de dados, será 10.

É de referir que, para além dos métodos anteriormente enumerados, existem muitos outros que lidam igualmente com dados ausentes e que para esta dissertação optou-se por se fazer somente uso da técnica *listwise deletion*.

3.3 Modelos de Regressão Logística Simples

Na análise de regressão linear simples, o valor médio de Y dado x, $E(Y|x)$ conhecido por valor esperado condicional - *conditional mean*, é muito importante pois tem como objetivo avaliar a dependência da variável dependente em relação a todas as variáveis independentes, determinado desta forma, o valor da variável dependente (David W. Hosmer & Lemeshow, 2000).

Na regressão linear simples, assume-se que o valor médio de Y é expresso na seguinte equação linear em x:

$$E(Y|x) = \beta_0 + \beta_1 x \tag{3.1}$$

Segundo essa expressão, é possível que $E(Y|x)$ tome valores de x entre $-\infty$ e $+\infty$. Assim, é evidente que, quando a variável resposta é binária, a expressão (3.1) não pode ser aplicada. Com dados dicotómicos, o *conditional mean* deve ser maior ou igual a zero e menor ou igual a um. Isto pode ser visto pelo gráfico presente na Figura 1. Segundo este gráfico, é possível observar-se que este valor médio atinge 0 e 1 de forma gradual. Para além disso, a mudança no $E(Y|x)$ por unidade mudada em x torna-se progressivamente menor à medida que o *conditional mean* se aproxima de 0 ou 1. Esta curva é conhecida por ser do tipo *S-shaped* e assemelha-se a uma projeção de distribuição cumulativa de uma variável aleatória.

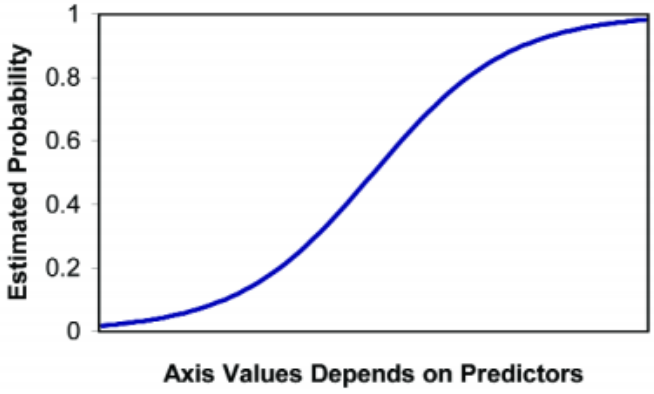

Figura 1 - Representação de uma curva do tipo *S-shaped* de uma função de regressão logística (Fonte: Bielecki & White, 2005).

Segundo Hosmer & Lemeshow (2000), algumas distribuições cumulativas têm vindo a ser usadas para proporcionar um modelo para $E(Y|x)$ no caso de Y ser dicotómico e muitas funções de distribuição têm a vindo ser propostas no uso de análise de variáveis de resposta dicotómicas, sendo uma das mais usadas a distribuição logística. Assim, de modo a simplificar a notação, consideremos $\pi(x) = \mathrm{E(Y|}x)$ para representar o *conditional mean* de Y dado x quando a distribuição logística é utilizada. A forma específica de modelo de regressão logística utilizada é:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \tag{3.2}$$

Sendo que a transformação de $\pi(x)$, conhecida por *logit transformation*, é definido, em termos de $\pi(x)$, como:

$$g(x) = ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x \tag{3.3}$$

Uma vantagem desta transformação é que $g(x)$ apresenta muitas propriedades desejáveis de um modelo de regressão linear. O *logit*, $g(x)$, é linear nos seus parâmetros, que podem ser contínuos, e pode variar entre $-\infty$ e $+\infty$.

Uma outra questão que diferencia os modelos de regressão linear dos de regressão logística está relacionada com a distribuição condicional de uma variável resposta. Relativamente à regressão linear, assume-se que a observação de uma variável resposta se expressa pela seguinte equação:

$$y = E(Y|x) + \varepsilon \tag{3.4}$$

Sendo, ε o erro que expressa o desvio da observação do *conditional mean.* Para além disso, ε segue a distribuição normal com média 0 e alguma variância que é frequente sobre níveis de variáveis independentes. Segue-se que, a distribuição condicional de uma variável resposta dado x**,** irá ser normal com média $E(Y|x)$ e com variância constante. Porém, este não é o caso quando se fala de uma variável resposta dicotómica. Para este caso, deve-se expressar o valor da variável resposta dado x como $y = \pi(x) + \varepsilon$, sendo que ε poderá assumir dois valores possíveis:

$$\varepsilon = \begin{cases} -\pi(x) & com\ probabilidade\ de\ 1-\pi(x) \quad se\ Y=0 \\ 1-\pi(x) & com\ probabilidade\ de\ \pi(x) \quad se\ Y=1 \end{cases} \tag{3.5}$$

Deste modo, ε apresenta uma distribuição cujo valor médio é nulo e a variância é igual a $\pi(x)[1-\pi(x)]$. A distribuição condicional da variável resposta segue uma distribuição binomial com probabilidade de sucesso $\pi(x)$.

É de destacar que, se quisermos representar a probabilidade de um evento ocorrer por $\pi(x)$, então o *odds* respetivo é dado pela seguinte expressão:

$$Odds = \frac{\pi(x)}{1-\pi(x)} \tag{3.6}$$

Entende-se por *odds* de um dado evento como sendo o quociente entre a probabilidade da ocorrência do evento e a probabilidade da sua não ocorrência (Stare & Maucort-Boulch, 2016).

3.4 Modelos de Regressão Logística Multivariada

Considerando uma coleção de *p* variáveis independentes indicadas por um vetor $\mathbf{x}' = (x_1, x_2, \dots, x_p)$ e que, $\mathrm{E}(\mathrm{Y}|\mathbf{x}) = \pi(\mathbf{x})$, o modelo da regressão logística multivariada é dado por:

$$\pi(\mathbf{x}) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}} \tag{3.7}$$

onde $\beta_0, \beta_1, \beta_2 \ \dots \ \beta_p$ são coeficientes do modelo *logit* (David W. Hosmer & Lemeshow, 2000).

Consequentemente, o *logit* do modelo logístico multivariado, ou seja, a sua linearização é dada pela seguinte equação:

$$g(\mathbf{x}) = \ln\left[\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p \tag{3.8}$$

A importância desta transformação *logit* é que $g(\mathbf{x})$ tem muita das propriedades desejadas de um modelo de regressão linear. O *logit*, $g(\mathbf{x})$, é linear nos seus parâmetros, pode ser contínua, e pode variar de $-\infty$ a $+\infty$, dependendo do valor de $\mathbf{x}_i$.

Se alguma variável independente for uma variável de escala nominal ou discreta, como por exemplo, a raça, sexo, entre outros, será inapropriado incluí-la no modelo como se fosse uma variável de escala intervalar. Os números utilizados para representar os vários níveis dessa variável de escala nominal são meros identificadores, não apresentando um significado numérico. Neste caso a melhor opção será utilizar um conjunto de variáveis *dummy*.

3.5 Estimação dos Coeficientes do Modelo

A estimação dos coeficientes de um modelo de regressão logística é possível através da aplicação do método da máxima verosimilhança (David W. Hosmer & Lemeshow, 2000). Assim, assumindo-se que há um conjunto de n observações independentes $(\mathbf{x}_i, y_i), i = 1,2, \dots, n$, a função de verosimilhança define-se pela seguinte expressão:

$$L(\beta) = \prod_{i=1}^{n} \left(\frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}} \right)^{y_i} \left[1 - \left(\frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}} \right) \right]^{1-y_i} \quad (3.9)$$

Para que seja possível obter-se os estimadores de máxima verosimilhança, será necessário resolver um sistema de $p + 1$ equações, constituídas por derivadas parciais da log verosimilhança relativamente a cada um dos parâmetros. A função de verosimilhança resultante pode ser expressa da seguinte maneira:

$$\sum_{i=1}^{n} [y_i - \pi(\mathbf{x}_i)] = 0 \quad (3.10)$$

e

$$\sum_{i=1}^{n} x_{ij} [y_i - \pi(\mathbf{x}_i)] = 0$$

para $j = 1,2, \dots, p$ (David W. Hosmer & Lemeshow, 2000).

Contudo, para se obter a solução de equações de verosimilhança será necessário recorrer a um algoritmo adequado.

3.6 Teste à Significância dos Coeficientes

Uma vez estimados os coeficientes do modelo, é necessário avaliar a significância dos mesmos, de modo a identificar quais as variáveis que apresentam uma maior influência no modelo estimado. De um modo geral, esta avaliação requer a realização de um teste de hipóteses, que visa avaliar se as variáveis independentes do modelo estão "significativamente" relacionadas com a variável dependente (David W. Hosmer & Lemeshow, 2000).

Segundo Hosmer & Lemeshow (2000), uma abordagem comummente utilizada para testar a significância do coeficiente do modelo de regressão logística passa pelo seguinte princípio: "Comparar valores observados de uma variável resposta para predizer valores obtidos de modelos com e sem a variável em questão". Na regressão logística a comparação entre os valores observados e previstos para a variável resposta é baseada na função de log-verosimilhança, que se dá pela seguinte expressão:

$$D = -2\log \left[\frac{verosimilhança\ do\ modelo\ ajustado}{verosimilhança\ do\ modelo\ saturado} \right] \quad (3.11)$$

em que o modelo saturado diz respeito a um modelo que contém todas as variáveis e o modelo ajustado diz respeito ao modelo que apresenta apenas as variáveis desejadas para o estudo. O quociente entre as verosimilhanças, presente na expressão (3.11), é designada por razão de verosimilhanças e a

aplicabilidade de $-2\log$ serve para que a quantidade D siga uma distribuição conhecida, de modo a que seja possível aplicar-lhes testes de hipóteses. Este teste é chamado de *likelihood ratio test* (LR).

Para além disso, a equação (3.12) é conhecida por *deviance,* onde $\hat{\pi}_i = \hat{\pi}(\mathbf{x_i})$ (David W. Hosmer & Lemeshow, 2000).

$$D = -2\sum_{i=1}^{n}[y_i \ln\left(\frac{\hat{\pi}_i}{y_i}\right) + (1-y_i)\ln\left(\frac{1-\hat{\pi}_i}{1-y_i}\right)] \tag{3.12}$$

Como na regressão logística a variável resposta é binária, o valor da função de verosimilhança do modelo saturado é 1. Deste modo, a expressão do *deviance* resume-se a:

$$D = -2\log\left(verosimilhança\ do\ modelo\ ajustado\right) \tag{3.13}$$

De modo a avaliar se uma variável é ou não significativa para o modelo, compara-se a *deviance* do modelo com e sem a variável, através da seguinte expressão:

$$G = D(modelo\ sem\ a\ variável) - D(modelo\ com\ a\ variável) \tag{3.14}$$

Como a verosimilhança do modelo saturado é comum a ambos os modelos, G pode ser reescrita da seguinte forma:

$$G = -2\log\left[\frac{(modelo\ ajustado\ sem\ a\ variável)}{(modelo\ ajustado\ com\ a\ variável)}\right] \tag{3.15}$$

Tendo como base a estatística G é possível realizar-se um teste para avaliar se uma certa variável é ou não relevante para o modelo. Deste modo, podemos colocar as seguintes hipóteses:

$$H_0: \beta_1 = \cdots = \beta_p = 0 \tag{3.16}$$
$$H_1: \beta_1 = \cdots = \beta_p \neq 0$$

Na regressão logística multivariada, a significância da estatística do teste da razão de verosimilhanças é testada utilizando a distribuição qui-quadrado cujo número de graus de liberdade é igual ao número de variáveis independentes no modelo, não se incluindo a constante. Para $P[\chi_k^2 > G] < 0{,}05$ rejeita-se H_0 concluindo-se assim que, pelo menos uma das variáveis é significativa, e, por conseguinte, será necessário testar cada um dos coeficientes. Para isso, recorre-se ao teste de *Wald*, de modo a averiguar se cada um dos coeficientes são significativamente diferentes de zero.

Na regressão logística multivariada, o teste de *Wald* é obtido pela seguinte expressão:

$$W = \hat{\beta}'[\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\beta}})]^{-1}\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}'(\mathbf{X}'\mathbf{V}\mathbf{X})\widehat{\boldsymbol{\beta}} \tag{3.17}$$

em que, sob a hipótese de cada um dos $p+1$ coeficientes ser igual a zero, segue uma distribuição χ_{p+1}^2. Para um nível de significância α, a hipótese nula será rejeitada caso $W > \chi_{1-\alpha}^2(p+1)$, ou seja, o parâmetro associado á variável não é zero e consequentemente a variável deve ser incluída no modelo. Porém, o teste de Wald em algumas situações poderá não ser eficaz podendo por vezes rejeitar variáveis

que na verdade são estatisticamente significativas. Nestes casos, é aconselhável testar a significância dos seus coeficientes de acordo com o teste da razão de verosimilhança, que se apresenta como sendo uma melhor estatística quando, por exemplo, a amostra é mais pequena (Hauck & Donner, 1977).

Hosmer & Lemeshow (2000) também referenciam o teste de *Score* como um possível teste para avaliar a significância dos coeficientes de um modelo. Este teste baseia-se na teoria da distribuição das derivadas do logaritmo da máxima verosimilhança e tem como principal vantagem, em comparação ao teste anterior, a questão de fazer uso de um menor esforço computacional no seu cálculo.

3.7 Seleção de Variáveis

De forma a determinar o subconjunto de varáveis independentes a serem incluídas no modelo de regressão final, deve-se recorrer à análise dos dados. Este processo é conhecido por problema de seleção de variáveis e tem como objetivo duas questões contraditórias. Segundo NCSS, empresa que desenvolveu o pacote estatístico NCSS, é requerido primeiramente que o modelo seja o mais completo e o mais realista possível, apresentando todas as variáveis independentes que estejam relacionadas com a variável dependente. Segundo, é solicitado que o modelo inclua o menor número possível de variáveis, uma vez que, cada variável independente irrelevante diminui a precisão dos coeficientes estimados e dos valores previstos (NCSS).

Apesar de existirem diversas estratégias de seleção de variáveis para um modelo de regressão, nesta secção será abordado somente o método de seleção *stepwise,* uma vez que, será o método adotado neste trabalho.

Na regressão *stepwise*, o procedimento de seleção é automaticamente executado por pacotes estatísticos. Existem vários critérios de seleção de variáveis, como R-quadrado ajustado, *Akaike information criterion* (AIC), *Bayesian information criterion* (BIC), Cp de Mallows, PRESS, entre outros (Zhang, 2016).

Este método de seleção automática de variáveis pode ser utilizado em três direções diferentes:

- *Forward*: O método *Forward* começa com um modelo nulo que não apresenta qualquer variável, somente a constante, sendo que, em cada interação, irá adicionar a variável que seja mais significativo ao modelo atual e que apresentar uma melhoria significativa no ajuste do modelo (Austin & Tu, 2004). Este procedimento vai-se repetindo até que nenhuma das variáveis restantes proporcionar uma melhoria significativa no ajuste ou até que uma dada regra de paragem seja satisfeita. Segundo Lee e Koval (1997), um dos critérios de paragem que possa

ser utilizada neste método passa pela utilização do teste χ^2 baseado num nível fixo α ($\chi^2_{(\alpha)}$). Este mesmo autor indica que, para fins de previsão, o melhor valor para α varia entre 0,05 e 0,40 para casos de regressão logística binária multivariada. Porém Bendel e Afifi (1977) recomendam que, para a regra de paragem deste modelo, o critério de entrada deverá apresentar um valor entre 0,15 e 0,25 (Zellner et al., 2004).

- *Backward*: Neste caso, este método inicia com o modelo que apresenta todas as variáveis e, em cada iteração, irá eliminar a variável menos relevante do modelo de regressão até atingir o modelo reduzido que melhor descreve os dados (Austin & Tu, 2004). As variáveis são eliminadas sequencialmente do modelo até que uma regra de paragem pré-especificada seja satisfeita. Uma das regras de paragem mais utilizadas é quando todas as variáveis que permanecem no modelo são significativas em um nível de significância pré-especificado, sendo que, começa por eliminar as variáveis que são menos significativas do modelo inicial.
- *Both*: Também conhecido por *stepwise*, é uma técnica de seleção automática de modelos que resulta da combinação das outras duas anteriores. Este é considerado uma modificação do procedimento *forward* pois, as variáveis são introduzidas uma a uma, sendo que, em cada passo é realizado uma avaliação de forma a garantir que as variáveis continuam a ser relevantes após a introdução de uma nova variável no modelo (NCSS). Se após a introdução de uma nova variável o valor das estatísticas de teste de uma dada variável passar a ser inferior a um certo valor estabelecido, então essa variável será removida do modelo, considerando-se a próxima e repete-se todo esse processo.

Por vezes, este tipo de procedimento de seleção de variáveis torna-se instável. Um exemplo é quando existe uma adição ou uma exclusão de um pequeno número de dados ou quando as covariáveis estão correlacionadas (Steyerberg et al., 2000). Para além disso, a seleção de variáveis apresenta um poder limitado para selecionar covariáveis importantes em prognósticos em pequenos conjuntos de dados, o que conduzirá a uma perda na capacidade preditiva do modelo.

3.8 Interpretação do Modelo de Regressão Logística Ajustado

Após se realizar uma avaliação de ajuste do modelo relativamente aos dados, o passo seguinte passa por realizar uma interpretação dos valores dos seus coeficientes estimados.

Segundo Hosmer & Lemeshow (2000), os coeficientes estimados para as variáveis independentes representam a taxa de variação de uma função da variável dependente por unidade de mudança na

variável independente. Assim, esta interpretação envolve duas questões: determinar a relação funcional entre a variável dependente e a variável independente, e definir de forma apropriada, a unidade de mudança para a variável independente.

Para a primeira questão, deve-se determinar que tipo de função da variável dependente produz uma função linear da variável independente, ou seja, que função *link*. Assim sendo, como se trata de um modelo de regressão logística a função *link* será a transformação *logit* representada na equação (3.18).

$$g(x) = ln\left\{\frac{\pi(x)}{[1-\pi(x)]}\right\} = \beta_0 + \beta_1 x \tag{3.18}$$

Neste caso, os incrementos dos coeficientes representam a mudança no *logit*, que corresponde a uma mudança de uma unidade na variável independente. Uma interpretação bem feita de um coeficiente no modelo de regressão logística, depende se é possível colocar significado na diferença entre dois *logits*. Um dos parâmetros mais utilizados para a interpretação dos valores de coeficientes de um modelo de regressão logística é o *odds ratio*, OR.

Porém, as interpretações dos valores dos coeficientes do modelo são feitas de forma diferente conforme de que tipo de variável independente se trate.

3.8.1 Variável independente nominal e dicotómica

O valor estimado de log dos *odds ratio* de qualquer variável independente com dois níveis diferentes, diz-se $x = a$ versus $x = b$, é a diferença entre os *logits* estimados nesses dois valores,

$$\begin{aligned} ln\left[\widehat{OR}(a,b)\right] = \hat{g}(x=a) - \hat{g}(x=b) &= (\hat{\beta}_0 + \hat{\beta}_1 \text{ x } a) - (\hat{\beta}_0 + \hat{\beta}_1 \text{ x } b) \\ &= \hat{\beta}_1 \text{ x } (a-b) \end{aligned} \tag{3.19}$$

O valor estimado de *odds ratio* é obtido pela exponenciação da diferença de *logit*,

$$\widehat{OR}(a,b) = \exp\left[\hat{\beta}_1 \text{ x } (a-b)\right] \tag{3.20}$$

É de referenciar que, nas equações (3.19) e (3.20) a notação de $\widehat{OR}(a,b)$ corresponde aos *odds ratio*.

$$\widehat{OR}(a,b) = \frac{\dfrac{\hat{\pi}(x=a)}{[1-\hat{\pi}(x=a)]}}{\dfrac{\hat{\pi}(x=b)}{[1-\hat{\pi}(x=b)]}} \tag{3.21}$$

3.8.2 Variável independente nominal e policotómica

Para este caso, onde a variável independente tem mais do que 2 níveis, para que seja possível fazer a sua interpretação esta variável terá de ser recodificado para uma variável *dummy*, sendo que poderá assumir valores 0 ou 1. Caso a variável nominal em estudo apresentar m categorias, então ter-se-á de

criar $m - 1$ variáveis *dummy* indexadas a essa categoria. O cálculo dos *odds ratio* é feito da mesma forma como se tratasse de uma variável dicotómica.

3.8.3 Variável independente contínua

Quando um modelo de regressão logística contém uma variável independente contínua, a interpretação dos coeficientes estimados depende de como essa variável entrou para o modelo e da unidade da mesma. Assumindo de que o *logit* é linear para esse tipo de variável,x, a equação do *logit* é $g(x) = \beta_0 + \beta_1 x$. Além disso, o incremento do coeficiente, β_1, fornece a chance no log *odds* para um acréscimo de 1 unidade em x, ou seja, $\beta_1 = g(x+1) - g(x)$ para qualquer valor de x. Por isso, para se fornecer uma interpretação útil para uma covariável de escala contínua, será necessário construir um método e estimar um intervalo para uma mudança arbitrária de "c" unidades na covariável.

O log *odds ratio* para a mudança de "c" unidades em x é obtida através da diferença de *logit* $g(x+c) - g(x) = c\beta_1$ e o *odds ratio* associado é obtido pela exponenciação da diferença do *logit,* $OR(c) = OR(x+c, x) = \exp(c\beta_1)$.

4. REGRESSÃO LOGÍSTICA – CLASSIFICAÇÃO E PREVISÃO

4.1 Classificadores

Atualmente, em muitas áreas da medicina são recolhidas um número cada vez maior de dados, sendo isto um desafio para o campo da estatística, na extração de conhecimentos úteis destes dados. Com o desenvolvimento das novas tecnologias, estimulou-se um maior interesse na criação de modelos preditivos e de classificadores. Muitos desses modelos preditivos foram desenvolvidos de forma a fornecer novas formas de diagnóstico de doenças, prever prognósticos, selecionar o melhor tratamento para os pacientes, entre outros (Pepe, 2005).

Toda a previsão de um dado evento exige um processo de classificação. Na estatística, os problemas de classificação são os que, por exemplo, atribuem um dado indivíduo a um determinado grupo, de acordo com o conjunto de características que apresenta. Assim, qualquer função matemática que implementa tal procedimento é chamada de classificador. São exemplos de classificadores os testes de diagnóstico, os biomarcadores, testes de *screening* e testes médicos (Alonzo & Pepe, 2007).

Um classificador pode apresentar resultados binários, ordinais ou contínuos (Alonzo & Pepe, 2007). Como exemplos de testes que apresentam resultados binários pode-se incluir os testes de gravidez e os de cultura de bactérias para detetar doenças infeciosas. Já, por exemplo, as interpretações de imagens de um radiologista para quantificar a incerteza de existência de cancro, geralmente são baseadas numa escala ordinal do tipo: 1= normal, 2=benigno, 3=provavelmente benigno, 4=suspeito de cancro, 5=altamente suspeito de cancro. Testes que incluem concentrações de marcadores tumorais, como o PSA para deteção de cancro da próstata, são considerados como classificadores contínuos.

Relativamente ao tipo de método de classificação que se possa implementar num classificador em construção, é possível encontrar-se na literatura vários autores que abordam os vários métodos de classificação existentes, desde os mais clássicos, como o Análise Discriminante Linear, Análise Discriminante Quadrática e Regressão Logística até aos mais recentes métodos de *Machine Learning* e *Data Mining*, como *Support Vector Machines, Neural Networks, Random Forests* e *Boosting* (Y. Liu et al., 2011; Maroco et al., 2011).

No que diz respeito à sua classificação, Liu et al. (2011) indicam que os vários métodos de classificação se dividem em dois principais grupos, os *soft* que estimam explicitamente as probabilidades condicionais da classe, prevendo a classe com base na maior probabilidade estimada, e as *hard*, que ignoram o requisito de estimativa de probabilidade de classe estimando diretamente o limite de classificação. Por

outro lado, outros autores declaram que, em termos gerais, os vários métodos de classificação se podem dividir em métodos probabilísticos, como os classificadores Bayesianos, em métodos de regressão, como a Regressão Logística e em métodos geométricos, como os *Support Vector Machines* (W. Liu et al., 2019).

Modelos como a regressão logística têm sido os mais utilizados em problemas de classificação médica (Dreiseitl & Ohno-Machado, 2002), principalmente os que apresentam a variável resposta como dicotómica, pois na medicina é mais frequente encontrar-se problemas que envolvem a classificação dos dados em dois grupos, como por exemplo, pacientes doentes ou não doentes, pacientes que reagem ao tratamento ou não reagem ao tratamento, recém-nascidos normais ou anormais, entre outros. Esse tipo de modelos, em que existem apenas duas categorias, são conhecidos como classificadores binários.

É de referir que, o próprio modelo de regressão logística por si só, simplesmente modela a probabilidade de saída em termos de entrada, não executando nenhuma classificação estatística, ou seja, nessa fase ainda não se pode considerar como um classificador. Para que se possa usar a regressão logística para criar um classificador binário, será necessário escolher um valor de corte e classificar os valores de *output* com probabilidade maior que o ponto de corte de uma classe ou abaixo do ponto de corte (W. Liu et al., 2019). Para além disso, um classificador é desenvolvido no conjunto de dados de treino, ou seja, num conjunto de amostras cujas classes sejam previamente conhecidas, e posteriormente ao seu treinamento, é exposto num conjunto de amostras cujas classes são desconhecidas, também conhecido por dados de teste, de modo a que ele comece a prever as classes dessas amostras (Dobbin & Simon, 2011).

Para que um futuro classificador apresente resultados de boa qualidade é necessário ter em atenção alguns fatores durante o desenvolvimento do mesmo, como por exemplo, o conjunto de dados a utilizar devem ser de boa qualidade, ter em atenção se a escolha dos parâmetros ajustáveis para o modelo foram os mais adequados e avaliar os critérios de avaliação utilizados para relatar os resultados do processo de modelação (Dreiseitl & Ohno-Machado, 2002). Para além disso, para garantir que os modelos sejam validados, é de preferência que se faça uso de um conjunto de dados externos de modo a verificar a plausibilidade dos modelos.

4.2 Desempenho de classificadores

Antes que um classificador seja adotado nos cuidados de saúde, será necessário determinar a sua precisão de classificação, de forma a evitar possíveis problemas na medicina. Um classificador mal

calibrado, no caso de sobrevida, poderá indicar que um indivíduo com possibilidade de vir a morrer ser classificado como sendo um indivíduo que irá sobreviver, e assim não receberá o tratamento vital, ou então, indicar que um indivíduo que virá a sobreviver apresente um resultado positivo para o óbito, sendo assim sujeito a procedimentos médicos desnecessários. Assim, de forma a evitar toda esta situação, deve-se realizar uma avaliação rigorosa do desempenho dos classificadores.

Na área médica, a sensibilidade, especificidade, valores preditivos e área sob a curva ROC são frequentemente utilizados para descrever o desempenho da classificação de um classificador e que devem ser considerados quando se pretende comparar classificadores (Dreiseitl & Ohno-Machado, 2002; Maroco et al., 2011).

4.2.1 Sensibilidade e especificidade

Para classificadores binários considera-se a seguinte analogia, onde a variável F corresponde a um estado de admissão de um recém-nascido (Pepe, 2003):

$$F = \begin{cases} 0 & Sobrevive \\ 1 & Falece \end{cases} \tag{4.1}$$

E a varável Y representa o resultado apresentado pelo classificador, sendo que, por convenção, valores mais altos de Y são indicativos de uma maior propensão para o recém-nascido vir a falecer.

$$Y = \begin{cases} 0 & Negativo\ para\ sobrevive \\ 1 & Positivo\ para\ \ falece \end{cases} \tag{4.2}$$

Assim, com base no resultado do classificador, os profissionais de saúde podem classificar um recém-nascido como sobrevivente ou como falecido.

Numa abordagem estatística, a variável Y sendo binária, o desempenho da classificação é geralmente resumido pela fração verdadeira positiva (FVP) e pela fração falsa positiva (FFP) definido por (Pepe, 2003):

$$TPF = P(Y = 1|F = 1) \quad e \quad FPF = P(Y = 1|F = 0) \tag{4.3}$$

Tendo em conta a analogia exposta, um verdadeiro positivo ocorre quando um classificador com resultado positivo classifica como falecido um recém-nascido que realmente irá falecer e um falso positivo ocorre quando o resultado do classificador dá positivo para falecido quando na realidade o recém-nascido irá sobreviver.

Assim, um bom classificador apresenta um valor alto de FVP (sensibilidade) e um valor baixo de FFP (1-especificidade), sendo considerado ideal quando o valor de FVP for igual a um e o valor de FFP igual a zero (Alonzo & Pepe, 2007). Em contrapartida, quando o valor de FVP e FFP forem iguais, isto significa que o classificador não é conclusivo sobre a previsão do estado do recém-nascido.

4.2.2 Curva ROC

A curva ROC é uma representação gráfica de fração de verdadeiros positivos, no eixo nas ordenadas, e de fração de falsos positivos, no eixo das abcissas, para um limiar de classificação, sendo que pode ser interpretada como sendo uma curva que resume as informações das funções de distribuição cumulativas das pontuações das duas classes consideradas (M. F. Mourão et al., 2015).

Tomamos como ponto de partida a existência de duas populações de recém-nascidos, uma população positiva/falece, que denominamos por F e uma população negativa/sobrevive, que denominamos por S, assim como, uma regra de classificação que permite atribuir cada indivíduo a cada um desses dois grupos. Assumimos que a regra de classificação se encontre em uma função contínua $t(x)$ de um vetor aleatório X de variáveis medidas em cada recém-nascido, do qual foi arranjada de forma a que, altos valores na função correspondessem à população F, enquanto que, valores mais baixos correspondessem à população S (M. F. Mourão et al., 2015). Logo, se x for o valor observado de X para um recém-nascido em particular e $t(x)$ a função de pontuação para esse recém-nascido, então esse bebé é atribuído à população F ou S, caso $t(x)$ ultrapasse ou não o valor limiar ou de corte c, respetivamente.

Para que seja possível fazer uma avaliação eficaz desse classificador, é necessário calcular a probabilidade de fazer uma atribuição incorreta de um recém-nascido nos grupos de população (M. F. Mourão et al., 2015). Para isso, pode-se definir quatro probabilidades e as suas taxas associadas para o classificador:

- A probabilidade de um recém-nascido de F ser classificado corretamente, ou seja, a fração verdadeira positiva, $TPF = P(t > c|F)$;
- A probabilidade de um recém-nascido de S ser classificado incorretamente, ou seja, a fração falsa positiva, $FPF = P(t > c|S)$;
- A probabilidade de um recém-nascido de S ser classificado corretamente, ou seja, a fração verdadeira negativa, $TNF = P(t \leq c|S)$;
- A probabilidade de um recém-nascido de F ser classificado incorretamente, ou seja, a fração falsa negativa, $FNF = P(t \leq c|F)$.

Dadas as densidades de probabilidades $P(t|F)$, $P(t|S)$ e o valor de c, valores numéricos entre zero e um podem ser obtidos pelas quatro frações representadas acima, o que fornece uma descrição completa do desempenho do classificador (M. F. Mourão et al., 2015).

Apesar que, para um bom desempenho se exige altas frações verdadeiras e baixas frações falsas, isso só acontece para uma escolha particular do limite c, sendo que, a melhor escolha para o limite não é conhecida antecipadamente, mas sim durante o processo de construção do classificador. Variando o valor de c e avaliando as quatro probabilidades mencionadas anteriormente, irá fornecer informações acerca do desempenho dos classificadores.

O objetivo da curva ROC passa por fornecer uma avaliação do classificador em toda a faixa de valores limite potenciais. Evidentemente, o valor de um classificador pode ser julgado pela extensão em que as duas distribuições de suas pontuações $P(t|F)$ e $P(t|S)$ diferem sendo que, quanto mais elas se distanciam, menos sobreposição haverá entre estas e consequentemente menos alocações incorretas existiram (M. F. Mourão et al., 2015). Por outro lado, quanto mais as duas distribuições se assemelham, maior será a sobreposição entre elas e, consequentemente maior serão as alocações incorretas. Analogamente, quando considerarmos a curva ROC, quanto mais próximo a curva estiver do canto superior esquerdo do gráfico, mais se aproxima de uma situação de completa separação entre as duas populações e, consequentemente melhor será o desempenho do classificador.

Apesar da curva ROC apresentar uma descrição abrangente do desempenho de um classificador em todos os seus valores limite possível, a sua análise poderá ser mais complicada quando se pretender comparar vários classificadores diferentes. Uma das melhores alternativas passa pela análise da área sob a curva ROC (AUC) de cada classificador. Quando nos depararmos com um caso de perfeita separação das distribuições F e S, o valor de AUC será 1.

4.3 Medidas de Diagnóstico do Modelo de Previsão

Normalmente, quando nos deparamos com uma tarefa de modelação, somente criar o modelo preditivo e confrontá-lo com um conjunto de novos dados não é suficiente. Para que o modelo desenvolvido possa ser utilizado, será necessário avaliar o seu desempenho preditivo. Na literatura é possível encontrar-se diversas métricas de avaliação de desempenho de modelos de regressão logística, tais como, área abaixo da curva ROC, AIC, BIC, pseudo R^2, entre outras.

4.3.1 Curva ROC

Uma das formas para avaliar o diagnóstico dos modelos estatísticos obtidos, neste caso de classificadores binários, passa pela utilização da curva ROC (*Receiver Operating Characteristic*).

A análise ROC, baseada em princípios fundamentais da teoria da decisão estatística e da teoria da deteção de sinal, foi criada no início dos anos 50 para avaliar a deteção de sinal em radar e na psicologia sensorial (Charles E. Metz, 2008). No entanto, a primeira utilização prática da curva ROC apareceu durante a segunda guerra mundial e foi desenvolvida pela força aérea real britânica para analisar sinais de radar e estudar a relação entre sinal e ruído, de modo a diferenciar com precisão os sinais que correspondessem, por exemplo, a um avião inimigo, ou então a um ruído provocado, por exemplo, por pássaros (Carter et al., 2016).

Desde então, as curvas ROC têm sido totalmente associadas à teoria da deteção de sinal, sendo o conceito posteriormente popularizado em diversos campos como imagens médicas (C.E. Metz, 1986), *machine learning* (Spackman, 1989), informática médica (Lasko et al., 2005), psicofísica, controlo de qualidade industrial, finanças, sociologia, entre outros (Alemayehu & Zou, 2012).

A sua aplicação na medicina foi descrita originalmente por Lusted em 1971, no qual colocou em prática na análise de imagens radiográficas. Hoje em dia, esta metodologia foi adaptada a outras áreas clínicas que dependem de testes de triagem e diagnóstico, nomeadamente de testes de laboratório, epidemiologia, radiologia e bioinformática (Zou et al., 2007). Por exemplo, na cardiologia, o teste diagnóstico desempenha um papel fundamental na prática clínica, nomeadamente na utilização de marcadores séricos para rastrear necrose miocárdica e em exames de imagem cardíacas que são utilizados para diagnosticar várias complicações cardiovasculares (Alemayehu & Zou, 2012). Além disso, as curvas ROC também têm sido utilizadas na avaliação do desempenho de modelos preditivos para a identificação de subgrupos de pacientes com riscos diferenciados para resultados específicos, como por exemplo, a mortalidade.

A curva ROC é uma abordagem gráfica no qual os seus eixos são representados pela fração de verdadeiros positivos de um classificador (sensibilidade), no eixo das ordenadas, e pela fração de falsos positivos (1-especificidade),no eixo das abcissas, em que cada ponto é gerado por um valor limite diferente, ou seja, ponto de corte (Goksuluk et al., 2016). Entende-se por sensibilidade como sendo a proporção de positivos que foram identificados corretamente e por especificidade como sendo a proporção de negativos que foram identificados corretamente (Carter et al., 2016).

Uma das principais tarefas passa por determinar o valor ideal do ponto de corte que corresponda aos valores razoáveis de FVP (fração de verdadeiros positivos) e FFP (fração de falsos positivos) (Goksuluk et al., 2016). Assim, recorre-se à curva ROC de modo a encontrar o ponto de corte ideal localizado na curva, que é o ponto mais próximo do canto superior esquerdo. Contudo, esse ponto de corte ideal nem sempre corresponde àquele em que se maximiza a sensibilidade e a especificidade.

Atualmente, existem vários métodos capazes de determinar o ponto de corte ideal, sendo que a maioria se baseia nas medidas de sensibilidade e especificidade. No entanto, existem outros métodos que têm como base no custo-benefício, valores preditivos, razões de probabilidade de diagnóstico e índice de *Youden* (Goksuluk et al., 2016).

A implementação da análise da curva ROC pode ser útil, nomeadamente, se se pretender avaliar o desempenho geral de um classificador utilizando para isso várias medidas de desempenho, comparar o desempenho de classificadores e determinar o ponto de corte ideal para um certo classificador (Goksuluk et al., 2016).

Na Figura 2 encontra-se representado um gráfico com a curva ROC, em que cada ponto da curva representa um possível ponto de corte e a sua respetiva sensibilidade e especificidade.

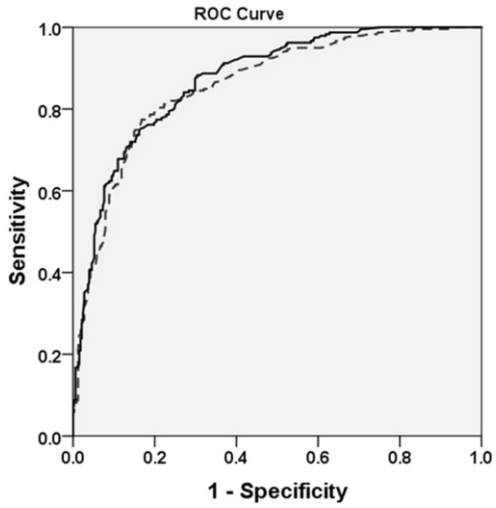

Figura 2 - Representação gráfica da curva ROC (Fonte: Gulliver & Yoder, 2018).

A representação gráfica das curvas ROC é considerada uma ferramenta útil para visualizar, organizar e selecionar classificadores de acordo com o seu desempenho, sendo que, curvas que se aproximam mais do canto superior esquerdo, correspondem a classificadores com capacidades discriminatórias maiores. Segundo vários autores (Alemayehu & Zou, 2012; Zou et al., 2007), a área abaixo da curva (AUC) é um dos índices mais utilizados para medir a *accuracy* de um classificador. Quando a área sob a curva ROC de um dado classificador apresentar valor de 1, isto indica que este tem uma capacidade de discriminação perfeita. Contudo, quando o valor de AUC for de 0,5, a sua capacidade de discriminação é nula (Fan et al., 2006). Apesar de existirem várias escalas de interpretação do valor de AUC, segundo Fan et al. (2006), as curvas ROC com um valor de AUC igual ou inferior a 0,75, não são clinicamente úteis, e uma AUC de 0,97 apresenta um valor clínico muito alto.

Segundo Braga (2000), os métodos estatísticos mais utilizados para o cálculo da AUC de um classificador são a estatística de Wilcoxon-Mann-Whitney, regra do trapézio, área binormal e estimação de máxima verosimilhança.

4.3.2 *Akaike Information Criterion*

O *Akaike information criterion* (AIC), proposto por Akaike em 1973, é um critério amplamente utilizado na seleção do melhor modelo entre um conjunto de modelos candidatos (Yanagihara et al., 2012). Cada modelo é caracteriza por apresentar um valor de AIC, pelo que, a seleção do melhor modelo passará por aquele que apresentar o menor valor de AIC. Este critério é frequentemente utilizado na sua forma original devido à sua simplicidade, contudo já foram propostas várias variantes deste mesmo critério ao longo dos últimos anos (Commenges et al., 2008). Segundo a abordagem teórica da informação defendida por Akaike (1973), a discrepância de informação de Kullback-Leibler (1951) é considerada como um critério básico para avaliar e medir a discrepância entre o modelo verdadeiro e o modelo candidato (Konishi & Kitagawa, 2008). Assim, o AIC derivou como uma estimativa aproximada assintótica da discrepância de informações de Kullback-Leibler e fornece uma ferramenta útil para avaliar modelos estimados pelo método da máxima verossimilhança. O critério AIC é definido pela seguinte equação:

$$AIC = -2(\ln(L)) + 2k \tag{4.4}$$

onde k representa o número de parâmetros independentes e L representa o valor obtido da função de máxima verosimilhança do modelo.

Porém, o AIC poderá não apresentar um bom desempenho em algumas situações, como por exemplo, um modelo com muitos parâmetros tende a ser escolhido como o melhor modelo quando o tamanho da amostra for pequeno (Yanagihara et al., 2012).

4.3.3 *Bayesian Information Criterion*

O *Bayesian information criterion* (BIC), proposto por Schwarz em 1978, também é um critério igualmente utilizado na seleção do melhor modelo, sendo que, aquele que apresentar o menor valor de BIC será o escolhido (Konishi & Kitagawa, 2008). O BIC é um critério baseado na probabilidade bayesiana e representa-se pela seguinte expressão:

$$BIC = -2lnL\left(\underset{\sim}{\hat{\theta}}\right) + pln(n) \tag{4.5}$$

onde $L\left(\underset{\sim}{\hat{\theta}}\right)$ é a função de máxima verosimilhança, p o número de parâmetros no modelo e n o tamanho da amostra (Stylianou et al., 2013).

Apesar dos critérios AIC e BIC serem ambos baseados no máximo da função de verossimilhança (MFV), existem algumas diferenças nas premissas por trás desses dois critérios. Por exemplo, enquanto que o AIC admite que entre os modelos avaliados nenhum é considerado o que realmente descreve o "modelo verdadeiro", o BIC assume que existe o modelo que descreve a relação entre as variáveis envolvidas, isto é, o "modelo verdadeiro" e tenta maximizar a probabilidade da sua escolha (Stylianou et al., 2013).

4.3.4 Pseudo R^2

Na regressão logística, a adequação de um modelo não se avalia da mesma forma que na regressão linear, que se pode avaliar pelo R^2. De forma a se poder avaliar até que ponto uma resposta binária pode ser prevista por um determinado modelo de regressão logística, muitas estatísticas diferentes de R^2 foram propostas nas últimas décadas, contudo nenhum deles foi aceite como padrão. Por exemplo, Mittlböck e Schemper (1996) citaram 12 medidas diferentes de R^2 e Menard (2000) considerou outras tantas medidas, sendo que concluiu que o índice *McFadden* (McFadden, 1974), também conhecido por R^2logit, seria o preferido dos cinco índices de pseudo-R^2 em estudo em um contexto de regressão logística.

O índice *McFadden* expressa a variação percentual entre o *Likelihood Value* do modelo, que considera somente a constante, e o *Likelihood Value*, que incorpora as variáveis explicativas, conforme se segue na seguinte expressão:

$$\mathrm{R^2logit} = \frac{[-2\mathrm{LLnulo} - (-2\mathrm{LLmodelo})]}{(-2\mathrm{LLnulo})} \tag{4.6}$$

onde LL representa o logaritmo das funções de verosimilhança.

Modelos que apresentem valores de *McFadden*-R^2 compreendidos entre 0,2 e 0,4 são considerados modelos bem ajustados (Z. Hu & Lo, 2007).

Outros pseudos R^2 que vêm a ser utilizado em diferentes estudos em que se realizam o ajuste de modelos de regressão logística são o Cox & Snell R^2 e Nagelkerke R^2. Segundo Walker & Smith (2016), o índice Cox & Snell representa-se pela seguinte expressão:

$$R^2_{CS} = 1 - \left[\frac{\mathrm{L(Null)}}{\mathrm{L(Full)}}\right]^{\frac{2}{N}} \tag{4.7}$$

onde $\mathrm{L(Null)}$ e $\mathrm{L(Full)}$ são as funções de verossimilhança para o modelo somente com a constante e o modelo com as variáveis independentes, respetivamente, e N corresponde ao tamanho da amostra. Já o índice Nagelkerke, que corresponde a uma versão melhorada de Cox & Snell, uma vez que, restringe o seu valor para que não exceda o valor de 1, é expresso pela seguinte equação:

$$R_N^2 = \frac{1 - \left(\frac{L(Null)}{L(Full)}\right)^{\frac{2}{N}}}{1 - L(Null)^{\frac{2}{N}}} \tag{4.8}$$

onde, neste caso, é realizado a divisão entre o índice de Cox & Snell pelo seu valor máximo possível.

Essas estatísticas, que geralmente são idênticas ao R^2 padrão quando aplicadas a um modelo de regressão linear geralmente enquadram-se em categorias de entropia, também conhecidas por pseudo-R^2, e de variância (B. Hu et al., 2006).

Apesar de existirem outras técnicas de pseudo R^2, neste trabalho serão somente citados o pseudo R^2 de *MacFadden*, o Cox & Snell R^2 e Nagelkerke R^2.

4.4 Diagnóstico de Pontos Influentes e de *Outliers*

Antes de se concluir de que o modelo se encontra ajustado, será essencial examinar outras medidas de forma a se avaliar se o modelo se encontra ajustado a todo o conjunto de padrões de covariáveis. Considera-se que o modelo ajustado contém p variáveis independentes, $\mathbf{x}' = (x_1, x_2, x_3, \dots, x_p)$, e que J represente o número distinto de valores observados de $\mathbf{x}'$(David W. Hosmer & Lemeshow, 2000). Para além disso, denota-se o número de casos com $\mathbf{x} = \mathbf{x}_j$ por $m_j, j = 1,2,3, \dots, J$ e que $\sum m_j = n$. Considerando-se igualmente que y_j representa o número de respostas positivas, $y = 1$, entre os casos m_j com $\mathbf{x} = \mathbf{x}_j$.

As quantidades chaves para os diagnósticos de regressão logística são os componentes de "soma dos quadrados residuais". Além disso, nesse tipo de regressão existe erros binomiais e, como resultado, a variância do erro é uma função de média condicional:

$$var\left(Y_j|\mathbf{x}_j\right) = m_j E\left(Y_j|\mathbf{x}_j\right) \mathrm{x} \left[1 - E\left(Y_j|\mathbf{x}_j\right)\right] = m_j \pi(\mathbf{x}_j)[1 - \pi\left(\mathbf{x}_j\right)] \tag{4.9}$$

Por isso, começa-se com os resíduos definidos nas equações (4.10) e (4.11) que foram "divididos" pela estimação dos seus erros padrões.

$$r(y_j, \hat{\pi}_j) = \frac{(y_j - m_j \hat{\pi}_j)}{\sqrt{m_j \hat{\pi}_j (1 - \hat{\pi}_j)}} \tag{4.10}$$

$$d(y_j, \hat{\pi}_j) = \pm\left\{2\left[y_j \ln\left(\frac{y_j}{m_j\hat{\pi}_j}\right) + (m_j - y_j)ln\left(\frac{(m_j - y_j)}{m_j(1-\hat{\pi}_j)}\right)\right]\right\}^{\frac{1}{2}} \quad (4.11)$$

Dado que cada resíduo foi dividido por uma estimativa aproximada de seu erro padrão, é espectável que, caso o modelo de regressão logística seja correto, essas quantidades tem uma média aproximadamente igual a zero e uma variância aproximadamente igual a 1.

Tal como acontece para os modelos de regressão linear, uma métrica utilizada para diagnosticar *outliers* em modelos de regressão logística passa pela utilização de valores *leverage* que derivaram do *"hat" matrix* (David W. Hosmer & Lemeshow, 2000). Sendo $\mathbf{X}$ a matriz J x $(\mathrm{p}+1)$ que contem os valores para todos os padrões de covariáveis J formados a partir dos valores observados das covariáveis p e $\mathbf{H}$ o *hat matrix*. Na equação (4.12) encontra-se representado a expressão do *hat matrix* para regressão logística:

$$\mathbf{H} = \mathbf{V}^{\frac{1}{2}}\mathbf{X}(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{\frac{1}{2}} \quad (4.12)$$

onde $\mathbf{V}$ é uma matriz diagonal J x J com elemento geral $v_j = m_j\hat{\pi}(\mathbf{x}_j)[1 - \hat{\pi}(\mathbf{x}_j)]$.

Sendo o j^{th} elemento da diagonal principal da matriz $\mathbf{H}$ denominado por h_j, pode ser mostrado que:

$$h_j = m_j\hat{\pi}(\mathbf{x}_j)\left[1 - \hat{\pi}(\mathbf{x}_j)\right]\mathbf{x}'_j(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{x}'_j = v_j \text{ x } b_j \quad (4.13)$$

onde $b_j = \mathbf{x}'_j(\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}\mathbf{x}'_j$ e $\mathbf{x}'_j = (1, x_{1j}, x_{2j}, \ldots,, x_{pj})$ é um vetor de valores de covariáveis definindo o j^{th} padrão de covariáveis. A soma dos elementos diagonais de $\mathbf{H}$ é $\sum h_j = (p+1)$, número de parâmetros no modelo. Se formularmos o *hat matrix* para a regressão logística como uma matriz n x n , cada elemento diagonal é delimitado de cima por $\frac{1}{m_j}$, onde m_j é o número total de casos com o mesmo padrão de covariáveis. Quando o *hat matrix* é baseado em dados agrupados por padrões de covariáveis, o limite superior para qualquer elemento diagonal é 1.

Contudo, caso o número de padrões de covariáveis for muito menor a n, poderá haver o risco de se falhar na identificação de pontos influentes e/ou um fraco ajustamento dos padrões de covariáveis. Considerando um padrão de covariáveis com m_j casos, $y_j = 0$ e probabilidade logística estimada $\hat{\pi}_j$. O resíduo de Pearson ao ser calculado individualmente para cada caso com este padrão de covariáveis, segundo a equação (4.10) é:

$$r_i = \frac{(0 - \hat{\pi}_j)}{\sqrt{\hat{\pi}_j(1-\hat{\pi}_j)}} = -\sqrt{\frac{\hat{\pi}_j}{(1-\hat{\pi}_j)}} \quad (4.14)$$

enquanto que, o resíduo de Pearson baseado em todos os casos com este padrão de covariáveis é:

$$r_i = \frac{(0 - m_j\hat{\pi}_j)}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}} = -\sqrt{m_j}\sqrt{\frac{\hat{\pi}_j}{(1-\hat{\pi}_j)}} \tag{4.15}$$

que aumenta negativamente conforme m_j aumenta.

Uma das questões em ter em mente quando se realiza a interpretação da magnitude de *leverage*, diz respeito ao efeito que v_j tem em h_j na equação (4.13). Segundo a literatura, o ajustamento determina os coeficientes estimados e, desde que o coeficiente estimado determina a probabilidade estimada, pontos com valores grandes de h_j são extremos no espaço da covariável e por isso, ficam afastados da média.

Hosmer & Lemeshow (2000) apresenta igualmente um exemplo de um gráfico de *leverage* versus probabilidades estimadas para uma amostra de 100 observações de um modelo de regressão logística com $g(x) = 0.8x$ e $x \sim N(0,9)$ (Figura 3).

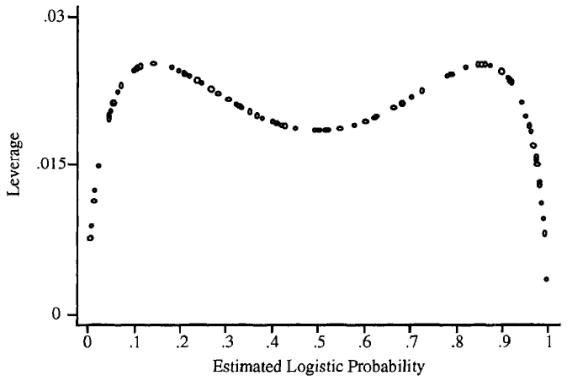

Figura 3 - Gráfico explicativo da relação entre *leverage* com probabilidade estimada (Fonte: Hosmer & Lemeshow, 2000).

Tal como é possível verificar-se na Figura 3, o *leverage* aumenta assim quando a probabilidade estimada se vai afastando do valor 0,5, até que o valor de probabilidade estimada se torne menor que 0,1 ou superior a 0,9. A partir desses momentos, os valores de *leverage* diminuem drasticamente até que atingem o valor de 0. Através deste exemplo, é possível verificar-se que o ponto mais extremo no espaço da covariável, pode ter o valor mais baixo de *leverage*.

Uma outra estatística de diagnóstico muito útil é uma que avalia o efeito num valor de coeficiente estimado e em todas as medidas de ajustamento, X^2 e D, quando se eliminar todos os casos com uma covariável padrão particular. Isto é obtido como a diferença padronizada entre $\widehat{\boldsymbol{\beta}}$ e $\widehat{\boldsymbol{\beta}}_{(-j)}$, que representam a máxima verosimilhança estimada. Neste caso, utiliza-se todos os J padrões de covariáveis e exclui-se o caso m_j com padrão $\mathbf{x}_j$ respetivamente. Essa quantidade para regressão logística seria:

$$\Delta\widehat{\boldsymbol{\beta}}_j = \left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{-j}\right)'(\mathbf{X}'\mathbf{V}\mathbf{X})\left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{-j}\right) = \frac{r_j^2 h_j}{(1-h_j)^2} = \frac{r_{sj}^2 h_j}{(1-h_j)} \tag{4.16}$$

Utilizando aproximações lineares semelhantes, é possível verificar-se que o decréscimo no valor de estatística de Pearson chi-quadrado devido à eliminação dos casos com padrão de covariável $\mathbf{x}_j$ é:

$$\Delta X_j^2 = \frac{r_j^2}{(1-h_j)} = r_{sj}^2 \tag{4.17}$$

Uma quantidade parecida poderá ser obtida pela mudança da *deviance*:

$$\Delta D_j = d_j^2 + \frac{r_j^2 h_j}{(1-h_j)} \tag{4.18}$$

Se substituirmos r_j^2 por d_j^2 irá produzir a seguinte aproximação:

$$\Delta D_j = \frac{d_j^2}{(1-h_j)} \tag{4.19}$$

Este tipo de estatística de diagnóstico é bastante atraente, uma vez que, permite identificar os padrões de covariáveis que são fracamente ajustados, ou seja, que apresentem altos valores de ΔX_j^2 e/ou de ΔD_j, e que têm uma grande influência nos valores dos parâmetros estimados (com grandes valores de $\Delta\widehat{\boldsymbol{\beta}}_j$) (David W. Hosmer & Lemeshow, 2000). Relativamente à medida ΔX_j^2, esta apresenta um valor baixo quando y_j e $m_j\hat{\pi}(\mathbf{x}_j)$ apresentam valores próximos. Isto é mais frequente acontecer quando $y_j = 0$ e $\hat{\pi}(\mathbf{x}_j) < 0{,}1$ ou quando $y_j = m_j$ e $\hat{\pi}(\mathbf{x}_j) > 0{,}9$. Já ΔX_j^2 apresenta um valor alto quando os valores de y_j e $m_j\hat{\pi}(\mathbf{x}_j)$ se encontram muito distantes. Isto ocorre mais frequentemente quando $y_j = 0$ e $\hat{\pi}(\mathbf{x}_j) > 0{,}9$ ou quando $y_j = m_j$ e $\hat{\pi}(\mathbf{x}_j) < 0{,}1$. Esses mesmos padrões de covariáveis não apresentam um valor alto de $\Delta\widehat{\boldsymbol{\beta}}_j$, pois quando $\hat{\pi}(\mathbf{x}_j) < 0{,}1$ ou $\hat{\pi}(\mathbf{x}_j) > 0{,}9$, $\Delta\widehat{\boldsymbol{\beta}}_j \approx \Delta X_j^2 h_j$, e h_j tende para zero. Para que $\Delta\widehat{\boldsymbol{\beta}}_j$ apresente um valor alto, é necessário que ΔX_j^2 e h_j apresentem valores pelo menos moderados. Isto é mais frequente acontecer quando $0{,}1 < \hat{\pi}(\mathbf{x}_j) < 0{,}3$ ou $0{,}7 < \hat{\pi}(\mathbf{x}_j) < 0{,}9$.

5. APLICAÇÃO *SHINY*

Atualmente, existem vários *softwares* comerciais, como o IBM®, SPSS®, MedCalc e Stata® e *software* livre, como o R, que são usados de forma a orientar vários profissionais de saúde nas suas tomadas de decisão. Contudo, cada uma dessas ferramentas de análise estatística oferece recursos diferentes, sendo umas mais limitadas do que outras.

Entre estes, o R tem demonstrado ser um dos mais completos, graças aos milhares de pacotes que possui. Alguns exemplos de pacotes muito utilizados e que serão utilizados ao logo deste trabalho são:

- *dplyr*: este pacote faz parte do pacote *tidyverse* e trata-se de uma gramática de manipulação de dados (Wickham & Grolemund, 2016). Este pacote é utilizado para transformar e resumir dados presentes num *dataframe*, sendo que, algumas operações comuns de manipulação de dados passam pela filtragem de linhas, seleção de colunas, reordenamento de linhas, adição de novas colunas, entre outras;
- *ggplot2:* trata-se de um sistema para a criação de gráficos estatísticos baseados no "Grammar of Graphics", que é composto por um conjunto de componentes independentes que podem ser compostos de maneiras diferentes (Wickham, 2015). A terminologia de "Grammar of Graphics", que foi criada por Wickham (2015), refere-se a um conjunto de regras que permitem estruturar elementos matemáticos e estéticos em um gráfico. Segundo este autor, este pacote é constituido por "elementos gramaticais", sendo eles os dados, a estética, as geometrias, que se refere aos elementos visuais como pontos, linhas e barras para representar os dados no gráfico, as facetas, que se refere à divisão dos dados em vários subconjuntos e a sua representação em gráficos, as escalas, as coordenadas e os temas, que permitem alterar a aparência de elementos que não sejam os dados (Wickham, 2016). Com o *ggplot2* é possível criar gráficos personalizados, o que não nos deixa limitados a um conjunto de gráficos pré-especificados;
- *caret*: este pacote é um dos mais utilizados no pré-processamento de dados e possui um conjunto de funções que simplificam o processo de criação de modelos preditivos. *Caret*, que, é abreviação de ***C****lassification* ***a****nd* ***re****gression* ***t****raining*, contém um conjunto de ferramentas úteis para a divisão de dados, pré-processamento, ajuste e treinamento de modelos, entre outros (Kuhn, 2019).
- *ROCR*: este pacote permite avaliar, visualizar e criar curvas ROC com parâmetros de corte e permite combinar duas de 25 medidas de desempenho, sendo que, uma nova medida de desempenho pode ser adicionada usando para isso uma interface padrão (Sing et al., 2015).

Algumas dessas medidas são a precisão, fração de verdadeiros positivos, fração de falsos positivos, fração de verdadeiro negativo, fração de falso negativo, área sob a curva ROC, entre outros.

O R é uma linguagem e um ambiente de *software* livre e de código aberto muito utilizado por estatísticos e analistas de dados para desenvolver *software* de estatística. Como o R se trata de um programa que é conduzido por meio de digitação de linhas de comando, o que dificulta a sua operacionalização por pessoas que não têm qualquer conhecimento de programação, foram desenvolvidas interfaces gráficas, para facilitar o seu uso, como o RStudio. O RStudio é um ambiente de desenvolvimento integrado para programar em R, que reúne os vários componentes do R, como a consola, edição de código-fonte, gráficos, *knitr* entre outros, em um ambiente de trabalho único (Gandrud, 2013).

Porém, quando um profissional, que não tem conhecimento da programação R, pretender obter uma avaliação rápida dos seus dados, o uso de uma interface baseada em comandos como o R, poderá ser desafiador e demorado. Neste sentido muitos autores têm desenvolvido ferramentas *web* de análise rápida, gratuita e de fácil uso, que têm como base a linguagem R. Daí surgiu a necessidade de fazer uso do pacote *shiny*, presente na biblioteca do *software* R, que permite construir aplicações *web*. Por exemplo, na área clínica, Pallmann et al. (2019) desenvolveram uma aplicação *web* gratuita chamada de MoDEsT, que tem como objetivo estimar a dose máxima tolerada, na criação de novos tratamentos, a ser administrada a humanos. Esta aplicação usa um procedimento de decisão bayesiano baseada em regressão logística. Por outro lado, Goksuluk et al. (2016) desenvolveram a aplicação easyROC baseada na linguagem R, que fornece estatísticas ROC e ferramentas gráficas, sendo capaz de calcular o ponto de corte ideal da curva ROC, comparar vários marcadores, entre outras coisas, de modo a apoiar os pesquisadores na tomada das suas decisões, sem terem de escrever uma única linha de código em R.

Sendo o *shiny* um pacote que permite a criação de *web apps*, sem haver a necessidade prévia de conhecimentos de linguagens de desenvolvimento *web* como o HTML, CSS e JavaScript, os responsáveis pelo desenvolvimento conseguem implementar todas as ferramentas estatísticas do R em uma aplicação utilizando somente a linguagem de programação R (Costa, 2018). Isto é possível, pois o *shiny* faz uma espécie de conversão de linguagem, de R para HTML.

Relativamente à sua estrutura, o *shiny* divide-se basicamente em dois componentes, a interface do utilizador (UI) e o *server.* Estes dois componentes, que são constituídos por um conjunto de comandos R, podem ser escritos em dois *scripts* R diferentes, mas inseridos na mesma diretoria, ou então em um único *script* chamado *app.R.* Caso se opte pela opção de criar dois *scripts* diferentes, será necessário

criar um servidor *Rshiny*, que é constituído pelo *script server.R* ,e uma interface de utilizador, *ui.R*, que se encontram no mesmo local da diretoria (Seal & Wild, 2018).

A componente *ui.R*, que diz respeito ao arquivo de definição da interface do utilizador, é usado para configurar o que o utilizador irá ver na aplicação *web, ou seja, o* documento *web* em HTML, também conhecido por documento *user interface* (UI), e para aceitar informações que o utilizador vai inserindo (Sivaprakasam & Sadagopan, 2019). Já o *server.R*, contém as instruções em R necessárias para o computador criar a aplicação. Estas instruções serão principalmente necessárias para o servidor saber o que fazer quando o utilizador alterar as opções na aplicação.

O *shiny* vem com uma variedade de *widgets* que permite a criação rápida de interfaces dos utilizadores e também faz todo o trabalho em termos de configuração das interfaces interativas dos utilizadores (Beeley, 2013). Para além disso, é muito fácil integrar aplicações *Shiny* com conteúdos da *web*, usando para isso HTML, CSS, JavaScript e jQuery. Para além das funcionalidades fornecidas pelo pacote *shiny*, existem outros pacotes de extensão que complementam a funcionalidade do mesmo, como por exemplo, *shinythemes* (Chang et al., 2018), que permite implementar temas Bootstrap na aplicação, *shinydashboard* (Chang et al., 2018), que permite criação de dashboards, *shinyjs* (Attali, 2020), que permite implementar funcionalidades de JavaScript na aplicação *Shiny*, entre outros.

O *shiny* é ferramenta que faz uso da "programação reativa", o que garante que as alterações no *input* sejam imediatamente refletidas no *output*, possibilitando desta forma, a construção de uma ferramenta altamente interativa (Seal & Wild, 2018).

Por fim, é possível colocar uma aplicação *Shiny* acessível a outros utilizadores, sendo que algumas plataformas hospedeiras de tais aplicações poderão ser o "Shinyapps.io" ou até mesmo o GitHub (RStudio, 2020).

6. RESULTADOS

Toda a análise estatística realizada aos dados ao longo deste capítulo, assim como, a construção do algoritmo e da aplicação *web*, foram feitas utilizando a ferramenta RStudio versão 1.3.959 e R versão 4.0.1.

6.1 Caracterização da Base de Dados

O Registo Nacional de Muito Baixo Peso funciona em Portugal desde 1994, sendo que integra o registo de todos os recém-nascidos com peso inferior a 1500 gramas ao nascer e/ou idade gestacional inferior a 32 semanas, independentemente do peso de nascimento, e gémeos independentemente do peso e da idade gestacional. O seu principal objetivo consiste no registo da população sobrevivente dos RNMBP para avaliação prospetiva do neuro desenvolvimento e das sequelas (Cunha et al., 2010).

Para a realização desta dissertação, foi utilizada uma base de dados, fornecida gentilmente pelo Registo Nacional de Muito Baixo Peso, que engloba o registo de todos os recém-nascidos de muito baixo peso, nascidos no período entre 2010 e 2012, de todo o território português. Todos os recém-nascidos registados apresentavam um peso inferior a 1500 gramas e/ou uma idade gestacional inferior a 32 semanas.

A base de dados inicial era composta por 3496 registos e 105 variáveis. Porém, para este estudo, só se fez uso de uma amostra composta por 2306 registos, no qual se excluiu todos os gémeos e todos os recém-nascidos que morreram na sala de parto. Desta amostra, 263 (11,41%) dos recém-nascidos de muito baixo peso acabaram por morrer.

Neste subcapítulo fez-se uma análise exploratória aos dados, tendo-se para isso utilizado alguns pacotes presentes no R, nomeadamente *ggplot2* (versão 3.3.1) para a construção de gráficos personalizados, *tidyverse* (versão 1.3.0) e *dplyr* (verão 1.0.0) para transformação e manipulação dos dados, *scales* (versão 1.1.1) para visualização e *cowplot* (versão 1.0.0) para a junção de gráficos, construídos em *ggplot2,* numa só figura.

No Apêndice I – Base de Dados encontra-se representado uma breve descrição e a codificação de cada nível das variáveis do qual se fizeram uso na análise exploratória e na construção do classificador.

Na Tabela I e na Tabela II encontra-se representado a distribuição de registos de RNMBP por ano ,e a distribuição de registos de RNMBP segundo o estado de amissão em cada ano, respetivamente. Aqui, é possível verificar-se que, a tendência ao longo destes três anos, tem sido a diminuição do número de

nascimentos de recém-nascidos de muito baixo peso, sendo o ano 2010 o que apresenta o maior número. Para além disso, também se verifica que destes bebés, a maioria acabou por sobreviver, sendo que, o ano em que ocorreu mais mortes desta população foi em 2011.

Tabela I - Distribuição de registos de recém-nascidos por ano.

Ano	Número de recém-nascidos	Percentagem de recém-nascidos (%)
2010	790	34,26
2011	763	33,09
2012	753	32,65
Total	2306	100,00

Tabela II - Distribuição de registos de recém-nascidos sobrevivos e falecidos por ano.

	Sobrevivo		Falecido	
Ano	**Número de recém-nascidos**	**Recém-nascidos (%)**	**Número de recém-nascidos**	**Recém-nascidos (%)**
2010	716,00	90,63	74,00	9,37
2011	662,00	86,76	101,00	13,24
2012	665,00	88,31	88,00	11,69
Total	2043,00	88,60	263,00	11.40

Fazendo um enquadramento geográfico, segundo a distribuição destes recém-nascidos em cada distrito do país, tendo como base a residência das progenitoras, verifica-se que o distrito de Lisboa é o que apresenta um maior número de recém-nascidos de muito baixo peso, com um valor de 653 bebés, o que equivale a 30,47% dos recém-nascidos (RN) (Figura 4). Já o distrito da Guarda é o que apresenta o menor número desta população, com um valor de 15 RN, ou seja, 0,70% dos recém-nascidos de muito baixo peso. Nesse caso é evidente que, no geral, os distritos portugueses que apresentam uma maior densidade populacional, acabam por apresentar igualmente uma maior percentagem de RNMBP.

Por outro lado, no gráfico da Figura 5 encontra-se representado a distribuição dos recém-nascidos de muito baixo peso sobrevivos e falecidos segundo o distrito de residência. Para todos os distritos é evidente que dos RNMBP, a maioria acabou por sobreviver, sendo a percentagem de falecidos reduzido. Neste caso, está evidente que o distrito de Bragança é o que apresenta uma maior percentagem de recém-nascidos falecidos, com um valor de 29,4%, enquanto que o distrito de Évora é o que apresenta a percentagem de recém-nascidos falecidos mais baixo, com um valor de 4,7%.

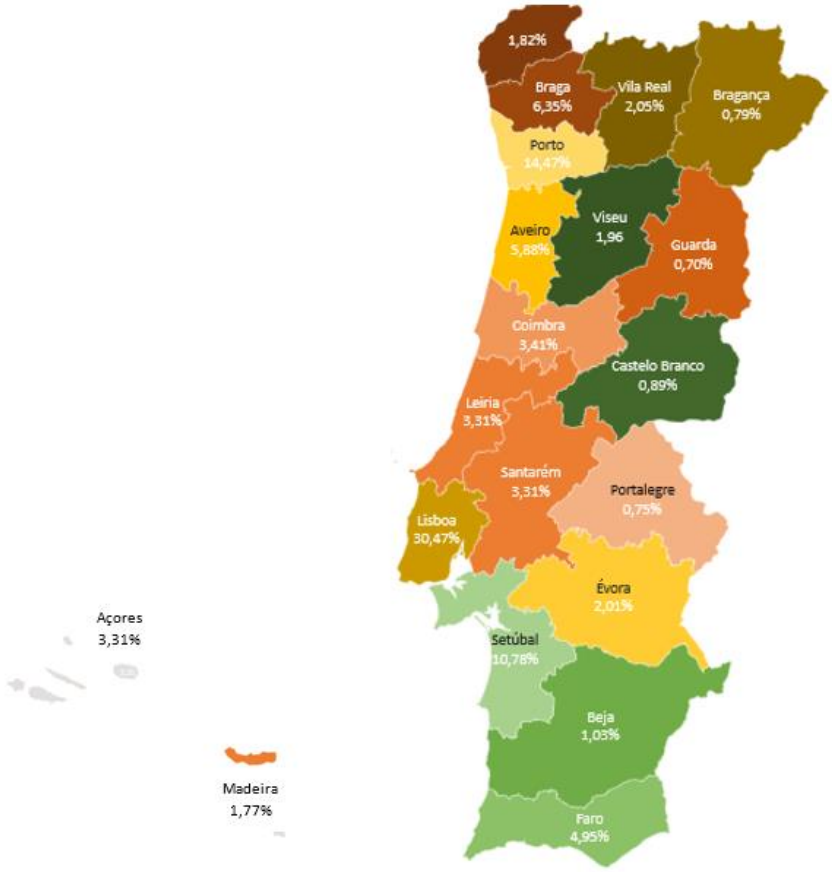

Figura 4 - Distribuição dos recém-nascidos de muito baixo peso segundo o distrito de residência.

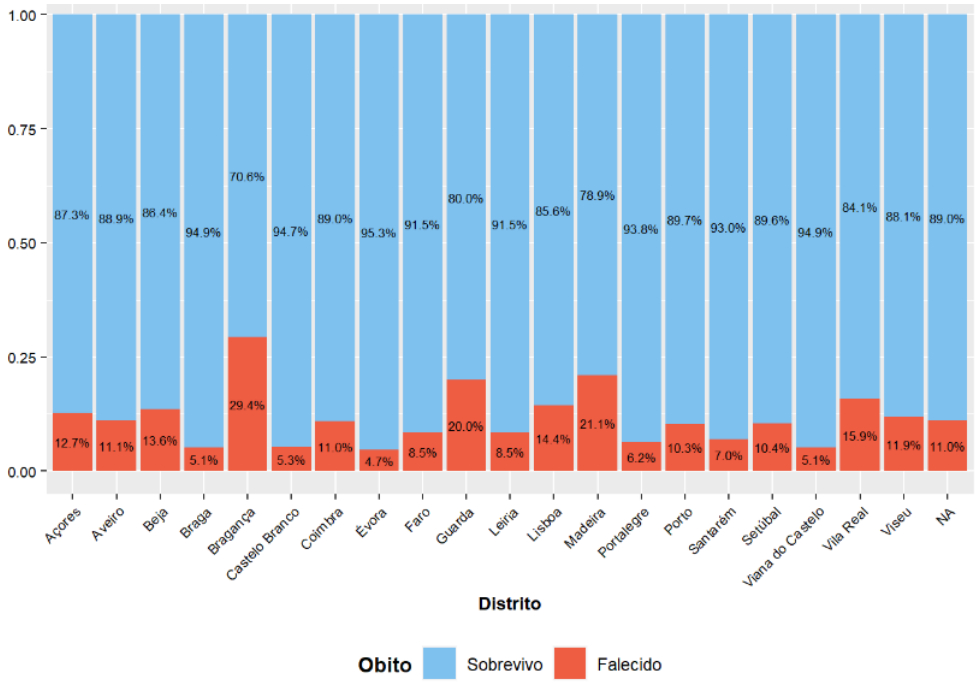

Figura 5 - Distribuição dos recém-nascidos de muito baixo peso falecidos e sobrevivos por distrito.

Na Tabela III é possível verificar-se a distribuição dos RNMBP quanto ao local de nascimento. De acordo com a classificação do registo dos recém-nascidos, os que foram classificados como "outborn", significa que o RNMBP nasceu fora do hospital de registo, podendo ter sido transferido de outro hospital ou ter nascido, por exemplo, em casa. Já se este for classificado como "inborn", significa que o recém-nascido nasceu no hospital responsável pelo registo. Neste caso, consta-se que 93,06% destes RNMBP nasceram no hospital responsável pelo registo (Tabela III). Igualmente se verifica que, dos bebés que não nasceram no hospital responsável pelo registo, uma grande parte deles acabaram por nascer num hospital de apoio perinatal (80,63%), sendo que, nenhum proveio de um hospital privado. Destes 129 bebés que nasceram num hospital de apoio perinatal, somente 17 deles acabaram por falecer.

Tabela III - Distribuição dos recém-nascidos de muito baixo peso quanto ao local de nascimento.

		Número de RNMBP	RNMBP (%)
Local de nascimento	*Inborn*	2146,00	93,06
	Outborn	160,00	6,94
	Total	**2306**	
Outborn	Hospital de Apoio Perinatal Diferenciado	9,00	5,63
	Hospital de Apoio Perinatal	129,00	80,63
	Instituição de Saúde sem Apoio Perinatal	1,00	0,63
	Local Extra Hospitalar	21,00	13,13
	Hospital Privado	0,00	0,00
	Total	**160**	

Relativamente às mães destes recém-nascidos, que se encontram registados nesta base de dados, apresentam uma idade média de 30,25 anos com um desvio padrão de 6,15 anos, sendo que, o valor mínimo registado para a idade das mães é de 14 anos e a máxima de 47 anos. Para além disso, 50% destas mães apresentam uma idade igual ou inferior a 31 anos, 25% idade inferior ou igual a 26 anos e 25% idade superior a 35 anos.

Considerando o estado de admissão dos recém-nascidos, a idade média das progenitoras para os sobreviventes e falecidos é de 30,30 e 29,88 anos, respetivamente.

No que diz respeito à gravidez e ao parto, na Tabela IV encontra-se representado algumas das principais características a eles referentes. Nota-se que a maior parte das progenitoras apresentavam patologias na gravidez (66,10%) e que a cesariana foi o tipo de parto mais frequente (69,12%). Também se verifica que o motivo de parto espontâneo foi o mais predominante entre as mães (44,93%), seguido pela

patologia fetal (28,83%), IVG (25,90%) e a patologia materna (0,35%). Da mesma forma, no estudo liderado por Lee & Gould (2006), a cesariana foi o tipo de parto mais frequente entre os RNMBP.

Tabela IV – Distribuição dos recém-nascidos segundo as características associadas à gravidez e parto.

		Número de RN	RN (%)	RN sobrevivos (%)	RN falecidos (%)
Patologias na gravidez	Sim	1521,00	66,10	89,15	10,85
	Não	780,00	33,90	87,56	12,44
	Total	**2301**			
Tipo de parto	Vaginal	712,00	30,88	83,84	16,15
	Cesariana	1594,00	69,12	90,72	9,28
	Total	**2306**			
Motivo do parto	Espontâneo	1027,00	44,93	86,56	13,44
	Patologia Materna	8,00	0,35	50,00	50,00
	Patologia Fetal	659,00	28,83	90,29	9,71
	IVG	592,00	25,90	90,71	9,29
	Total	**2286**			

Relativamente a cuidados e tratamentos realizados durante o período pré-natal, é possível verificar-se pela Tabela V, que 93,40% das progenitoras receberam cuidados antes do nascimento dos seus filhos e somente 4,09% das mães estiveram sujeitas à conceção medicamente assistida. No que diz respeito à administração de corticoides antes do parto, 12,91% das progenitoras não foram administradas com corticoide, 25,96% delas tiveram o seu parto em menos de 24 horas após a 1ª dose de corticoides ou mais de uma semana após a última dose ,e por fim, a maioria das mães (61,13%), tiveram o seu parto em mais de 24 horas e menos de uma semana, após a administração de pelo menos uma dose de corticoides. Na mesma tabela encontra-se representado a distribuição dos recém-nascidos por estado de admissão para cada caso.

Tabela V - Distribuição dos recém-nascidos segundo os cuidados e tratamentos efetuados ou não antes dos seus nascimentos.

		Número de RN	RN (%)	RN sobrevivos (%)	RN falecidos (%)
Cuidados pré-natais	Sim	2151,00	93,40	88,89	11,11
	Não	152,00	6,60	85,53	14,47
	Total	**2303,00**			
Conceção Assistida	Sim	94,00	4,09	82,98	17,02
	Não	2207,00	95,91	88,81	11,19

Tabela V - Distribuição dos recém-nascidos segundo os cuidados e tratamentos efetuados ou não antes dos seus nascimentos. (Continuação)

		Número de RN	RN (%)	RN sobrevivos (%)	RN falecidos (%)
	Total	**2301,00**			
Corticoides pré-natais	Não	296,00	12,91	82,43	17,57
	Parcial	595,00	25,96	88,91	11,09
	Completo	1401,00	61,13	89,65	10,35
	Total	**2292,00**			

Os recém-nascidos que se incluem neste estudo apresentam algumas características referidas na Tabela VI. Estas características foram registadas nas primeiras 24 horas de vida de cada recém-nascido de muito baixo peso, quer este tenha nascido no hospital de registo ou tenha sido transferido para o mesmo. Assim, segundo a Tabela VI, estes bebés apresentam uma média e um desvio padrão de idade gestacional de 207,80 ± 18,13 semanas e um peso ao nascer com uma média e desvio padrão de 1196 ± 356,97 g. Para além disso, a média e desvio padrão que os RN apresentam, ao nível de comprimento ao nascer, é de 37,16 ± 3,62 cm e o perímetro cefálico ao nascer com uma média e desvio padrão de 26,57 ± 2,66 cm.

Tabela VI - Resumo das estatísticas sumárias para a avaliação dos RN registadas nas primeiras 24 horas de vida.

	Idade gestacional (dias)	Peso (g)	Comprimento (cm)	Perímetro cefálico (cm)
Média	207,80	1196,00	37,16	26,57
Desvio padrão	18,13	356,97	3,62	2,66
Mediana	210,00	1210,00	37,50	27,00
Máximo	280,00	2810,00	49,00	36,00
Mínimo	160,00	370,00	25,00	18,00
Percentil 25	196,00	930,00	35,00	25,00
Percentil 75	220,00	1440,00	40,00	28,50

Sendo o peso uma das variáveis com mais expressão nos indicadores de mortalidade, pretende-se fazer uma análise mais elaborada do mesmo. Esta variável que apresenta como valor máximo 2810,00 gramas e valor mínimo 370,00 gramas, também indica que nesta amostra de RNMBP, 25% deles apresentam um peso ao nascer inferior ou igual a 930,00g e 25% um peso superior a 1440,00 (Tabela VI). Ao analisar o seu valor de mediana, verifica-se que 50% destes recém-nascidos apresentam um peso ao nascer igual ou inferior a 1210,00g. A distribuição dos valores do peso ao nascer dos 2306 recém-

nascidos em estudo é possível verificar-se na Figura 6. Nesta representação gráfica é possível concluir-se, que uma grande parte dos RN em estudo, apresentam um peso ao nascer muito baixo.

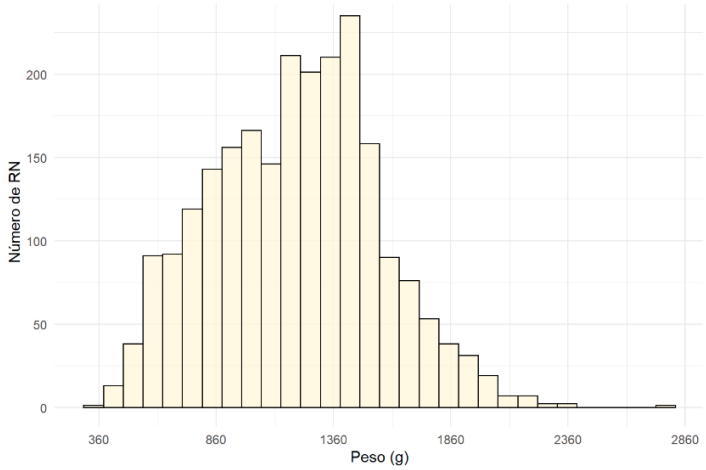

Figura 6 - Histograma que representa a distribuição do peso ao nascer da amostra em estudo.

Relativamente à variação dos pesos dos RNMBP segundo o estado de admissão, sobrevivo e falecido, apura-se que, no geral, os sobrevivos apresentam um peso à nascença maior do que os recém-nascidos que acabam por falecer (Figura 7). Na Tabela VII é possível observar-se um resumo das estatísticas sumárias dos RNMBP sobreviventes e falecidos em função do peso ao nascer, o que acaba por comprovar o que foi dito anteriormente. Tal conclusão também foi verificada no estudo levado a cabo por Kardum et al. (2019)

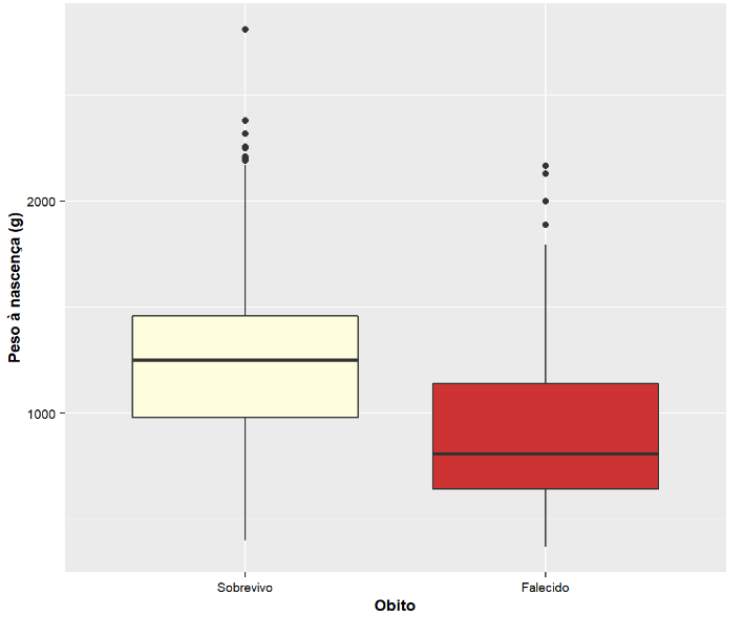

Figura 7 – *Box plot* que representa a distribuição do peso ao nascer, por sobrevivo e falecido.

Tabela VII - Resumo das estatísticas sumárias da variável peso ao nascer, tendo em conta o estado de admissão dos recém-nascidos de muito baixo peso.

Peso ao nascer		
	Sobrevivo	**Falecido**
Média	1233,65	903,28
Desvio Padrão	341,43	339,81
Mediana	1250,00	810,00
Máximo	2810,00	2167,00
Mínimo	400,00	370,00

Uma outra variável muito analisada no estudo da mortalidade destes recém-nascidos é a idade gestacional. Tal como acontece com o fator peso à nascença, também é mais comum os sobrevivos apresentarem um valor de idade gestacional maior do que os recém-nascidos que acabam por falecer (Figura 8). Este resultado também foi atingido por Zile et al. (2017). Relativamente aos RN sobrevivos, 50% deles apresentam uma idade gestacional igual ou inferior a 212 dias, enquanto 50% dos falecidos apresentam uma idade gestacional igual ou inferior a 186 dias, ou seja, 27 semanas. Para além disso, segundo esta base de dados em estudo, não existem RN falecidos com idades gestacionais superiores a 268 dias, isto é, 38 semanas.

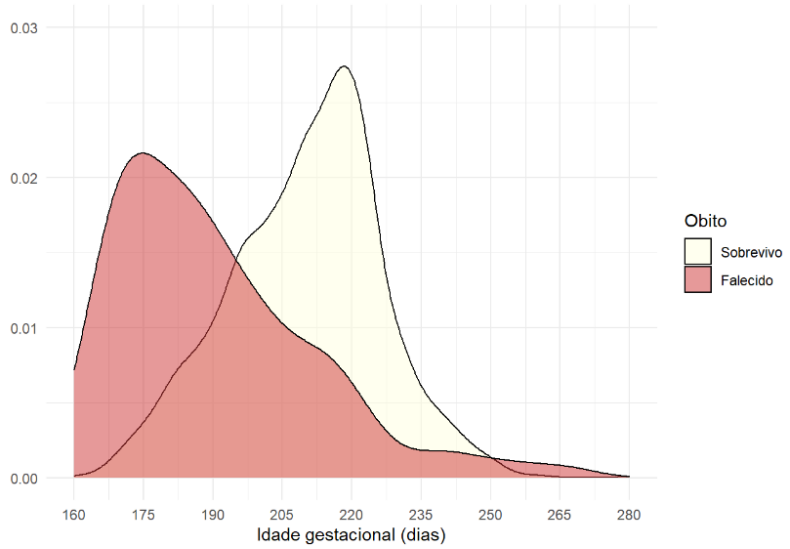

Figura 8 – Representação gráfica da distribuição da idade gestacional em dias, por sobrevivo e falecido.

Nesta amostra de 2306 bebés, 1207 (52,34%) deles correspondem a recém-nascidos do sexo masculino e 1099 (47,66%) do sexo feminino. Além disso, consta-se que a incidência da mortalidade é ligeiramente superior no sexo masculino com um valor de 6,7%, em comparação com o sexo feminino que apresenta um valor de 4,7% de recém-nascidos falecidos (Figura 9). A questão de a mortalidade ser mais comum

nos recém-nascidos do sexo masculino do que do sexo feminino, foi igualmente verificado por Naskar et al. (2014).

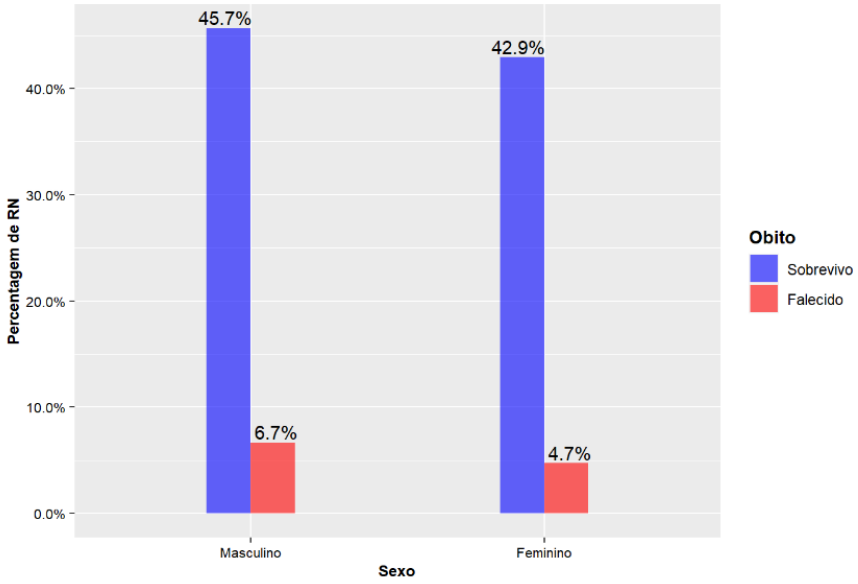

Figura 9 – Distribuição dos recém-nascidos de muito baixo peso segundo o sexo e o resultado.

Um dos primeiros exames a ser realizado aos recém-nascidos é feito logo no primeiro minuto de vida. Este teste, conhecido como índice de apgar, tem como objetivo avaliar o nível de adaptação do bebé à vida extrauterina, ajudando a identificar se existe qualquer necessidade de se recorrer a algum tipo de tratamento extra após o nascimento. Este teste, que é feito logo no primeiro minuto de nascimento e é repetido novamente no quinto e décimo minutos, é composto por cinco categorias, sendo elas a cor de pele, a pulsação arterial, a frequência respiratória, a atividade muscular e a resposta reflexa. A cada uma destas categorias é atribuído um valor compreendido entre 0 e 2, segundo o estado do bebé, sendo que a soma destes cinco sinais varia entre 0 e 10. Quanto mais próximo do valor 10 estará a avaliação, melhor será o estado de saúde do bebé.

Uma vez que, dos três índices de agpar (1°min,5°min,10°min), o índice apgar ao 10° minuto é o mais importante, optou-se por criar um índice novo que engloba a média destes três índices. Este índice apresenta uma média de 7,94 e uma mediana de 8. O valor mais baixo presente na amostra de estudo é de valor 1, sendo que três recém-nascidos apresentaram essa pontuação, e 419 recém-nascidos receberam uma pontuação de 10, o valor máximo do índice. Dentro desta escala, o valor 9 foi o mais frequente, tendo 599 recém-nascidos recebido essa pontuação.

Na Figura 10 encontra-se esquematizado um gráfico que representa a distribuição dos valores do índice apgar segundo o estado de admissão do RNMBP, sobrevivo ou falecido. Neste caso verifica-se, que bebés ao qual lhes foram atribuídos uma pontuação de 1,2 e 4 no índice de apgar, uma grande parte deles

correspondem a recém-nascidos que acabaram por falecer. Já bebés que apresentam um maior valor no índice apgar, uma grande parte deles acabaram por sobreviver, sendo que, a pontuação 10 no índice de apgar é o que apresenta o menor número de RN falecidos (2,4%). Relativamente aos RN sobrevivos, 50% deles apresentaram um valor de índice de apgar igual ou inferior a 8, enquanto que 50% dos RN falecidos apresentaram um valor igual ou inferior a 6.

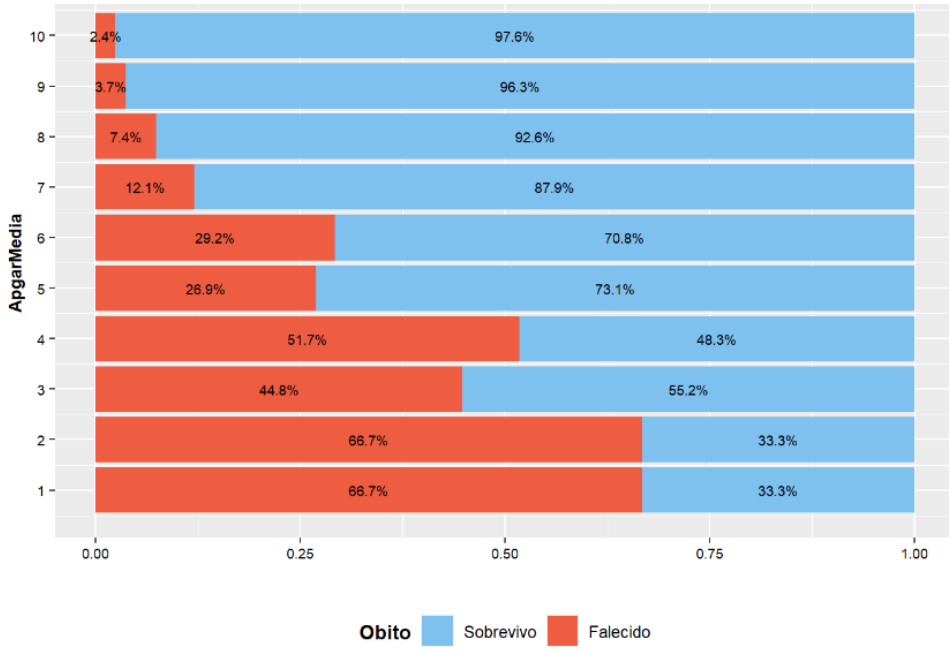

Figura 10 - Distribuição dos valores do índice apgar segundo o estado de admissão do recém-nascido.

Atualmente, é muito frequente os profissionais de saúde recorrerem a escalas de avaliação de gravidade clínica que permitem avaliar a perspetiva de sobrevivência e a qualidade de vida dos recém-nascidos de muito baixo peso, sendo o CRIB e o SNAPPE II muito utilizados. Na secção 6.2 encontra-se explicado mais detalhadamente em que consiste estas duas escalas de gravidade que serão alvo de estudo nesta dissertação, assim como, outras escalas que são igualmente muito faladas na literatura.

Para esta amostra em estudo, o índice CRIB apresenta uma média de pontuação de 2,74, uma mediana de 1 e uma pontuação final compreendida entre 0 e 20. Já o índice SNAPPE II apresenta uma média de pontuação de 21,64, uma mediana de 15 e uma pontuação final que varia entre 0 e 131.

Nas Figura 11 e Figura 12 estão representadas as distribuições das escalas CRIB e SNAPPE II, respetivamente, segundo o estado de admissão dos recém-nascidos.

Para o CRIB, no qual foram considerados válidos 2163 respeitantes à classificação atribuída pela escala, verifica-se que a maioria dos bebés sobrevivos apresentam uma pontuação igual ou abaixo de 10 valores, sendo o valor 1 na escala o mais frequente. Para além disso, a todos os bebés que lhes foram atribuídos uma pontuação compreendida entre os valores 17 e 20 acabaram por falecer.

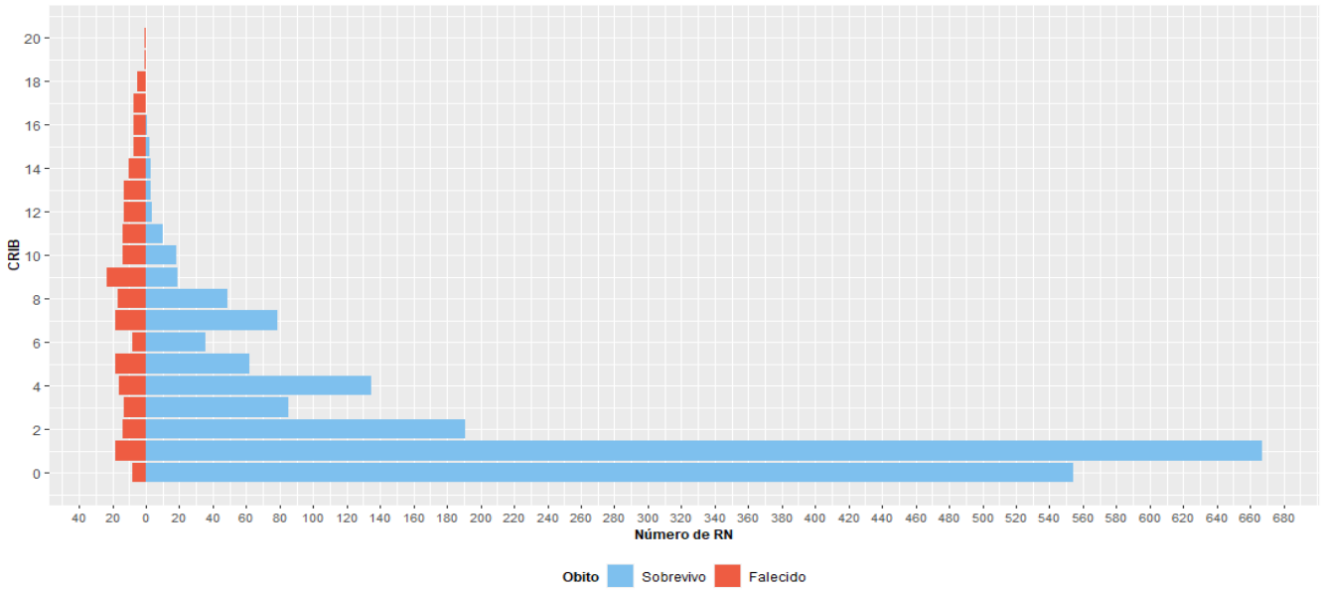

Figura 11 - Distribuição do CRIB segundo o estado de admissão dos recém-nascidos.

Da mesma forma, para o SNAPPE II, em que foram considerados 1687casos válidos, a maioria dos recém-nascidos sobrevivos em estudo foram pontuados com valores mais baixos da escala, sendo o valor 0 o mais frequente.

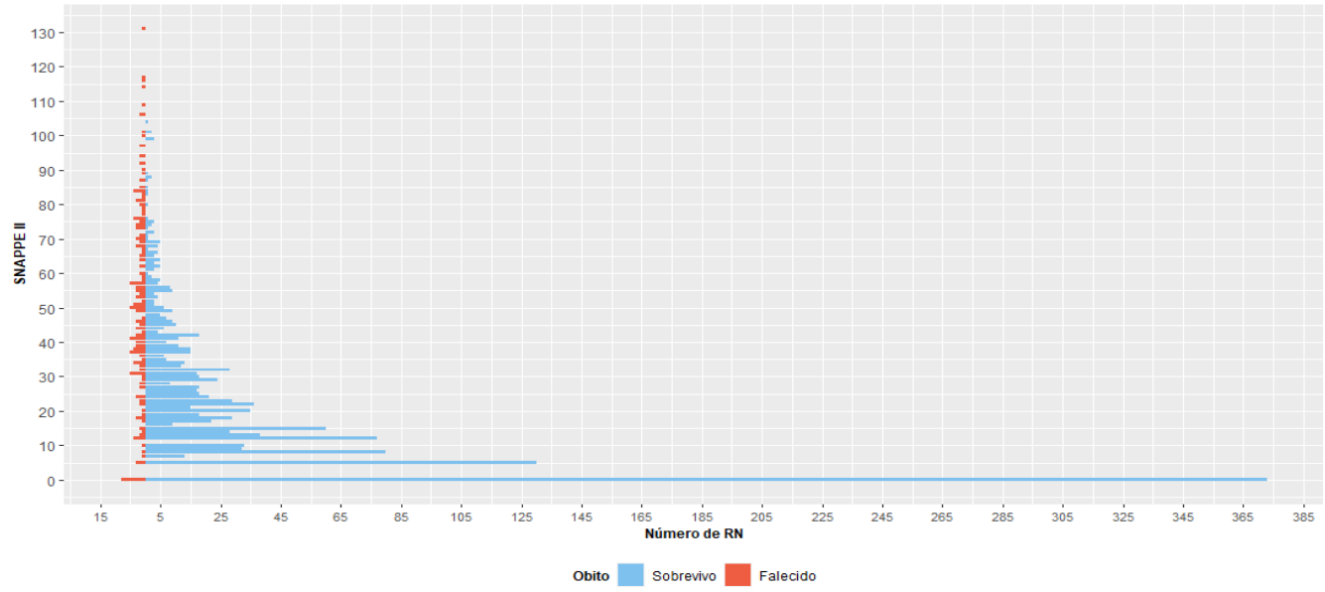

Figura 12 - Distribuição do SNAPPE II segundo o estado de admissão dos recém-nascidos.

Por vezes, alguns recém-nascidos necessitam de serem reanimados de forma a ajudá-los a se adaptarem à vida extrauterina. Na Tabela VIII encontra-se representado cinco diferentes processos de reanimação que o recém-nascido pode ou não ter enfrentado na sala de parto ou em qualquer outro local onde o bebé tenha nascido, caso a situação o tenha pedido. Também se encontra representado a percentagem de sobrevivos e falecidos para cada processo de reanimação caso tenha ou não a ele recorrido. Consta-se que, dos métodos de reanimação disponíveis, somente o processo de ressuscitação com oxigénio foi o mais utilizado, sendo que 60,72% destes recém-nascidos em estudo tiverem de recorrer a este sistema. Todavia as ressuscitações por compressão cardíaca e por adrenalina foram os métodos pelo qual os profissionais de saúde menos optaram, sendo que somente 4,27% e 3,70% dos bebés estiveram sujeitos a tais procedimentos, respetivamente.

Tabela VIII - Distribuição dos recém-nascidos de muito baixo peso segundo os diferentes tipos de ressuscitação.

Ressuscitação		Número de RN	RN (%)	RN sobrevivos (%)	RN falecidos (%)
Oxigénio	Sim	1388,00	60,72	83,72	16,28
	Não	898,00	39,28	96,55	3,45
	Total	**2286,00**			
Insuflador	Sim	1065,00	46,43	84,41	15,59
	Não	1229,00	53,57	92,43	7,57
	Total	**2294,00**			
Entubação Et	Sim	925,00	40,20	77,62	22,38
	Não	1376,00	59,80	96,08	3,92
	Total	**2301,00**			
Compressão Cardíaca	Sim	98,00	4,27	67,35	32,65
	Não	2199,00	95,73	89,68	10,32
	Total	**2297,00**			
Adrenalina	Sim	85,00	3,70	61,18	38,82
	Não	2211,00	96,30	89,69	10,31
	Total	**2296,00**			

No que diz respeito aos diferentes diagnósticos que estes recém-nascidos poderão realizar, verifica-se na Tabela IX, que entre os oito tipos de diagnósticos, a doença mais frequentemente detetada entre estes recém-nascidos é o síndrome dificuldade respiratória (SDR) com um valor de 75,24%. Tal resultado foi obtido igualmente por Bernstein et al. (2000), em que 76% dos RN foram diagnosticados com essa doença. Já as restantes doenças apresentam uma taxa de diagnóstico muito mais reduzida, sendo que a menos comum entre estes bebés é a perfuração gastrointestinal (GI) com um valor de 2,69%.

Segundo Hull et al. (2014), a enterocolite necrotizante é uma das emergências gastrointestinais que mais contribui para a mortalidade, sendo que, no seu estudo, dos 9% dos bebés que apresentavam NEC, 28% deles acabaram por falecer.

Na Tabela IX também se encontra representado a percentagem de sobrevivos e falecidos para cada tipo de doença que possa ter sido diagnosticada.

Tabela IX - Distribuição dos recém-nascidos de muito baixo peso segundo o diagnóstico.

Diagnóstico		Número de RN	RN (%)	RN sobrevivos (%)	RN falecidos (%)
SDR (Síndrome Dificuldade Respiratória)	Sim	1735,00	75,24	86,22	13,78
	Não	571,00	24,76	95,80	4,20
	Total	**2306,00**			
Pneumotorax	Sim	123,00	5,33	65,85	34,15
	Não	2183,00	94,67	89,88	10,12
	Total	**2306,00**			
PDA	Sim	612,00	26,80	81,21	18,79
	Não	1672,00	73,20	92,11	7,89
	Total	**2284,00**			
Enterocolite Necrotizante - NEC	Sim	173,00	7,50	69,94	30,06
	Não	2133,00	92,50	90,11	9,89
	Total	**2306,00**			
Perfuração gastrointestinal (GI) focal	Sim	62,00	2,69	67,74	32,26
	Não	2244,00	97,31	89,17	10,83
	Total	**2306,00**			
Malformação Congénita Major	Sim	125,00	5,43	73,60	26,40
	Não	2178,00	94,57	89,49	10,51
	Total	**2303,00**			
Sepsis Meningite Tardia	Sim	794,00	34,43	89,17	10,83
	Não	1512,00	65,57	88,29	11,71
	Total	**2306,00**			
Sepsis Meningite Precoce	Sim	271,00	11,75	80,44	19,56
	Não	2035,00	88,25	89,68	10,32
	Total	**2306,00**			

Do mesmo modo, na Tabela X encontra-se representado a distribuição dos recém-nascidos de muito baixo peso segundo o exame imagiológico, que possam ou não ter realizado, os respetivos resultados a nível dos graus que apresentam para HPIV (hemorragia peri ou intraventricular) e para LPV (leucomalácia

periventricular), assim como, o registo relativo a se tiveram ou não algum enfarte venoso e dilatação ventricular pós-hemorrágica . Para além disso, também se encontra representado a percentagem de sobrevivos e falecidos para cada fator.

Daqui, observa-se que, 97,04% dos recém-nascidos em estudo realizaram o exame imagiologia cerebral ao 28º dia. Relativamente ao exame HPIV, 69,27% dos RNMBP não apresentaram qualquer evidência de HIV, apesar de terem falecido 4,82% desses bebés, e 8,80% dos recém-nascidos apresentaram o grau mais avançado de HPIV, sendo que 51,28% desses bebés acabaram por falecer. Resultados semelhantes foram obtidos por Bernstein et al. (2000), que indica que a incidência de hemorragia intraventricular foi de 28%, sendo que 9% dos RNMBP apresentaram HPIV grave (grau III ou IV). Por outro lado, 84,33% dos bebés não apresentaram evidência de LPV, onde 9,02% acabaram por morrer. Também se verifica que, 85,71% dos bebés não tiveram qualquer enfarte venoso, porém dos restantes bebés que tiveram, 51,61% deles acabaram por falecer. Por fim, 81,94 dos RN não tiveram nenhuma dilatação ventricular pós-hemorrágica.

Tabela X - Distribuição dos recém-nascidos de muito baixo peso segundo o exame imagiológico e os seus resultados a nível dos graus que apresentam para HPIV e LPV.

Imagem		Número de RN	RN (%)	RN sobrevivos (%)	RN falecidos (%)
Imagiologia Cerebral Dia 28	Sim	2230,00	97,04	90,00	10,00
	Não	68,00	2,96	44,12	55,88
	Total	**2298,00**			
HPIV	0	1535,00	69,27	95,18	4,82
	1	291,00	13,13	94,50	5,50
	2	195,00	8,80	87,69	12.31
	3	195,00	8,80	48,72	51,28
	Total	**2216,00**			
LPV grau	0	1841,00	84,33	90,98	9,02
	1	239,00	10,95	91,63	8,37
	2	52,00	2,38	96,15	3,85
	3	33,00	1,51	84,85	15,15
	4	18,00	0,82	72,22	27,78
	Total	**2183,00**			
Enfarte Venoso (EVHP)	Sim	93,00	14,29	48,39	51,61
	Não	558,00	85,71	84,95	15,05
	Total	**651,00**			

Tabela X - Distribuição dos recém-nascidos de muito baixo peso segundo o exame imagiológico e os seus resultados a nível dos graus que apresentam para HPIV e LPV. (Continuação)

Imagem		Número de RN	RN (%)	RN sobrevivos (%)	RN falecidos (%)
Dilatação Ventricular pós-hemorrágica	Sim	117,00	18,06	70,94	29,06
	Não	531,00	81,94	82,11	17,89
	Total	**648,00**			

Relativamente a possíveis procedimentos e tratamentos que os RNMBP podem estar sujeitos, encontra-se representado na Tabela XI a distribuição dos recém-nascidos segundo cada procedimento e tratamento, assim como a distribuição de sobrevivos e falecidos para cada caso.

Tabela XI - Distribuição dos recém-nascidos de muito baixo peso segundo os procedimentos e tratamentos.

Procedimentos e Tratamentos		Número de RN	RN (%)	RN sobrevivos (%)	RN falecidos (%)
Exame Oftalmológico	Sim	1622,00	71,30	98,77	1,23
	Não	653,00	28,70	63,25	36,75
	Total	**2275,00**			
Ecografia Tf	0	13,00	0,62	100,00	0,00
	1	279,00	13,23	81,72	18,28
	2	453,00	21,48	89,40	10,60
	3	480,00	22,76	90,83	9,17
	10	884,00	41,92	92,19	7,81
	Total	**2109,00**			
SR Oxigénio	Sim	1842,00	79,88	86,32	13,68
	Não	464,00	20,12	97,63	2,37
	Total	**2306,00**			
SR Cpap	Sim	1715,00	74,40	95,22	4,78
	Não	590,00	25,60	69,32	30,68
	Total	**2305,00**			
SR Vppni	Sim	488,00	21,58	92,21	7,79
	Não	1773,00	78,42	87,42	12,58
	Total	**2261,00**			
SR Vaf	Sim	433,00	18,79	69,05	30,95
	Não	1872,00	81,21	93,11	6,89
	Total	**2305,00**			

Tabela XI - Distribuição dos recém-nascidos de muito baixo peso segundo os procedimentos e tratamentos. (Continuação)

Procedimentos e Tratamentos		Número de RN	RN (%)	RN sobrevivos (%)	RN falecidos (%)
SR Vafni	Sim	6,00	0,26	50,00	50,00
	Não	2300,00	99,74	88,70	11,30
	Total	**2306,00**			
Surfactante Inicial	Sim	194,00	8,42	80,41	19,59
	Não	2109,00	91,58	89,33	10,67
	Total	**2303,00**			
Surfactante Posterior	Sim	1009,00	43,77	81,17	18,83
	Não	1296,00	56,23	94,37	5,63
	Total	**2305,00**			
Pda Profilático	Sim	5,00	0,22	100,00	0,00
	Não	2301,00	99,78	88,57	11,43
	Total	**2306,00**			
Pda Terapeutico	Sim	321,00	13,92	86,92	13,08
	Não	1985,00	86,08	88,87	11,13
	Total	**2306,00**			
Cirurgia Pda	Sim	81,00	3,51	88,89	11,11
	Não	2225,00	96,49	88,58	11,42
	Total	**2306,00**			
Cirurgia Nec	Sim	63,00	2,73	55,56	44,44
	Não	2243,00	97,27	89,52	10,48
	Total	**2306,00**			
Cirurgia Major Outra	Sim	103,00	4,47	89,32	10,68
	Não	2203,00	95,53	88,56	11,44
	Total	**2306,00**			

Por fim, apresenta-se na Figura 13 a distribuição do tempo de internamento dos RNMBP sobrevivos num histograma (A) e num *box plot* (B). Segundo esta amostra de bebés, o tempo médio de internamento é de 49,12 dias com um desvio padrão de 41,00 dias. O valor mínimo registado para o tempo de internamento é 0 dias, enquanto que, o valor máximo é de 524, ou seja, cerca de 75 semanas. Por outro lado, 50% destes RNMBP estiveram internados nos centros hospitalares 41 dias ou menos, 25% tiveram internados 27 dias ou menos e outros 25% estiveram internados 63 dias ou mais.

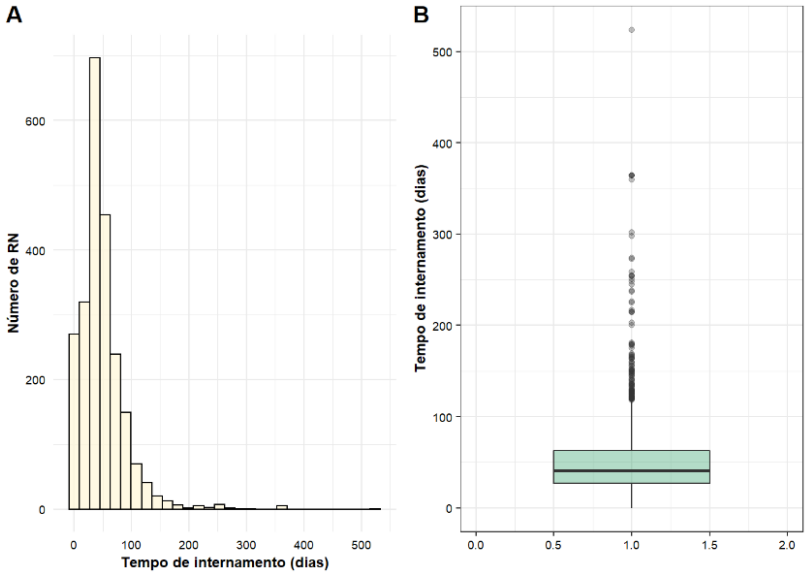

Figura 13 - Histograma que representa a distribuição do tempo de internamento dos recém-nascidos de muito baixo peso sobrevivos (A) e representação de *box plot* do tempo de internamento dos mesmos recém-nascidos (B).

6.2 Escalas de gravidade clínica para bebés recém-nascidos de muito baixo peso

Durante muito tempo, o peso do recém-nascido e a idade gestacional foram considerados as medida mais importante de risco neonatal inicial, pois eram facilmente avaliadas (Gooden et al., 2014; Mourão, 2016). Contudo, devido às limitações que apresentavam na avaliação do risco clínico de recém-nascidos com muito baixo peso, aliado à necessidade crescente de obter informação cada vez mais rigorosa, começou por surgir novas formas de avaliação mais precisas para o risco de mortalidade neonatal, conhecidas por escalas de gravidade clínica.

Estes novos sistemas de pontuação surgem assim para apoiar a melhoria progressiva dos indicadores de mortalidade neonatal já existentes, que são frequentemente utilizados na avaliação da qualidade dos cuidados prestados nas unidades de cuidados intensivos neonatais (Mourão, 2016). Por outro lado, graças a estas novas escalas, foi possível realizar a comparação entre serviços, regiões e mesmo países (Braga, 2000).

As escalas de gravidade clínica podem envolver o uso de dados demográficos, fisiológicos e clínicos dos bebés recém-nascidos, de modo a que seja possível calcular uma pontuação que quantifica a sua morbidade e mortalidade (Dorling et al., 2005).

A criação destas escalas poderá ser feita de duas formas diferentes (Dorling et al., 2005). Uma delas são as escalas "médicas", em que, um conjunto de especialistas recorrem ao conhecimento clínico de forma a selecionar as variáveis a serem incluídas na pontuação e os seus pesos relativos. Um outro caso

passa pela utilização de modelos estatísticos que irão identificar as variáveis que apresentam uma forte associação com o resultado de interesse e os seus pesos relativos. De facto, existem evidências que, a longo prazo, as pontuações "estatísticas" superam as pontuações "médicas". Contudo, o conhecimento clínico pode contribuir igualmente para a escolha das variáveis a ser incluídas num modelo final, pois poderá ser visto como mais confiável por parte de utilizadores.

Entre as diferentes escalas, algumas são mais simples, apresentando poucas variáveis, o que torna a sua aplicação mais rápida, e outras mais completas, apresentando mais variáveis contudo, apresentam uma aplicação mais demorada (Brito et al., 2003).

Segundo Dorling et al. (2005), para uma boa escala neonatal, esta deverá apresentar algumas características, tais como, ser de fácil uso, deve ser aplicável no início do processo de hospitalização, deve apresentar a capacidade preditiva da mortalidade ou morbidade para várias categorias de neonatos e deve ter utilidade para todos os grupos de recém-nascidos. Contudo, nem sempre é possível alcançar completamente essas propriedades. Também é preciso ter em consideração que, nem as melhores escalas são completamente precisas, pois não existem fórmulas matemáticas que consigam captar completamente todos os processos clínicos de um neonato.

Posto isto, será feito uma breve descrição de algumas escalas de gravidade que medem o risco de mortalidade de recém-nascidos com muito baixo peso.

6.2.1 CRIB - *Clinical Risk Index for Babies*

Esta escala, que foi criada com o intuito de prever a mortalidade de bebés nascidos com menos de 32 semanas de gestação, tem como base dados de bebés admitidos em quatro unidades neonatais terciárias do Reino Unido de 1988 a 1990 (Dorling et al., 2005).

Para a sua construção, os autores utilizaram a regressão logística de modo a identificar as seis variáveis mais preditivas de mortalidade, sendo estas o peso ao nascer, idade gestacional, malformações congénitas, excesso de base máximo nas primeiras 12 horas e os valores máximos e mínimos de FiO_2, isto é, fração inspirada de oxigénio, nas primeiras 12 horas após o parto (A. C. da S. Braga & Oliveira, 2003).

A sua pontuação final consiste na soma dessas seis variáveis, sendo que o seu valor final poderá estar compreendido entre 0 e 23. Cada variável da escala CRIB tem um valor numérico pré-determinado que varia conforme a gravidade (Tabela XII). Quanto mais alto for a pontuação final da escala, maior será a probabilidade de o recém-nascido vir a morrer.

Uma das vantagens desta escala deve-se à sua facilidade de recolha de dados, necessitando de pouco tempo para o cálculo da pontuação de cada recém-nascido (Dorling et al., 2005). Para além disso, o facto do CRIB ser avaliado nas primeiras 12 horas de vida, torna-o menos suscetível aos efeitos do tratamento a que as crianças possam estar sujeitas (M. F. Mourão et al., 2015).

Tabela XII - Escala CRIB com os possíveis valores de pontuação que cada variável pode tomar, adaptado de (Sarquis et al., 2002).

VARIÁVEL		PONTUAÇÃO
PESO AO NASCER (G)	>1350	0
	851 - 1350	1
	701 - 850	4
	≤ 700	7
IDADE GESTACIONAL (SEMANAS)	> 24	0
	≤ 24	1
MALFORMAÇÃO CONGÉNITA	Ausente	0
	Sem risco de vida imediata	1
	Com risco de vida imediata	3
EXCESSO DE BASE MÁXIMO (MMOL/L)	> -7,0	0
	-7,0 a -9,9	1
	-10,0 a -14,9	2
	< -15,0	3
MÍNIMO DE FIO_2	≤ 0,40	0
	0,41 - 0,60	2
	0,61 - 0,90	3
	0,91 - 1,00	4
MÁXIMO DE FIO_2	≤ 0,40	0
	0,41 - 0,80	1
	0,81 - 0,90	3
	0,91 – 1,00	5

6.2.2 CRIB II - *Clinical Risk Index for Babies II*

O facto de a adequação do CRIB aos dados ter sido questionada, pois o *score* pode ser mal calibrado com a mortalidade após a terapia intensiva neonatal, levou Parry et al. (2003) a desenvolver uma nova escala, o CRIB II. Esta surgiu como uma versão aperfeiçoada do CRIB, de modo a prever a mortalidade de recém-nascidos com 32 semanas de gestação.

A escala CRIB II prevê a mortalidade dos recém-nascidos por 5 variáveis, o peso ao nascer, idade gestacional, sexo , excesso de base máximo nas primeiras 12 horas e temperatura no momento da hospitalização, tendo assim, excluído variáveis que poderiam ter influencia pelos cuidados prestados ao bebé (Dorling et al., 2005; Jafrasteh et al., 2017). O resultado desta escala é baseado na soma destas 5 variáveis, sendo que, o seu resultado final poderá estar compreendido entre 0 e 27 (Jafrasteh et al., 2017).

6.2.3 SNAP - *Score for Neonatal Acute Physiology*

O SNAP, que é a principal alternativa à escala CRIB, foi desenvolvido em 1990 e tem como base dados de três unidades de Boston e EUA relativos a 1643 bebés, no qual, 154 pesavam menos de 1500g ao nascer (Dorling et al., 2005). Este sistema de pontuação diferencia-se do CRIB respetivamente à sua aplicabilidade. Enquanto que a escala CRIB tem aplicabilidade em neonatos com peso de nascimento inferior a 1500g, o SNAP pode ser utilizado em todas as idades gestacionais e todos os pesos (Jafrasteh et al., 2017). O facto do SNAP apresentar uma reduzida sensibilidade às diferenças entre os prematuros, deve-se ao reduzido número de recém-nascidos com muito baixo peso existentes nos dados utilizados para o desenvolvimento da escala (Dorling et al., 2005).

Os scores do SNAP baseiam-se em 28 variáveis recolhidas durante as primeiras 24 horas de vida, sendo elas pressão sanguínea, frequência cardíaca, frequência respiratória, temperatura, PO_2 (pressão parcial de oxigénio), razão entre PO_2 e FiO_2, PCO_2 (pressão parcial de gás carbónico no sangue arterial), índice de oxigenação, volume de glóbulos, contagem de células brancas do sangue, rácio total de imaturidade, contagem absoluta de neutrófilos, contagem de plaquetas, uréia no sangue, creatinina, produção de urina, bilirrubina indireta, bilirrubina direta, sódio, potássio, cálcio ionizado, cálcio total, glicose, bicarbonato sérico, pH sérico, convulsão, apneia e *stool guaiac test* (métodos que deteta a presença de sangue oculto nas fezes) (Mourão, 2016).

Ao contrário da escala CRIB, onde os parâmetros são ponderados de acordo com sua relação estatística com a morte, na escala SNAP, as variáveis foram ponderadas de acordo com a opinião de especialistas, que atribuem a cada variável uma pontuação de 0,1,3 ou 5 (Dorling et al., 2005). Esta escala poderá apresentar um valor final compreendida entre 0 e 123, sendo que, quanto maior for o seu valor, maior a probabilidade do recém-nascido vir a falecer (Mourão, 2016).

6.2.4 SNAPPE - *Score for Neonatal Acute Physiology Perinatal Extension*

Devido ao difícil uso e ao grande número e complexidade das varáveis que a escala SNAP apresentava, em 1993 Richardson desenvolveu o SNAPPE, no qual mais tarde, em 2001, veio a simplificá-lo (Harsha & Archana, 2015; Mia et al., 2005). A escala SNAPPE é constituída por seis variáveis fisiológicas e três variáveis de risco de mortalidade perinatal (Mia et al., 2005).

6.2.5 SNAP II – *Score for Neonatal Acute Physiology II*

Devido à dificuldade na recolha de dados para o SNAP e para o SNAPPE, os autores desenvolveram uma nova escala, o SNAP II, que tem como base, dados recolhidos de 30 unidades do Norte de América de 10 819 recém-nascidos (Dorling et al., 2005).

Esta nova escala distingue-se pela redução do período de recolha de dados para 12 horas e pela redução de número de variáveis. Assim, o SNAP II é constituído por seis parâmetros fisiológicos, nomeadamente a pressão média arterial, razão entre pressão parcial de oxigénio e fração de oxigénio inspirado, temperatura mais baixa registada (em °F), pH sérico, ocorrência de múltiplas convulsões e produção de urina (<1mL / kg / h), sendo estas recolhidas dentro de um período de 12 horas (Sundaram et al., 2009). Este sistema de pontuações admite valores compreendidos entre 0 e 115, sendo que, maiores valores corresponde a uma maior probabilidade do recém-nascio vir a falecer (Mourão, 2016).

6.2.6 SNAPPE II – *Score for Neonatal Acute Physiology Perinatal Extension II*

Pelos mesmos motivos de dificuldade na recolha de dados para a escala SNAPPE, os mesmos autores que criaram o SNAP, desenvolveram uma nova escala de gravidade, o SNAPPE II (Dorling et al., 2005). Para a sua criação, foram utilizados dados das mesmas 30 unidades norte-americanas referidas na subsecção anterior, tendo neste caso, utilizado dados de 14 610 recém-nascidos.

Este sistema de pontuação envolve o registo de nove parâmetros, como a pressão arterial média, a razão entre pressão parcial de oxigénio e fração de oxigénio inspirado, temperatura mais baixa em °F, pH sérico, convulsões múltiplas, produção de urina, peso ao nascer, pontuação de apgar e idade gestacional, ou seja, houve uma inclusão de 3 novas variáveis relativamente à escala apresentada anteriormente. (Rachuri et al., 2019).

6.2.7 NTISS - *National Therapeutic Intervention Scoring System*

Esta escala foi desenvolvida por Gray et al.(1992), tendo surgido como uma modificação do índice de TISS (*Therapeutic Intervention Scoring System*), que era utilizado nas unidades de cuidados intensivos para adultos (Dorling et al., 2005). Das 76 variáveis do TISS, 42 foram excluídas e 28 variáveis foram adicionadas para formar o NTISS (Gray et al., 1992). Esta escala pode apresentar um resultado final compreendido entre 0 e 47. A capacidade do escore do NTISS, avaliado 24 horas após a admissão, para prever a mortalidade em recém-nascidos a termo e prematuros tem sido boa (Wu et al., 2015).

6.2.8 NICHHD - *National Institute of Child Health and Human Development*

Para o desenvolvimento da escala NICHHD, foram incluídos fatores observados em sete unidades neonatais dos Estados Unidos de 1823 crianças, nascidas entre os anos 1987 e 1989, e que apresentavam um peso entre 501g e 1500g (Dorling et al., 2005). Esta escala não tem sido muito utilizada desde o seu desenvolvimento. A regressão logística foi utilizada de forma a selecionar as variáveis que fazem parte desta escala, nomeadamente o peso ao nascer, baixa idade gestacional, raça, sexo e *score* de Apegar ao 1° minuto.

6.2.9 *Berlin Score*

Esta escala alemã, que permite a classificação precoce na admissão, foi desenvolvida usando o método de regressão logística aplicado a dados de recém-nascidos de muito baixo peso nascidos entre 1988 e 1991 (Dorling et al., 2005). Este sistema de pontuação tem em consideração os seguintes fatores: peso ao nascer, índice de apgar aos cinco minutos de idade, o início da ventilação artificial, o grau de síndrome do desconforto respiratório e excesso de base na admissão (Maier et al., 2002).

6.2.10 NEOCOSUR - *Neonatal del Cono Sur*

Esta escala de risco de mortalidade neonatal para recém-nascidos de muito baixo peso foi desenvolvida em 2005 por Marshall et al., tendo como base, algumas variáveis obtidos no momento do nascimento, antes da admissão na unidade de terapia intensiva neonatal (Marshall et al., 2005). Para tal, foi utilizado dados de bebés com peso de nascimento de 500 a 1500 g, nascidos de 1 de outubro de 2000 a 30 de maio de 2003 em 16 centros da Rede Neonatal del Cono Sur (NEOCOSUR). Esta escala demonstrou ter um bom desempenho de previsão em uma população da rede sul-americana, tendo sido considerada o melhor modelo para ser aplicado em países em desenvolvimento.

6.3 Tratamento da Base de Dados

Uma vez que, o objetivo principal desta dissertação passa pelo desenvolvimento de um modelo capaz de prever o risco de morte de um recém-nascido de muito baixo peso, será necessário, numa primeira fase, tratar a base de dados de modo a excluir as variáveis que não devem entrar na construção do modelo como variáveis preditoras, assim como, tratar os valores omissos.

Primeiramente, excluiu-se todos os dados relativos a recém-nascidos gémeos, pois esses dados podem causar enviesamento na informação. Depois optou-se por eliminar as 16 observações relativos aos recém-nascidos que acabaram por morrer na sala de partos, pois seria incorreto, durante o processo de tratamento da base de dados, imputar valores para variáveis que só podem ser medidas depois do recém-nascido sair da sala de parto. Também se optou por eliminar um registo de um recém-nascido cujo sexo era indeterminado, devido à ambiguidade de ser ou não hermafrodita.

Posteriormente, começou-se por realizar uma análise mais direcionado às variáveis, tendo-se eliminado numa primeira fase, 51 variáveis da base de dados inicial, devido à falta de relevância que apresentavam para o caso de estudo. Muitas delas foram eliminadas por serem variáveis com pouca variabilidade, não sendo medidas em todos os recém-nascidos. Outras foram excluídas por dependerem do registo de outras variáveis ou por simplesmente não apresentarem expressão para modular, tais como, variáveis relativas a datas. Por outro lado, as variáveis que correspondiam a índices de gravidade clínica (*Crib* e *Snape II*), foram igualmente excluídas da base de dados, pois, como apresentam um bom desempenho na avaliação de risco de morte para bebés com muito baixo peso, poderão ser utilizadas na comparação com o modelo a ser desenvolvido. Além disso, não seria igualmente aconselhável introduzir estas duas variáveis no modelo, pois esses indicadores já apresentam um conjunto de variáveis, o que poderia provocar efeitos de colinariedade muito fortes. Variáveis como "Transporte" e "NascimentoTipoLocal", que vão ao encontro da variável "NascimentoOutborn", foram igualmente eliminadas de forma a evitar a ocorrência de enviesamento.

Após se ter eliminado as observações e variáveis anteriores, passou-se à fase de recodificação da base de dados. Durante o processo de análise da base de dados, encontraram-se alguns registos com valores e denominações estranhos, tais como, "-999", "999", "9", "Indeterminado", "Não aplicável", "Desconhecido" e "Causa Desconhecida". Apesar destas entradas se encontrarem preenchidas, não nos fornecem qualquer tipo de informação relevante para que seja possível realizar uma análise. Deste modo, todas estas entradas foram recodificadas para NA's, ou seja, como valores omissos. Relativamente à variável "MaeIdade", recodificaram-se 37 registos com valor zero para valores omissos, uma vez que

estes valores refletem erros cometidos aquando do registo. Também se fizeram algumas recodificações de níveis de algumas variáveis para que fosse mais percetível o seu significado.

Como todo processo de modelação requer uma variável dependente, para este caso a variável "ObitoData" foi a escolhida. Contudo, foi necessário fazer algum trabalho de limpeza pois esta só apresentava valores correspondentes a datas de óbito. Assim, todo o registo que apresentasse a data de morte passou a tomar o valor de "1", ou seja, o recém-nascido faleceu, e todas as células que se encontravam vazias passaram a tomar valor de "0", ou seja, o recém-nascido sobreviveu. Além disso, renomeou-se esta variável para "Obito".

Por fim, visto que seria muito provável que, as variáveis relativas aos índices de apgar ao 1º, 5º e 10º minuto iriam contribuir para a mesma situação, optou-se por criar uma nova variável que correspondesse à média destes três índices, dando assim origem à variável "ApgarMedia".

O passo seguinte passou por analisar a percentagem de NA's que cada variável apresentava, eliminando aquelas que apresentassem uma percentagem acima de 60%. Este valor estipulado foi naturalmente uma escolha arbitrária. Para se ter uma perceção visual de como se encontra a base de dados a nível de valores omissos, fez-se uso do pacote *naniar* (versão 0.6.0) presente no RStudio. Este pacote possui a função gráfica *vis_miss* que fornece uma melhor perceção visual dos valores omissos na base dados, assim como, a percentagem de *missing value* presente em cada variável. Deste modo, obteve-se a Figura 14 que nos mostra que, a base de dados inicial com 54 variáveis que se preparou para se fazer uso no processo de modelação, apresenta um valor de 6,3% *missing value*, sendo que, três dessas variáveis apresentam uma percentagem acima dos 60% de *missing values*.

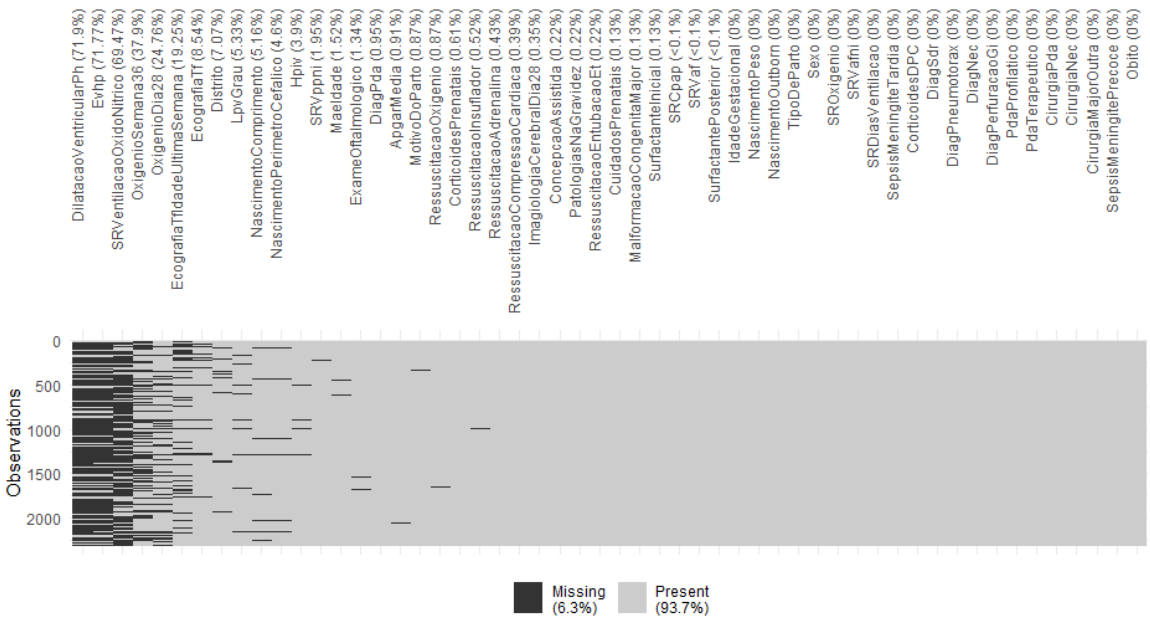

Figura 14 - Representação gráfica dos valores omissos presentes em cada variável da base de dados.

Após se ter eliminado essas três variáveis da base de dados, e antes de se ter prosseguido para os métodos de tratamento dos valores ausentes, analisou-se as restantes variáveis que apresentavam uma maior percentagem de *missing values*, de forma a perceber se teriam alguma importância para o estudo. Daqui, verificou-se que as variáveis "OxigenioSemana36" e "OxigenioSemana36" deveriam ser eliminadas da base de dados, pois quando relacionadas com a variável dependente, os valores presentes nessas variáveis tinham correspondência a um número muito reduzido de recém-nascidos falecidos, o que não permitia extrair conclusões muito fidedignas no que diz respeito a se essas variáveis de facto teriam algum impacto na mortalidade desses recém-nascidos.

Posto isto, chegou-se ao fim com uma base de dados constituída por 49 variáveis e 2306 observações, assim como, uma percentagem de 1,3% *missing values*.

6.4 Tratamento dos Valores Omissos

Para que a construção do modelo preditivo seja possível se realizar, é necessário que a base de dados se encontre livre de valores omissos. Uma vez que, a base de dados candidata para a modelação possuía 1,3% de *missing values*, tornou-se inevitável neste estudo a aplicação de metodologias de tratamento de valores omissos.

Assim, para este trabalho, começou-se por criar diferentes base de dados, no qual se aplicou a cada uma delas, um método de tratamento de valores ausentes diferentes. Neste caso, testou-se os métodos de deleção de casos, imputação simples, tendo-se aplicado a imputação pela moda em variáveis qualitativas e imputação pela média e pela mediana para variáveis quantitativas, imputação múltipla e k-vizinhos mais próximos.

Contudo, dado o baixo número de valores omissos presentes na base de dados, e visto que, ao aplicar o método de deleção nos dados corresponde a trabalhar com uma base de dados original que representa a realidade, pois correspondem a dados reais, optou-se por dar continuidade ao desenvolvimento do modelo preditivo tendo como ponto de partida a base de dados deletada. É de salientar que, a utilização de uma base de dados no qual se implementa um método de tratamento de valores omissos, que não seja a deleção, sem conhecer a condição real, nas variáveis qualitativas muitas vezes a sua imputação poderá levar com que essa variável não apresente variabilidade.

De seguida, criou-se para a base de dados escolhida, um conjunto de treino constituído por 70% dos dados iniciais e um conjunto de teste constituído por 30% dos dados iniciais. Para isso, recorreu-se ao pacote *caret* (versão 6.0-86), presente no R, que permite a divisão dos dados em dados treino e teste

para o processo de criação de modelos preditivos. Os dados de treino serão utilizados para treinar o modelo, enquanto que, os dados de teste serão utilizados para validar o modelo desenvolvido. Neste caso, a base de dados tratada segundo a metodologia de deleção, passou a conter um conjunto de treino composto por 1075 registos e um conjunto de teste por 460 registos, o que equivale a um total de 1535 casos e 49 variáveis.

6.5 Construção do Modelo de Regressão Logística

O modelo de regressão logística a ser desenvolvido nesta dissertação, terá como finalidade ser usado como um classificador por parte dos profissionais de saúde, para que estes consigam prever qual o resultado final de um recém-nascido de muito baixo peso, ou seja, se este irá sobreviver ou falecer. Para isso, será integrado no modelo um conjunto de variáveis que são ou não rotineiramente medidas nestes bebés. Por outro lado, pretende-se implementar este algoritmo (modelo) numa aplicação *Shiny,* de forma a facilitar todo o processo de cálculo de previsão do estado de vida dos RN, por parte destes profissionais. Para se conseguir obter o modelo final a ser implementado na aplicação, existe um conjunto de cinco passos principais a serem realizados, tal como se pode observar na Figura 15.

Figura 15 - Representação esquemática dos passos a seguir para a construção do modelo preditivo final.

Assim, antes de se proceder à construção do modelo de regressão logística propriamente dita, foi necessário concretizar os três primeiros passos, presentes na Figura 15, de forma a escolher o melhor conjunto de varáveis independentes a introduzir no modelo preditivo, eliminando aquelas que não apresentam variabilidade. Para isso, foram utilizadas técnicas estatísticas e exploratórias que permitiram eliminar variáveis que à priori não iriam contribuir em nada matematicamente para o modelo.

Para o primeiro passo, começou-se por realizar um procedimento univariado variável a variável, tendo-se eliminado aquelas que não eram significativas, ou seja, que apresentavam um *p-value* acima de 0,25. A escolha deste valor foi fundamentada no estudo feito por Hosmer & Lemeshow (1989) em que utilizou este valor como critério para escolha das variáveis a implementar na análise multivariada. De seguida, efetuou-se uma análise exploratória através de tabelas de frequências simples para todas as 49 variáveis, de forma a observar a distribuição dos seus dados em cada nível. Todas as variáveis que apresentavam uma diferença significativa na distribuição dos dados foram eliminadas da base de dados. Por fim, construiu-se tabelas cruzadas para cada variável com a variável dependente, "Obito", de forma a se analisar a correspondência de recém-nascidos sobrevivos e falecidos para cada variável. Para este ponto foi necessário do pacote *gmodels* (versão 2.18.1) presente no R para a construção de tabelas cruzadas. Posto isto, obteve-se o conjunto de 19 variáveis independentes, as possíveis candidatas a serem implementadas no modelo a desenvolver (Tabela XIII).

Tabela XIII - Sumário das características das variáveis independentes candidatas a incluir no modelo de regressão logística.

Variável	Tipo	Códigos/Níveis	Valor do coeficiente	Valor p *
IdadeGestacional	Quantitativa, proporcional, contínua	—	-0,052648	1,09e-11
NascimentoPeso	Quantitativa, proporcional, contínua	—	-0,0022376	5,06e-08
NascimentoComprimento	Quantitativa, proporcional, contínua	—	-0,23396	5,35e-11
NascimentoPerimetroCefalico	Quantitativa, proporcional, contínua	—	-0,31304	2,86e-10
CorticoidesPrenatais2 **CorticoidesPrenatais3**	Qualitativa e nominal	1 - Não 2 - Parcial 3 - Completo	-1,3516 -1,0060	0,00062 0,00113
PatologiasNaGravidez1	Qualitativa, binária e nominal	0 – Não 1 - Sim	-0,3150	0,222
TipoDeParto2	Qualitativa, binária e nominal	1 – Vaginal 2 - Cesariana	-0,5774	0,0228
MotivoDoParto2 **MotivoDoParto3** **MotivoDoParto4**	Qualitativa e nominal	1– Espontâneo 2 – Patologia materna 3 – Patologia Fetal 4 - IVG	2,4781 -0,5008 -0,4048	0,0146 0,1121 0,1998

Tabela XIII - Sumário das características das variáveis independentes candidatas a incluir no modelo de regressão logística. (Continuação)

Variável	Tipo	Códigos/Níveis	Valor do coeficiente	Valor p *
Sexo2	Qualitativa, binária e nominal	1 – Masculino 2 - Feminino	-1,3008	1,32e-05
ApgarMedia	Quantitativa, proporcional e discreta	—	-0,55041	3,26e-14
RessuscitacaoOxigenio1	Qualitativa, binária e nominal	0 – Não 1 - Sim	1,5967	1,1e-05
RessuscitacaoInsuflador1	Qualitativa, binária e nominal	0 – Não 1 - Sim	0,4940	0,049
RessuscitacaoEntubacaoEt1	Qualitativa, binária e nominal	0 – Não 1 - Sim	1,9004	2,15e-10
MalformacaoCongenitaMajor1	Qualitativa, binária e nominal	0 – Não 1 - Sim	1,2408	0,000885
SepsisMeningiteTardia1	Qualitativa, binária e nominal	0 – Não 1 - Sim	0,5237	0,0361
DiagSdr1	Qualitativa, binária e nominal	0 – Não 1 - Sim	1,4036	0,00119
DiagPda1	Qualitativa, binária e nominal	0 – Não 1 - Sim	1,2518	6,92e-07
DiagNec1	Qualitativa, binária e nominal	0 – Não 1 - Sim	2,0044	1,19e-11
PdaTerapeutico1	Qualitativa, binária e nominal	0 – Não 1 - Sim	0,6163	0,0395

* O valor p corresponde à estatística de Wald

Após esta seleção das variáveis candidatas a serem implementadas no modelo de regressão logística, procedeu-se à construção de possíveis modelos multivariados.

Ao longo desta secção será apresentado 12 modelos diferentes de regressão logística multivariada que poderão ser possíveis candidatos a classificador. Será explicado como se fez a escolha do conjunto de variáveis independentes a fazer parte de cada modelo e será analisado os seus resultados relativos à qualidade de ajuste e à capacidade de desempenho, de forma a escolher qual deles será o melhor modelo a ser utilizado como classificador.

Posto isto, começa-se por apresentar o primeiro modelo (Tabela XIV) que é consistido pelas 19 variáveis independentes anunciadas anteriormente.

Tabela XIV - *Output* obtido para o primeiro modelo de regressão logística.

Variável	Valor do coeficiente	Erro padrão	Z observado	*p-value*
Constante	11,3911244	3,6727236	3,102	0,001925**
IdadeGestacional	-0,0082194	0,0148131	-0,555	0,578983
NascimentoPeso	0,0005619	0,0010297	0,546	0,585255
NascimentoComprimento	-0,1587680	0,0927251	-1,712	0,086852
NascimentoPerimetroCefalico	-0,1161499	0,1325950	-0,876	0,381044
CorticoidesPrenatais2	-2,1568765	0,5108605	-4,222	2,42e-05**
CorticoidesPrenatais3	-0,8881270	0,4069440	-2,182	0,029078**
PatologiasNaGravidez1	-0,4025280	0,3684541	-1,092	0,274623
TipoDeParto2	-0,7531725	0,4226365	-1,782	0,074736
MotivoDoParto2	1,6963305	1,3014182	1,303	0,192422
MotivoDoParto3	0,4029854	0,4870497	0,827	0,408010
MotivoDoParto4	0,9824735	0,5110728	1,922	0,054559
Sexo2	-1,6774525	0,3764301	-4,456	8,34e-06**
ApgarMedia	-0,4178052	0,1117325	-3,739	0,000185**
RessuscitacaoOxigenio1	0,2647841	0,5554500	0,477	0,633574
RessuscitacaoInsuflador1	-0,7314086	0,3420346	-2,138	0,032484**
RessuscitacaoEntubacaoEt1	0,6610875	0,4982011	1,327	0,184526
MalformacaoCongenitaMajor1	1,6538317	0,4895943	3,378	0,000730**
SepsisMeningiteTardia1	-1,1714909	0,3724024	-3,146	0,001657**
DiagSdr1	0,2451241	0,5537571	0,443	0,658014
DiagPda1	0,8499118	0,3758656	2,261	0,023746**
DiagNec1	2,5868460	0,4323391	5,983	2,19e-09**
PdaTerapeutico1	-1,0507271	0,4529255	-2,320	0,020348**

** *P-values* menores que o nível de significância de 5%, ou seja, correspondem a variáveis do modelo que são estatisticamente significativas.

Analisando os valores de coeficientes estimados para cada variável presente na Tabela XIV, verifica-se que todos esses resultados eram espectáveis exceto o resultado obtido pela variável "SepsisMeningiteTardia". Segundo a Tabela XIV, um recém-nascido que apresentar Sepsis Meningite Tardia, apresenta uma maior probabilidade de vir a sobreviver, o que não corresponde à realidade.

Posto isto, construiu-se um segundo modelo de 12 variáveis, resultado da aplicação de três diferentes métodos de seleção de variáveis de *stepwise,* ao modelo anterior. Este modelo foi construído de forma a verificar se a variável "SepsisMeningiteTardia" seria eliminada do modelo e assim obter um modelo válido. A aplicação do método *stepwise* necessitou da utilização dos pacotes *tidyverse* (versão 1.3.0), *leaps* (versão 3.1) e *MASS* (versão 7.3-51.6) presentes no R. Os resultados obtidos pelos métodos *Forward, Backward* e *Both* foram os mesmos (Tabela XV).

Tabela XV - *Output* obtido para o segundo modelo de regressão logística usando o método *Forward*, *Backward* e *Both*.

Variável	Valor do coeficiente	Erro padrão	Z observado	*p-value*
Constante	9,28871	2,20835	4,206	2,60e-05
NascimentoComprimento	-0,20332	0,04932	-4,123	3,75e-05**
CorticoidesPrenatais2	-1,94592	0,48867	-3,982	6,83e-05**
CorticoidesPrenatais3	-0,80143	0,38977	-2,056	0,03977**
TipoDeParto2	-0,61509	0,31852	-1,931	0,05347
Sexo2	-1,55088	0,35696	-4,345	1,39e-05**
ApgarMedia	-0,44212	0,10702	-4,131	3,61e-05**
RessuscitacaoInsuflador1	-0,64242	0,32065	-2,003	0,04513**
RessuscitacaoEntubacaoEt1	0,78157	0,43644	1,791	0,07333
MalformacaoCongenitaMajor1	1,46555	0,46235	3,17	0,00153**
SepsisMeningiteTardia1	-1,03690	0,35595	-2,913	0,00358**
DiagPda1	0,91641	0,36666	2,499	0,01244**
DiagNec1	2,40534	0,41289	5,826	5,69e-09**
PdaTerapeutico1	-1,03266	0,44373	-2,327	0,01995**

** *P-values* menores que o nível de significância de 5%, ou seja, correspondem a variáveis do modelo que são estatisticamente significativas.

Tal como aconteceu no modelo 1 (Tabela XIV), também no segundo modelo a variável "SepsisMeningiteTardia" apresentou um valor de coeficiente estimado que não está de acordo com a realidade. Desta forma, quer o modelo 1 quer o modelo 2 não serão considerados como modelos candidatos para a construção do classificador.

Posto isto, construiu-se mais dois modelos idênticos aos anteriores, mas desta vez sem a variável "SepsisMeningiteTardia".

Na Tabela XVI encontra-se representado o resultado obtido pela modelação do terceiro modelo, constituído pelas 18 variáveis independentes anunciadas na Tabela XIII, exceto a variável "SepsisMeningiteTardia".

Tabela XVI - *Output* obtido para o terceiro modelo de regressão logística.

Variável	Valor do coeficiente	Erro padrão	Z observado	*p-value*
Constante	8,9267943	3,4953200	2,554	0,01065**
IdadeGestacional	-0,00044	0,014226	-0,031	0,97556
NascimentoPeso	0,000895	0,001026	0,872	0,38304
NascimentoComprimento	-0,14944	0,092615	-1,61e+00	0,10663
NascimentoPerimetroCefalico	-0,13233	0,127622	-1,037	0,29977
CorticoidesPrenatais2	-2,05907	0,506809	-4,063	4,85e-05**
CorticoidesPrenatais3	-0,76177	0,401633	-1,897	0,05787
PatologiasNaGravidez1	-0,37485	0,367669	-1,02	0,30795
TipoDeParto2	-0,63515	0,405303	-1,567	0,11709
MotivoDoParto2	1,817444	1,354219	1,342	0,17958
MotivoDoParto3	0,182797	0,472712	0,387	0,69898
MotivoDoParto4	0,643161	0,488025	1,318	0,18754
Sexo2	-1,57928	0,365079	-4,326	1,52e-05**
ApgarMedia	-0,39711	0,109573	-3,624	0,00029**
RessuscitacaoOxigenio1	0,250619	0,54058	0,464	0,64293
RessuscitacaoInsuflador1	-0,73127	0,333987	-2,19	0,02856**
RessuscitacaoEntubacaoEt1	0,710856	0,491702	1,446	0,14826
MalformacaoCongenitaMajor1	1,531858	0,483348	3,169	0,00153**
DiagSdr1	0,269762	0,542981	0,497	0,61932
DiagPda1	0,702319	0,369441	1,901	0,0573
DiagNec1	2,100831	0,382505	5,492	3,97e-08**
PdaTerapeutico1	-0,97012	0,445251	-2,179	0,02934**

** *P-values* menores que o nível de significância de 5%, ou seja, correspondem a variáveis do modelo que são estatisticamente significativas.

Por sua vez, na Tabela XVII apresenta-se o quarto modelo com 11 variáveis independentes, resultado da aplicação de três diferentes métodos de seleção de variáveis de *stepwise*, ao terceiro modelo. Os resultados obtidos pelos métodos *Forward, Backward* e *Both* foram os mesmos, deste modo, escolheu-se aleatoriamente um destes três modelos gerados pela aplicação dos três métodos de seleção de variáveis, de forma a dar continuidade ao estudo posterior a realizar.

Tabela XVII - *Output* obtido para o quarto modelo de regressão logística usando o método *Forward, Backward* e *Both*.

Variável	Valor do coeficiente	Erro padrão	Z observado	*p-value*
Constante	7,24071	2,02585	3,574	0,000351**
NascimentoComprimento	-0,161	0,04566	-3,526	0,000421**
CorticoidesPrenatais2	-1,94005	0,48597	-3,992	6,55e-05**
CorticoidesPrenatais3	-0,75353	0,38462	-1,959	5,01e-02
TipoDeParto2	-0,60556	0,31471	-1,924	0,054332
Sexo2	-1,47104	0,34658	-4,245	2,19e-05**
ApgarMedia	-0,41903	0,10492	-3,994	6,50e-05**
RessuscitacaoInsuflador1	-0,6651	0,31693	-2,099	3,59e-02**
RessuscitacaoEntubacaoEt1	0,79146	0,43131	1,835	0,0665
MalformacaoCongenitaMajor1	1,44081	0,454	3,174	0,001506**
DiagPda1	0,73846	0,3593	2,055	0,039853**
DiagNec1	1,98942	0,36639	5,43	5,64e-08**
PdaTerapeutico1	-0,92864	0,43357	-2,142	0,032206**

** *P-values* menores que o nível de significância de 5%, ou seja, correspondem a variáveis do modelo que são estatisticamente significativas.

Dado que, os modelos 3 e 4 ainda apresentam algumas variáveis que não são significativas e que o modelo final que se pretende implementar na aplicação deverá ser parcimonioso, não apresentado um grande número de variáveis para realizar o cálculo da previsão, construiu-se mais quatro modelos diferentes:

- **Modelo 5:** modelo formado pelas 8 variáveis independentes (Tabela XVIII) que foram indicadas no modelo 3 com *p-value* menor que 0,058.
- **Modelo 6:** modelo formado pelas 9 variáveis independentes (Tabela XIX) que foram indicadas no modelo 4 com *p-value* menor que o nível de significância usual de 5%.
- **Modelo 7:** modelo formado pelas 7 variáveis independentes (Tabela XX) que foram indicadas no modelo 3 com *p-value* menor que o nível de significância usual de 5%.
- **Modelo 8:** modelo formado pelas 9 variáveis independentes obtidas pela análise da variância (ANOVA) ao modelo 3 (Tabela XXI). Este conjunto de 9 variáveis, correspondem às variáveis que foram indicadas na análise ANOVA com *p-value* menor que 0,05. Os dados estatísticos relativos ao modelo 8 encontram-se representados na Tabela XXII.

Tabela XVIII - *Output* obtido para o quinto modelo de regressão logística.

Variável	Valor do coeficiente	Erro padrão	Z observado	*p-value*
Constante	2,44643	0,78132	3,131	0,001741**
CorticoidesPrenatais2	-1,78649	0,47919	-3,728	0,000193**
CorticoidesPrenatais3	-0,64714	0,36769	-1,76	0,07841
Sexo2	-1,39081	0,33809	-4,114	3,89e-05**
ApgarMedia	-0,58336	0,08867	-6,579	4,74e-11**
RessuscitacaoInsuflador1	-0,57168	0,31289	-1,827	6,77e-02
MalformacaoCongenitaMajor1	0,96109	0,45559	2,11	3,49e-02**
DiagPda1	1,12094	0,34256	3,272	0,001067**
DiagNec1	2,05581	0,35096	5,858	4,69e-09**
PdaTerapeutico1	-0,64374	0,41521	-1,55	1,21e-01

** *P-values* menores que o nível de significância de 5%, ou seja, correspondem a variáveis do modelo que são estatisticamente significativas.

Tabela XIX - *Output* obtido para o sexto modelo de regressão logística.

Variável	Valor do coeficiente	Erro padrão	Z observado	*p-value*
Constante	8,91917	1,73294	5,147	2,65e-07**
NascimentoComprimento	-0,18778	0,04404	-4,264	2,01e-05**
CorticoidesPrenatais2	-1,9768	0,48379	-4,086	4,39e-05**
CorticoidesPrenatais3	-0,93038	0,3793	-2,453	1,42e-02**
Sexo2	-1,44066	0,34398	-4,188	2,81e-05**
ApgarMedia	-0,50342	0,08997	-5,595	2,20e-08**
RessuscitacaoInsuflador1	-0,59197	0,31068	-1,905	5,67e-02
MalformacaoCongenitaMajor1	1,35752	0,4525	3,000	2,70e-03**
DiagPda1	0,83091	0,35366	2,349	1,88e-02**
DiagNec1	1,9081	0,35722	5,342	9,21e-08**
PdaTerapeutico1	-0,77163	0,42272	-1,825	0,0679

** *P-values* menores que o nível de significância de 5%, ou seja, correspondem a variáveis do modelo que são estatisticamente significativas.

Tabela XX - *Output* obtido para o sétimo modelo de regressão logística.

Variável	Valor do coeficiente	Erro padrão	Z observado	*p-value*
Constante	2,9238	0,7766	3,765	0,000167**

Tabela XX - *Output* obtido para o sétimo modelo de regressão logística. (Continuação)

Variável	Valor do coeficiente	Erro padrão	Z observado	*p-value*
CorticoidesPrenatais2	-1,8296	0,4772	-3,834	0,000126**
CorticoidesPrenatais3	-0,6875	0,3622	-1,898	5,77e-02
Sexo2	-1,4551	0,3355	-4,338	1,44e-05**
ApgarMedia	-0,6031	0,0888	-6,792	1,11e-11**
RessuscitacaoInsuflador1	-0,6087	0,3114	-1,954	5,07e-02
MalformacaoCongenitaMajor1	1,0516	0,4424	2,377	1,74e-02**
DiagNec1	2,1494	0,3435	6,257	3,92e-10**
PdaTerapeutico1	0,1511	0,349	0,433	6,65e-01

** *P-values* menores que o nível de significância de 5%, ou seja, correspondem a variáveis do modelo que são estatisticamente significativas.

Tabela XXI – Resultados da análise da variância (ANOVA) do modelo 3.

Variável	ANOVA
IdadeGestacional	8,22e-13**
NascimentoPeso	0,371943
NascimentoComprimento	0,005786**
NascimentoPerimetroCefalico	0,630744
CorticoidesPrenatais	0,000168**
PatologiasNaGravidez	0,628371
TipoDeParto	0,670532
MotivoDoParto	0,613212
Sexo	5,65e-06**
ApgarMedia	1,01e-06**
RessuscitacaoOxigenio	0,793471
RessuscitacaoInsuflador	0,045507**
RessuscitacaoEntubacaoEt	0,11414
MalformacaoCongenitaMajor	0,000547**
DiagSdr	0,595269

Tabela XXI – Resultados da análise da variância (ANOVA) do modelo 3. (Continuação)

Variável	ANOVA
DiagPda	0,10831
DiagNec	1,80e-07**
PdaTerapeutico	0,027772**

** *P-values* menores que o nível de significância de 5%, ou seja, correspondem a variáveis do modelo que são estatisticamente significativas.

Tabela XXII - *Output* obtido para o oitavo modelo de regressão logística.

Variável	Valor do coeficiente	Erro padrão	Z observado	*p-value*
Constante	1,14e+01	1,96956	5,811	6,20e-09**
IdadeGestacional	-0,01893	0,0115	-1,646	9,97e-02
NascimentoComprimento	-0,14922	0,05432	-2,747	6,01e-03**
CorticoidesPrenatais2	-2,05321	0,48519	-4,232	2,32e-05**
CorticoidesPrenatais3	-0,98091	0,37874	-2,59e+00	9,60e-03**
Sexo2	-1,38401	0,34466	-4,016	5,93e-05**
ApgarMedia	-0,48216	0,09177	-5,254	1,49e-07**
RessuscitacaoInsuflador1	-0,59988	0,30947	-1,938	5,26e-02
MalformacaoCongenitaMajor1	1,59e+00	0,45867	3,462	5,36e-04**
DiagNec1	1,93e+00	0,35301	5,481	4,22e-08**
PdaTerapeutico1	-0,33204	0,36461	-0,911	0,362472

** *P-values* menores que o nível de significância de 5%, ou seja, correspondem a variáveis do modelo que são estatisticamente significativas.

Por fim, optou-se por construir mais quatro modelos, resultados da aplicação do método *stepwise* nos modelos 5,6, 7 e 8, de forma a se verificar se as variáveis se mantinham. Neste caso implementou-se a técnica *both* do *stepwise*, pois ao contrário do *forward* e *backward,* que dariam importância e peso às variáveis, podendo não ser a ordem pelo qual se acharia o modelo ótimo, o *both* fará a pesquisa nas duas direções.

Na Tabela XXIII, Tabela XXIV, Tabela XXV e Tabela XXVI encontram-se representado os *outputs* dos modelos 9, 10, 11 e 12 obtidos pela aplicação do método *stepwise* segundo a direção de pesquisa *both* nos modelos 5,6, 7 e 8, respetivamente.

Tabela XXIII - *Output* obtido pelo nono modelo de regressão logística.

Variável	Valor do coeficiente	Erro padrão	Z observado	*p-value*
Constante	2,44643	0,78132	3,131	0,001741**
CorticoidesPrenatais2	-1,78649	0,47919	-3,728	0,000193**
CorticoidesPrenatais3	-0,64714	0,36769	-1,76	0,07841
Sexo2	-1,39081	0,33809	-4,114	3,89e-05**
ApgarMedia	-0,58336	0,08867	-6,579	4,74e-11**
RessuscitacaoInsuflador1	-0,57168	0,31289	-1,827	0,067687
MalformacaoCongenitaMajor1	0,96109	0,45559	2,11	0,034898**
DiagPda1	1,12094	0,34256	3,272	0,001067**
DiagNec1	2,05581	0,35096	5,858	4,69e-09**
PdaTerapeutico1	-0,64374	0,41521	-1,55	0,121052

** *P-values* menores que o nível de significância de 5%, ou seja, correspondem a variáveis do modelo que são estatisticamente significativas.

Tabela XXIV - *Output* obtido pelo décimo modelo de regressão logística.

Variável	Valor do coeficiente	Erro padrão	Z observado	*p-value*
Constante	8,91917	1,73294	5,147	2,65e-07**
NascimentoComprimento	-0,18778	0,04404	-4,264	2,01e-05**
CorticoidesPrenatais2	-1,9768	0,48379	-4,086	4,39e-05**
CorticoidesPrenatais3	-0,93038	0,3793	-2,453	1,42e-02**
Sexo2	-1,44066	0,34398	-4,188	2,81e-05**
ApgarMedia	-0,50342	0,08997	-5,595	2,20e-08**
RessuscitacaoInsuflador1	-0,59197	0,31068	-1,905	0,0567
MalformacaoCongenitaMajor1	1,35752	0,4525	3,000	2,70e-03**
DiagPda1	0,83091	0,35366	2,349	0,0188**
DiagNec1	1,9081	0,35722	5,342	9,21e-08**
PdaTerapeutico1	-0,77163	0,42272	-1,825	0,0679

** *P-values* menores que o nível de significância de 5%, ou seja, correspondem a variáveis do modelo que são estatisticamente significativas.

Tabela XXV - *Output* obtido pelo décimo primeiro modelo de regressão logística.

Variável	Valor do coeficiente	Erro padrão	Z observado	*p-value*
Constante	2,9478	0,77525	3,802	0,000143**

Tabela XXV - *Output* obtido pelo décimo primeiro modelo de regressão logística. (Continuação)

Variável	Valor do coeficiente	Erro padrão	Z observado	*p-value*
CorticoidesPrenatais2	-1,80887	0,47472	-3,810	0,000139**
CorticoidesPrenatais3	-0,67206	0,36078	-1,863	0,062489
Sexo2	-1,45301	0,33506	-4,337	1,45e-05**
ApgarMedia	-0,60593	0,08856	-6,842	7,82e-12**
RessuscitacaoInsuflador1	-0,5933	0,30923	-1,919	5,50e-02
MalformacaoCongenitaMajor1	1,04581	0,4421	2,366	1,80e-02**
DiagNec1	2,17536	0,3383	6,43	1,27e-10**

** *P-values* menores que o nível de significância de 5%, ou seja, correspondem a variáveis do modelo que são estatisticamente significativas.

Tabela XXVI - *Output* obtido pelo décimo segundo modelo de regressão logística.

Variável	Valor do coeficiente	Erro padrão	Z observado	*p-value*
Constante	10,94559	1,88231	5,815	6,06e-09**
IdadeGestacional	-0,01691	0,01123	-1,505	1,32e-01
NascimentoComprimento	-0,14739	0,05426	-2,717	6,60e-03**
CorticoidesPrenatais2	-2,06893	0,48313	-4,282	1,85e-05**
CorticoidesPrenatais3	-0,99442	0,37728	-2,636	8,40e-03**
Sexo2	-1,40663	0,34519	-4,075	4,60e-05**
ApgarMedia	-0,48162	0,0918	-5,247	1,55e-07**
RessuscitacaoInsuflador1	-0,62937	0,30778	-2,045	0,040865**
MalformacaoCongenitaMajor1	1,57835	0,45824	3,444	5,72e-04**
DiagNec1	1,89133	0,34896	5,420	5,96e-08**

** *P-values* menores que o nível de significância de 5%, ou seja, correspondem a variáveis do modelo que são estatisticamente significativas.

Daqui, verificou-se que somente os modelos 9 e 10 mantiveram as mesmas variáveis ao se ter aplicado a técnica de *stepwise* aos modelos 5 e 6, respetivamente. Em contrapartida a aplicação desta técnica ao modelo 7 originou um modelo 11 com menos uma variável, "PdaTerapeutico1", e o modelo 8 originou o modelo 12 com menos uma variável, "PdaTerapeutico".

Seguidamente, efetuou-se um estudo comparativo no que diz respeito à qualidade dos modelos, assim como, às suas capacidades preditivas, de modo a escolher o modelo que melhor se ajusta aos dados e que consiga realizar uma melhor previsão do estado de admissão de um recém-nascido de muito baixo

peso. Na Tabela XXVII, mostra-se os valores de algumas características usualmente consideradas para avaliar a qualidade dos modelos a nível de ajuste e a capacidade preditiva, assim como, o número de variáveis utilizadas para cada modelo.

Para que fosse possível obter os valores para avaliar a qualidade de ajuste e a capacidade preditiva dos modelos construídos, recorreu-se a pacotes como, *ROCR* (versão 1.0-11), *broom* (versão 0.5.6), *ResourceSelection* (versão 0.3-5), *modEvA* (versão 2.0) e *InformationValue* (versão 1.2.3).

Olhando para a Tabela XXVII, em termos de indicador AIC, o modelo 4 é o que apresenta melhores resultados, pois quanto menor for o valor de AIC, melhor o modelo se ajusta aos dados. Relativamente ao indicador BIC, o modelo 12 é o que apresenta melhores resultados, pois tal como acontece para o AIC, quanto menor for o seu valor, melhor o modelo se ajusta aos dados.

Outra medida como o pseudo R^2 também explica o ajuste geral do modelo proposto, sendo que, o modelo com o maior R^2 é considerado o "melhor". Algumas dessas medidas mais utilizadas foram estipuladas por Cox e Snell, Nagelkerke e McFadden. Para este caso, o modelo 3 apresenta-se como o melhor modelo segundo estas três medidas, apresentando valores de 0,346, 0,152 e 0,401 para McFadden R^2, Cox e Snell R^2 e Nagelkerke R^2, respetivamente. Para além disso, todos os modelos apresentam um bom ajuste para a predição.

Relativamente às características que avaliam as capacidades preditivas dos modelos, verifica-se que todos eles apresentam um bom poder discriminante. Apesar dos valores obtidos serem semelhantes, os modelos 3 e 8 apresentam um valor de precisão relativamente superior aos restantes.

No que diz respeito à área abaixo da curva ROC, os modelos apresentam valores entre os 70% e 80% para a validação externa e valores próximos dos 90% para a validação interna, o que indica que todos os modelos são eficientes na distinção entre recém-nascidos de muito baixo peso sobrevivos e falecidos. O modelo 3 é o que apresenta o melhor valor de AUC para a validação interna e o modelo 4 o que apresenta o melhor valor de AUC para a validação externa. Para além disso, nos dez modelos o valor de AUC na validação interna apresentou um melhor valor do que na validação externa, o que era espectável. Isto deve-se ao facto de o conjunto de valores de teste, utilizados na validação externa, ser ligeiramente abaixo aos dos de treino (validação interna) e por a precisão que se calcula através da AUC para o conjunto de treino ser sempre subestimado em relação ao conjunto de teste.

Na Figura 16, Figura 17, Figura 18, Figura 19, Figura 20, Figura 21, Figura 22, Figura 23, Figura 24 e Figura 25 encontram-se representado as curvas ROC obtidas nos processos de validação interna (A) e externa (B) dos modelos 3, 4, 5, 6, 7, 8, 9, 10, 11 e 12, respetivamente. As curvas coloridas são constituídas por um conjunto de *cutoff*.

Tabela XXVII - Medidas de qualidade do ajustamento e de capacidade preditiva dos modelos de regressão logística 3 - 12.

	Modelo 3	Modelo 4	Modelo 5	Modelo 6	Modelo 7	Modelo 8	Modelo 9	Modelo 10	Modelo 11	Modelo 12
Número de variáveis	18 (Tabela XVI)	11 (Tabela XVII)	8 (Tabela XVIII)	9 (Tabela XIX)	7 (Tabela XX)	9 (Tabela XXII)	8 (Tabela XXIII)	9 (Tabela XXIV)	6 (Tabela XXV)	8 (Tabela XXVI)
Qualidade de ajuste										
AIC	378,958	367,477	387,995	371,017	396,048	373,498	387,995	371,017	394,233	372,352
BIC	488,520	432,218	437,795	425,798	440,869	428,279	437,795	425,798	434,073	422,152
McFadden	0,346	0,334	0,282	0,319	0,262	0,314	0,282	0,319	0,262	0,312
Cox e Snell	0,152	0,147	0,128	0,141	0,117	0,139	0,126	0,141	0,117	0,138
Nagelkerke	0,401	0,388	0,332	0,372	0,310	0,367	0,332	0,372	0,309	0,365
Capacidade preditiva										
Precisão	0,926	0,924	0,924	0,924	0,924	0,926	0,924	0,924	0,924	0,924
Especificidade	0,500	0,429	0,000	0,400	0,000	0,500	0,000	0,500	0,000	0,400
Sensibilidade	0,932	0,932	0,926	0,934	0,926	0,930	0,926	0,930	0,926	0,930
AUC										
Validação interna	0,903	0,901	0,878	0,895	0,870	0,891	0,878	0,895	0,868	0,893
Validação externa	0,790	0,799	0,718	0,790	0,695	0,797	0,718	0,790	0,698	0,793

Os melhores valores de *cutoff*, quando realizada a validação externa, foram 0,538, 0,432, 0,666, 0,456, 0,709, 0,626, 0,666, 0,456, 0,718 e 0,523 para os modelos 3, 4, 5, 6, 7, 8, 9, 10, 11 e 12, respetivamente. Estes valores são gerados pela conjugação dos melhores valores de sensibilidade e especificidade, e para os obter, recorreu-se à função *optimalCutoff* presente no R.

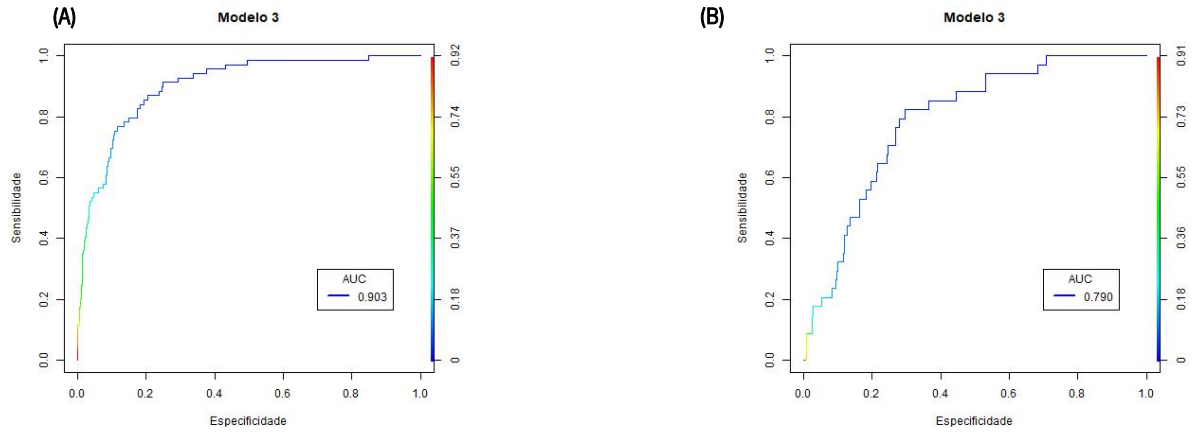

Figura 16 - Curvas ROC do modelo 3 obtidas na validação interna (A) e na validação externa (B).

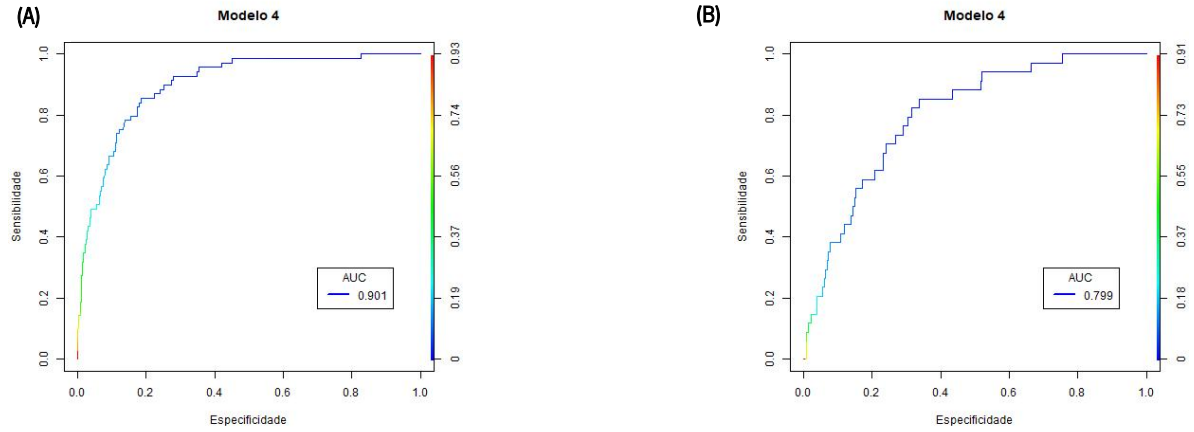

Figura 17 - Curvas ROC do modelo 4 obtidas na validação interna (A) e na validação externa (B).

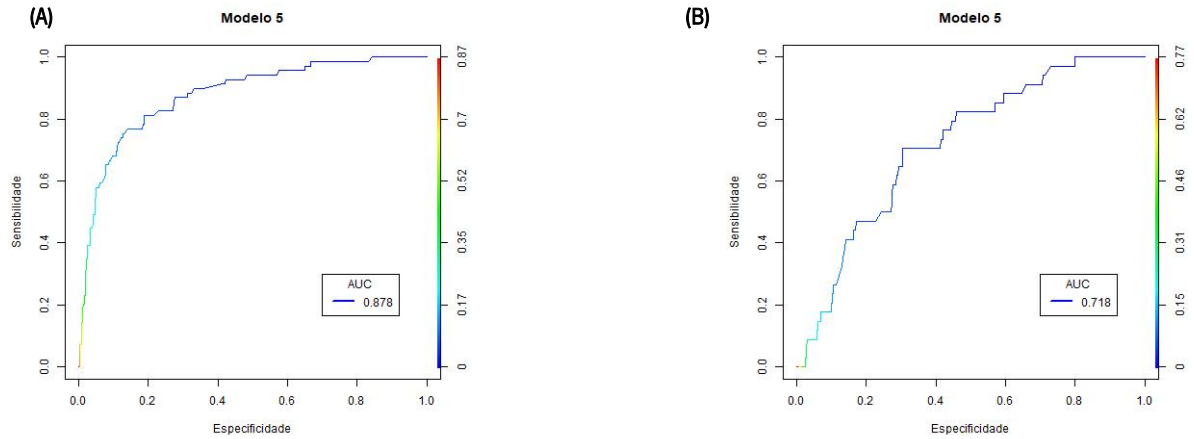

Figura 18 - Curvas ROC do modelo 5 obtidas pela validação interna (A) e pela validação externa (B).

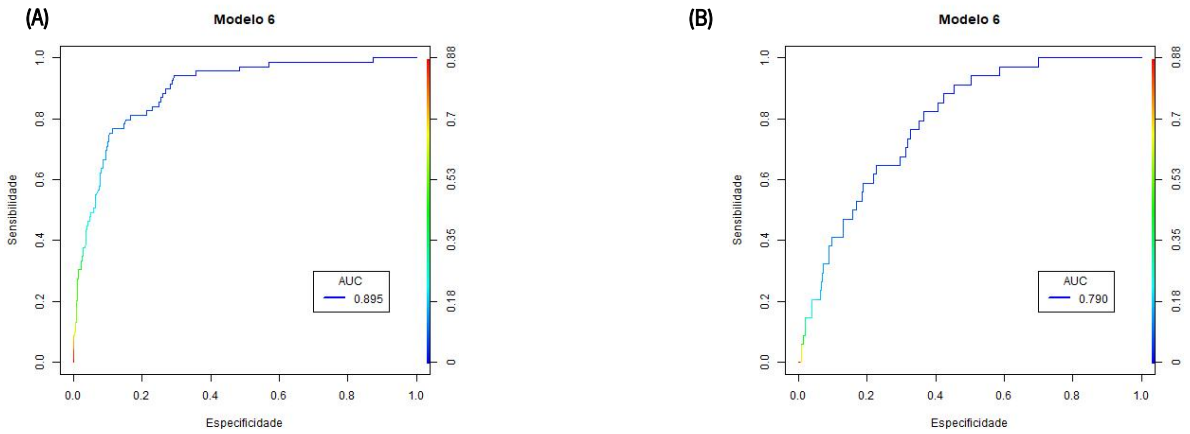

Figura 19 - Curvas ROC do modelo 6 obtidas na validação interna (A) e na validação externa (B).

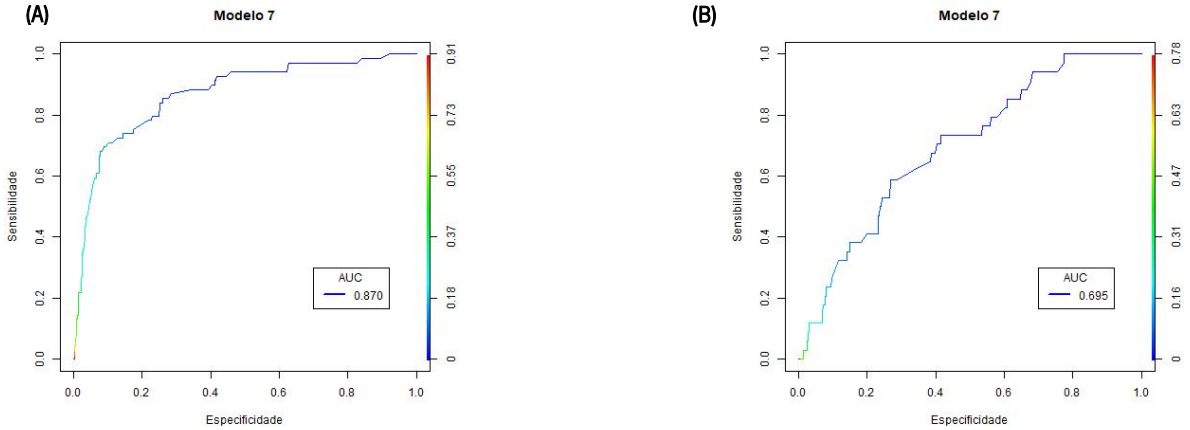

Figura 20 - Curvas ROC do modelo 7 obtidas na validação interna (A) e na validação externa (B).

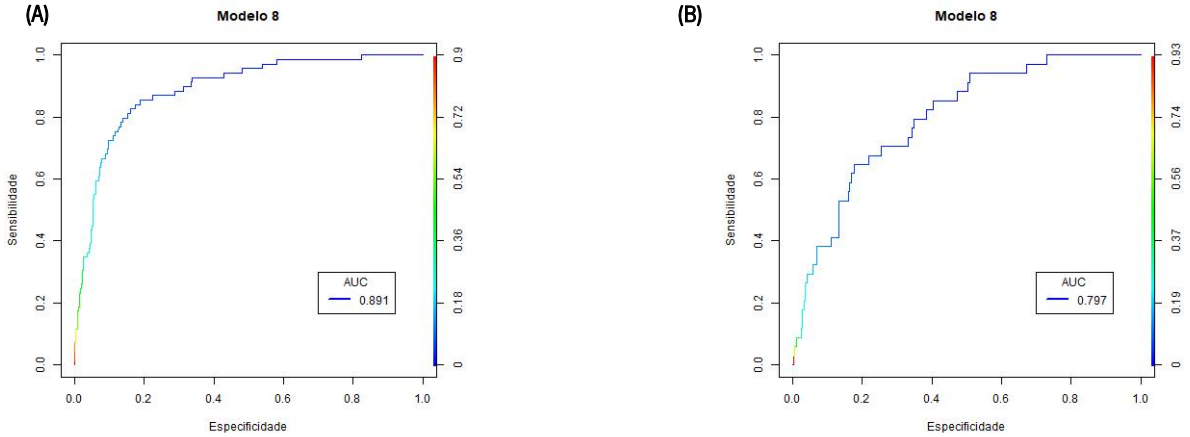

Figura 21 - Curvas ROC do modelo 8 obtidas na validação interna (A) e na validação externa (B).

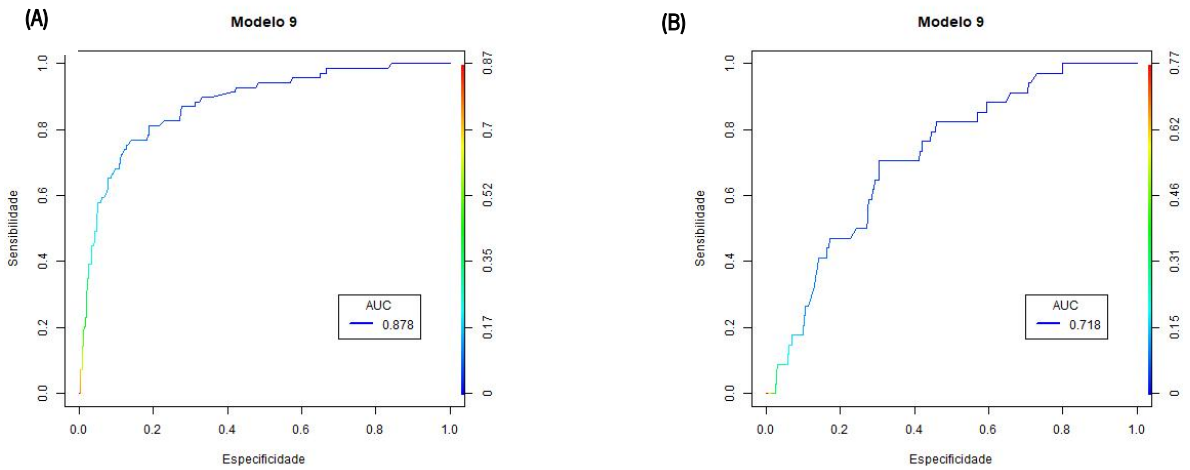

Figura 22 - Curvas ROC do modelo 9 obtidas na validação interna (A) e na validação externa (B).

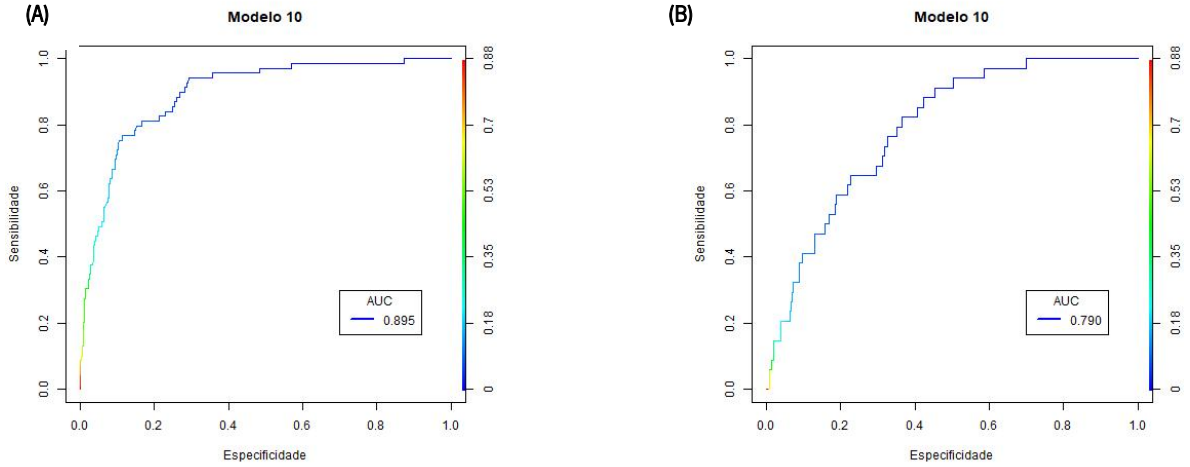

Figura 23 - Curvas ROC do modelo 10 obtidas na validação interna (A) e na validação externa (B).

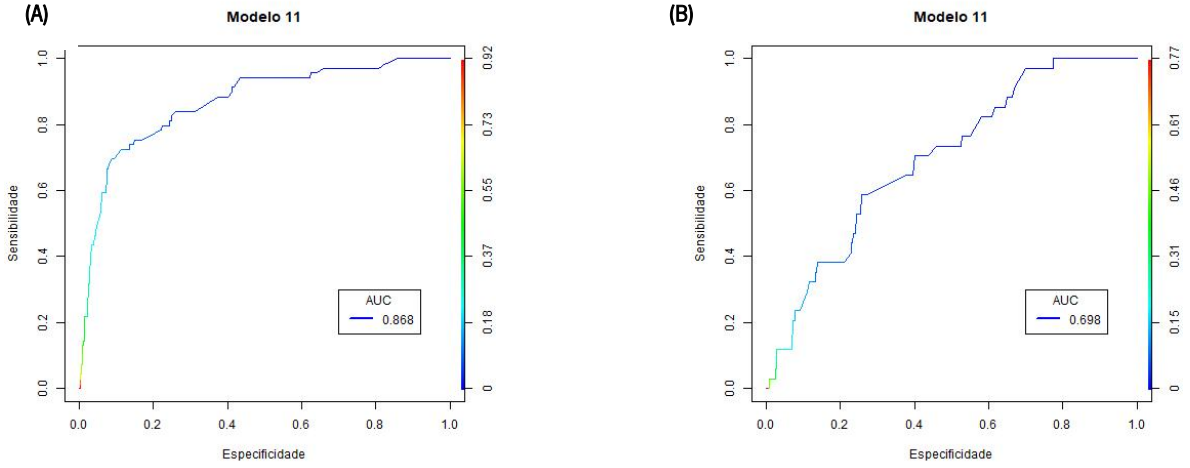

Figura 24 - Curvas ROC do modelo 11 obtidas na validação interna (A) e na validação externa (B).

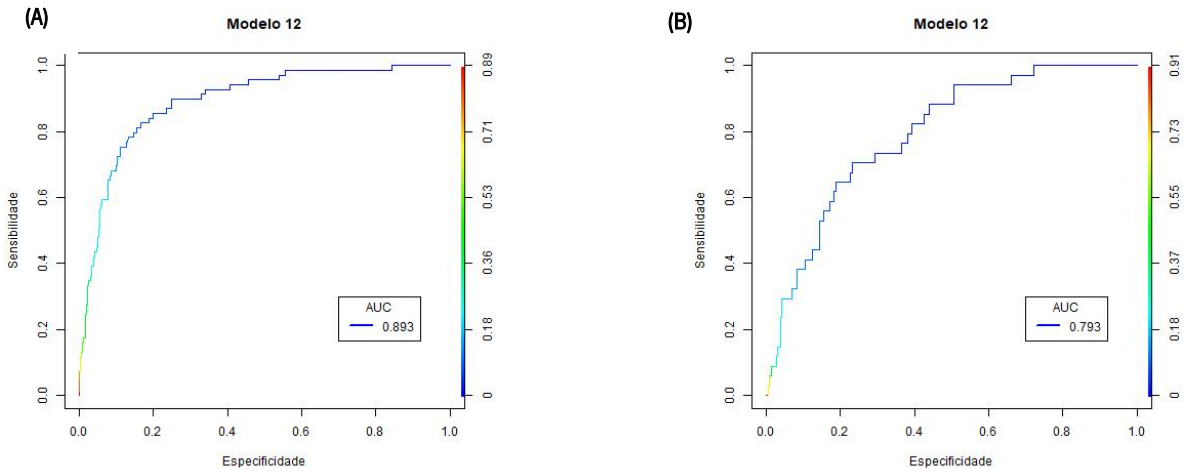

Figura 25 - Curvas ROC do modelo 12 obtidas na validação interna (A) e na validação externa (B).

Tendo como intuito implementar na aplicação *web* um modelo que seja capaz de predizer com eficácia o risco de morte de um recém-nascido, segundo um conjunto de variáveis que possam ou não ser medidas rotineiramente nas unidades hospitalares, escolheu-se o modelo 8 para dar continuidade ao trabalho. Uma vez que, todos os modelos construídos se ajustaram relativamente bem aos dados em estudo e apresentaram uma boa capacidade preditiva, a escolha do modelo final a ser implementado foi feita tendo em conta o seu valor de AUC, uma vez que também se realizou a validação externa, o número de variáveis pelo qual é constituído e o seu valor de precisão, sendo que se deu preferência a um modelo com um número mais reduzido de variáveis, com um melhor valor de precisão e com um dos melhores valores de AUC.

Comparando os resultados obtidos neste trabalho com outros estudos, verifica-se que é possível encontrar-se algumas semelhanças. Relativamente às variáveis que se apresentam como fortes candidatas a fatores associados ao risco de morte desses recém-nascidos, por exemplo, estudos liderados por de Vonderweid et al. (1991) indicam que variáveis como a idade gestacional, peso ao nascer, sexo e índice apgar são alguns exemplos de fatores que influenciam na mortalidade destes bebés. Por outro lado, Gera & Ramji (2001) concluíram no seu estudo que a necessidade de ventilação mecânica também se apresenta como um preditor significativo de mortalidade neonatal. Por sua vez, Ambalavanan & Carlo (2001) verificaram que o uso de corticoides pré-natais também é considerado como um possível fator de mortalidade, sendo que, por exemplo, a idade da mãe e tipo de parto não são significativos. Já Medlock et al. (2011), que realizaram uma revisão a 41 estudos que relatam o desenvolvimento de um modelo de previsão de mortalidade em recém-nascidos de muito baixo peso, declararam que, apesar da

variável malformação congénita ser um critério de exclusão no que diz respeito às variáveis de entrada em muitos estudos, quando testada é frequentemente significativa.

6.6 Identificação de pontos influentes e pontos mal ajustados

Após se ter selecionado o modelo a dar continuidade ao trabalho e antes de o implementar na aplicação, sentiu-se a necessidade de realizar uma análise mais profunda à qualidade do ajuste do modelo, sendo esta feita através de uma análise aos resíduos. Esta análise visa identificar a diferença existente entre os valores observados e os valores obtidos, sendo que, quanto maior for essa diferença pior será o ajustamento das observações, e consequentemente, pior será o desempenho do modelo de regressão logística.

Para além da análise aos resíduos realizada no subcapítulo anterior através de medidas de diagnóstico como o AIC e BIC, existem outras formas de se avaliar a validade do modelo de regressão logística, como por exemplo, através da verificação da existência de pontos influentes e/ou mal ajustados.

Uma das formas para se identificar pontos de influência nos modelos passa por se analisar gráficos de *Residuals vs Leverage*, sendo que, valores que se encontrarem fora da linha tracejada de distância de Cook, representada a vermelho, correspondem a pontos influentes. Segundo a Figura 26, o modelo a ser implementado na aplicação *web*, ou seja, o modelo 8, não aparenta apresentar pontos influentes, visto que todos os casos se encontram dentro da linha de distância de Cook.

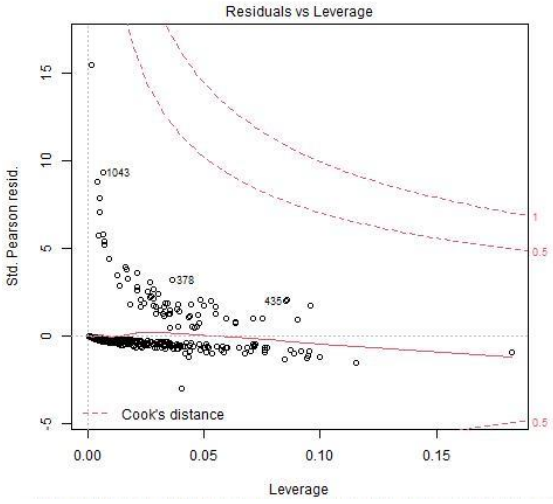

Figura 26 - Representação gráfica dos resíduos versos o *leverage* do modelo 8.

Contudo, ao aplicar a função *influencePlot* do pacote *car* (versão 3.0-9) presente no R, é possível gerar-se um gráfico de resíduos padronizados *versus leverage* mais pormenorizado, que salienta as observações atípicas com bolhas de diferentes tamanhos, sendo que, o tamanho dessas bolhas são

proporcionais à distância de Cook (Figura 27). Segundo esta figura, verifica-se que é possível haver algumas observações que geram maiores mudanças nos resíduos estandardizados, quando em relação ao *leverage*.

Para além disso, como complemento desse gráfico, também se obteve os valores dos resíduos, os valores da diagonal da matriz chapéu e distância de Cook das observações que foram identificadas como possíveis pontos de influência (Tabela XXVIII).

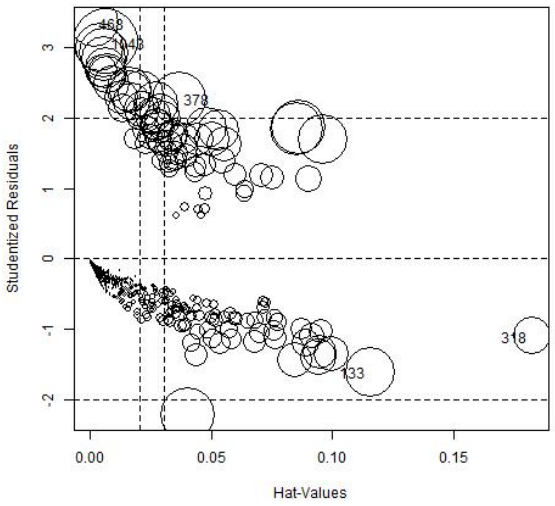

Figura 27- Representação gráfica dos resíduos versos o *leverage* do modelo 8, assim como, os possíveis pontos influentes.

Tabela XXVIII - Valores dos resíduos, da diagonal da matriz chapéu e da distância de Cook das observações que foram identificadas como possíveis pontos de influência.

Observação	Resíduo estandardizado	Diagonal da matriz chapéu	Distância de Cook
133	-1,609629	0,115404063	0,02909973
318	-1,094066	0,182296654	0,01699092
378	2,272172	0,036498019	0,03554931
468	3,359980	0,001334815	0,02918960
1043	3,082172	0,006316832	0,05043886

Como uma observação influente pode afetar o ajuste do modelo em outras observações, será melhor remover estes 5 possíveis pontos de influência e avaliar novamente o desempenho do modelo (Tabela XXIX) de forma a verificar se existem melhorias no desempenho do modelo.

Tabela XXIX - Medidas de qualidade do ajustamento e de capacidade preditiva do modelo 8 sem os possíveis pontos de influência.

	Modelo 8
AIC	343,157
BIC	397,886
McFadden	0,352
Cox e Snell	0,150
Nagelkerke	0,406
Precisão	0,924
Especificidade	0,400
Sensibilidade	0,930
AUC	
Validação interna	0,908
Validação externa	0,786

Apesar de, o modelo sem esses supostos cinco pontos influentes apresentar uma ligeira melhoria nos resultados das medidas que avaliam a qualidade de ajuste, no que diz respeito às medidas que avaliam o desempenho do modelo, não apresentaram uma grande alteração, sendo que o modelo original com esses cinco pontos continua a apresentar uma melhor precisão. Desta forma, decidiu-se que o modelo a implementar na aplicação *Shiny* será o modelo 8 original.

6.7 Interpretação dos coeficientes estimados

Uma vez ajustado o modelo final que funcionará como um classificador e ter-se avaliado a significância dos seus coeficientes estimados, será necessário agora interpretar os valores destes coeficientes.

O modelo de regressão logística, apresenta os seus resultados dos estimadores na forma logarítmica, o que torna a sua interpretação mais complicada. Assim, no sentido de facilitar a interpretação no que diz respeito à relação de cada variável presente no modelo final com a variável dependente (Obito), realizou-se uma transformação destes coeficientes, através da exponenciação das variáveis, dando origem aos *Odds Ratio* (OR). Para isso, utilizou-se a função *logitor* do pacote *mfx* (versão 1.2-2) presente no RStudio. Na Tabela XXX encontra-se representado os valores dos coeficientes estimados na forma logarítmica de todas as variáveis incluídas no modelo 8 (modelo final), assim como, os seus *odds ratio*, resultantes da transformação dos mesmos.

Tabela XXX - Interpretação dos coeficientes estimados das variáveis independentes do modelo final (modelo 8) segundo os *odds ratio*.

Variável	$\widehat{\beta}$	OR
IdadeGestacional	-0,01893	0,981245
NascimentoComprimento	-0,14922	0,861376
CorticoidesPrenatais2	-2,05321	0,128322
CorticoidesPrenatais3	-0,98091	0,374968
Sexo2	-1,38401	0,250572
ApgarMedia	-0,48216	0,61745
RessuscitacaoInsuflador1	-0,59988	0,548877
MalformacaoCongenitaMajor1	1,59e+00	4,893734
DiagNec1	1,93e+00	6,92357
PdaTerapeutico1	-0,33204	0,717461

O resultado acima evidencia que para uma alteração de uma unidade na variável IdadeGestacional, a chance de que a variável Obito tome valor de 1 diminui em 1,88%, ou seja, a chance de um recém-nascido vir a falecer é 0,98 vezes menor quando a idade gestacional aumenta em uma unidade, sendo que, aqui as demais variáveis independentes se mantêm constantes. Por outro lado, a probabilidade da variável dependente tomar valor de 1 com o aumento de uma unidade nas variáveis NascimentoComprimento e ApgarMedia, diminui em 13,86% e 38,26%, respetivamente, ou seja, é 0,86 e 0,62 vezes menos provável falecer quando existe um aumento de uma unidade no comprimento ao nascer e no índice de apgar, respetivamente.

Relativamente às variáveis categóricas, a leitura é feita de outra forma. Assim, no caso da variável CorticoidesPrenatais, como se trata de uma variável categórica, pois retoma as categorias de 1 a 3, as comparações das chances de que um recém-nascido possa vir a falecer são comparadas com a variável CorticoidesPrenatais1. Desta forma, caso o nascimento do recém-nascido ocorreu menos de 24 horas após a 1°dose de corticoide, ou mais de uma semana após a última dose de corticoides administradas por parte da progenitora (CorticoidesPrenatais2), diminuem-se as chances em 87,17% de que o recém-nascido venha a falecer. Já para recém-nascidos cujos nascimentos ocorreram mais de 24 horas e menos de uma semana, após pelo menos uma dose de corticoides administrada pela progenitora (CorticoidesPrenatais3) têm 62,50% menos chances de virem a falecer. Analogamente ao sexo dos bebés, os recém-nascidos do sexo feminino apresentam 74,94% menos probabilidade de virem a falecer. Recém-nascidos que tenham recebido qualquer tipo de pressão positiva por uma máscara e insuflador (RessuscitacaoInsuflador1) apresentam 45,11% menos chances de virem a falecer, assim como, aqueles

que tomaram ibuprofeno após o seu nascimento para o tratamento de persistência de ductos arteriosos (PDA) apresentam 28,25% menos chances de virem a falecer. Por fim, os recém-nascidos que apresentaram malformações congénitas major têm 4,89 vezes maior chance de virem a morrer e aqueles que realizaram diagnósticos que detetaram a presença de enterocolite necrotizante (NEC) têm 6,92 vezes maior probabilidade de virem a falecer.

6.8 Cálculo das previsões

Após a construção do modelo com um dos melhores poderes preditivos, um dos aspetos mais importantes a ser efetuado é testar o mesmo modelo com dados que não se encontram presentes na base de dados utilizada para a modelação do modelo, ou seja, com os dados teste. Neste caso, o aspeto mais importante a ser avaliado será a qualidade das previsões obtidas pelo modelo de regressão logística a implementar na aplicação *web*.

Posto isto, calculou-se as previsões do risco de morte para sete recém-nascidos de muito baixo peso, de modo a testar o poder preditivo do modelo escolhido. Neste caso, considerou-se como valor de *cutoff* de 0,626, sendo que, recém-nascidos que apresentam valores de previsão menores e iguais a 0,626 são considerados como bebés que têm probabilidade de virem a sobreviver (0), e recém-nascidos com um valor de previsão superior a 0,626 são considerados como indivíduos que poderão vir a falecer (1).

Na Tabela XXXI encontra-se destacado as características das variáveis, que constituem o modelo, de sete RN escolhidos aleatoriamente e na Tabela XXXII encontra-se representado os resultados de previsão do estado de admissão de cada recém-nascido e os seus valores reais.

Tabela XXXI - Representação das características dos sete recém-nascidos escolhidos aleatoriamente para testar o poder preditivo do modelo que funcionará como um classificador.

Variável	RN 1	RN 2	RN 3	RN 4	RN 5	RN 6	RN 7
IdadeGestacional	213,0	191,0	224,0	168,0	187,0	224,0	200,0
NascimentoComprimento	42,0	37,0	39,0	28,0	33,5	36,5	31,0
CorticoidesPrenatais	3,0	3,0	3,0	3,0	2,0	2,0	3,0
Sexo	1,0	1,0	2,0	1,0	2,0	1,0	1,0
ApgarMedia	9,0	10,0	10,0	8,0	5,0	5,0	6,0
RessuscitacaoInsuflador	1,0	1,0	0,0	1,0	1,0	1,0	0,0
MalformacaoCongenitaMajor	0,0	0,0	0,0	0,0	0,0	1,0	1,0
DiagNec	0,0	1,0	0,0	1,0	0,0	0,0	0,0
PdaTerapeutico1	0,0	0,0	0,0	0,0	0,0	0,0	0,0

Tabela XXXII - Representação dos resultados de previsão do estado de admissão de cada recém-nascido e os seus valores reais.

Recém-Nascido	Valor Real (Obito)	Valor Previsto em probabilidade (Obito)
1	0	0,008
2	0	0,103
3	0	0,003
4	1	0.642
5	1	0,028
6	0	0,152
7	1	0,679

Este conjunto de teste utilizado, apresenta uma precisão de 0,926. Contudo, segundo a Tabela XXXII, destes sete recém-nascidos escolhidos aleatoriamente, o modelo foi capaz de prever corretamente o estado de admissão de seis bebés. É de realçar que, este resultado poderá estar dependente da divisão manual dos dados que se realizou anteriormente.

6.9 Comparação dos indicadores CRIB e SNAPPE II com o classificador desenvolvido

Neste subcapítulo será feito uma análise comparativa entre o classificador desenvolvido neste projeto e os indicadores de mortalidade CRIB e SNAPPE II, de forma a avaliar se de facto o modelo desenvolvido poderá ser uma ferramenta alternativa a utilizar nas unidades hospitalares. Neste caso, foi feito um estudo de avaliação discriminativa que teve como prepósito estimar a capacidade preditiva de cada indicador em predizer entre dois estados, sobrevivência ou falecimento, dos recém-nascido de muito baixo peso. Para tal, fez-se comparações de curvas ROC e dos seus valores de AUC, tendo-se recorrido ao pacote *caTools* (versão 1.18.0) presente no RStudio.

Dado que, na amostra apresentada anteriormente nem todos os recém-nascidos continham simultaneamente informações referentes ao índice CRIB e SNAPPE II, para esta parte, selecionou-se um novo conjunto de dados teste no qual integra somente os recém-nascidos que preenchem estes critérios. Neste caso, fez-se uso de uma amostra de 350 registos.

Na Figura 28 encontra-se representado as três curvas ROC relativas aos indicadores CRIB e SNAPPE II, assim como, o classificador desenvolvido neste trabalho (Modelo_8).

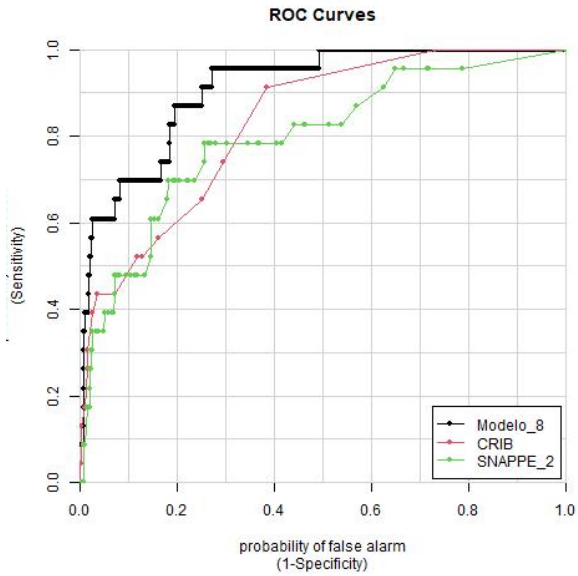

Figura 28 - Representação gráfica das curvas ROC dos três indicadores em estudo, obtidas através do *caTools*.

Para além do gráfico presente na figura anterior, ao utilizar o pacote *caTools* com recurso à função *colAUC*, foi igualmente possível obter os valores dos seus AUC que foram 0,909, 0,830 e 0,794 para o classificador construído, indicador CRIB e indicador SNAPPE II, respetivamente. Pela análise das curvas ROC e dos valores AUC, pode-se concluir que o classificador desenvolvido apresenta uma melhor capacidade preditiva que os restantes indicadores.

6.10 Aplicação *web*

Uma vez confirmado de que o modelo desenvolvido apresenta uma boa capacidade preditiva para o estado de admissão de um recém-nascido de muito baixo peso, o modelo 8, que funcionará como um classificador, esteve na base da construção de uma aplicação *web*. A ideia da construção desta aplicação surgiu de forma a facilitar a estimação, em tempo real, da probabilidade de risco de morte nestes recém-nascidos, ajudando assim os profissionais de saúde a tomarem as melhores decisões no planeamento da assistência pré e pós-natal, podendo assim atuarem atempadamente. A aplicação desenvolvida nesta dissertação pode ser encontrada em https://claudia-rodrigues.shinyapps.io/Previsao_do_risco_de_morte_em_RNMBP/ .

Para este projeto, optou-se por desenvolver a aplicação utilizando o pacote *shiny* (versão 1.5.0), presente no R, uma vez que, a construção de aplicações da *web* interativos neste ambiente é bastante simples,

sendo que, não são necessários conhecimentos prévios de HTML, CSS e JavaScript, linguagens muito utilizadas na construção de aplicações *web*.

Desta forma, começou-se por criar dois R *scripts*, o ui.R que corresponde ao objeto *user interface* e o server.R, relativo à função server, sendo que ambos se encontram no mesmo ficheiro. Como já foi mencionado em capítulos anteriores, o objeto *user interface* é o conjunto de código que define a estética da aplicação e que é responsável por exibir ao utilizador os *inputs* para realizar uma determinada tarefa. Por sua vez, a função server corresponde ao segmento de código que decide como os *inputs* introduzidos pelo utilizador serão utilizados e como o resultado será apresentado no *output*, ou seja, em outras palavras, será o responsável pelos cálculos que a aplicação irá realizar.

Apesar de, por padrão o *shiny* fazer uso do *framework* Bootstrap, o que contribui para que os seus aplicativos apresentem um *interface* limpo e com um design próprio, para este trabalho preferiu-se fazer-se uso do pacote *shinythemes* (versão 1.1.2), assim como, de CSS de forma a personalizar a aplicação.

A aplicação *Shiny* desenvolvida para este trabalho conta com um ambiente organizado em 3 abas, cada uma com sua finalidade. As funcionalidades de cada secção da aplicação serão explicadas nos próximos subcapítulos.

6.10.1 Página Inicial

Tal como explicado acima, e como se pode observar pela Figura 29, a aplicação divide-se em 3 abas diferentes, sendo que cada aba tem a sua finalidade. O acesso a cada aba é indicado na barra de navegação por um ícone específico, assim como os seus respetivos nomes. A primeira aba da barra de navegação corresponde à página inicial da aplicação pelo qual apresenta somente o título da aplicação.

Figura 29 - Página Inicial da aplicação desenvolvida.

6.10.2 Sobre

Tal como se encontra representado na Figura 30, na aba "Sobre" encontra-se explicado o propósito da aplicação, os autores responsáveis pelo seu desenvolvimento, as ferramentas utilizadas para a sua construção, assim como, uma breve descrição do modelo/ algoritmo implementado na mesma. Na parte da descrição do modelo é apresentado a *accuracy* do mesmo, os valores de AUC para a validação interna e externa e as variáveis que fazem parte do modelo, acompanhadas por uma pequena descrição de que valores e métricas é possível ser utilizado para cada uma delas.

Figura 30 - Segunda aba da aplicação onde é descrito o prepósito da aplicação, os autores, a descrição do modelo implementado e ferramentas utilizadas na sua construção.

6.10.3 Previsão

Na aba "Previsão" da aplicação é onde se realiza o cálculo da previsão, em tempo real, se um determinado recém-nascido de muito baixo peso irá sobreviver ou falecer. Para além disso, também indica a percentagem de probabilidade de um determinado bebé vir a falecer. Neste caso, a aplicação já está programada para que, resultados com valores de probabilidades iguais ou superiores a 63%, indicarão que o recém-nascido irá falecer, caso contrário, irá sobreviver.

Na Figura 31 encontra-se representado o formulário, que faz parte da aba "Previsão", com os 9 *inputs* a serem preenchidos pelo utilizador. Estes *inputs* correspondem às variáveis explicativas do modelo implementado na aplicação, sendo elas Idade Gestacional (em dias), Comprimento ao Nascer (em gramas), Corticóides Pré-natais, Género, Média dos Índices de Apgar, Ressuscitação Insuflador, Malformação Congénita Major, Diagnóstico de NEC, e Administração de Ibuprofeno para tratamento de PDA.

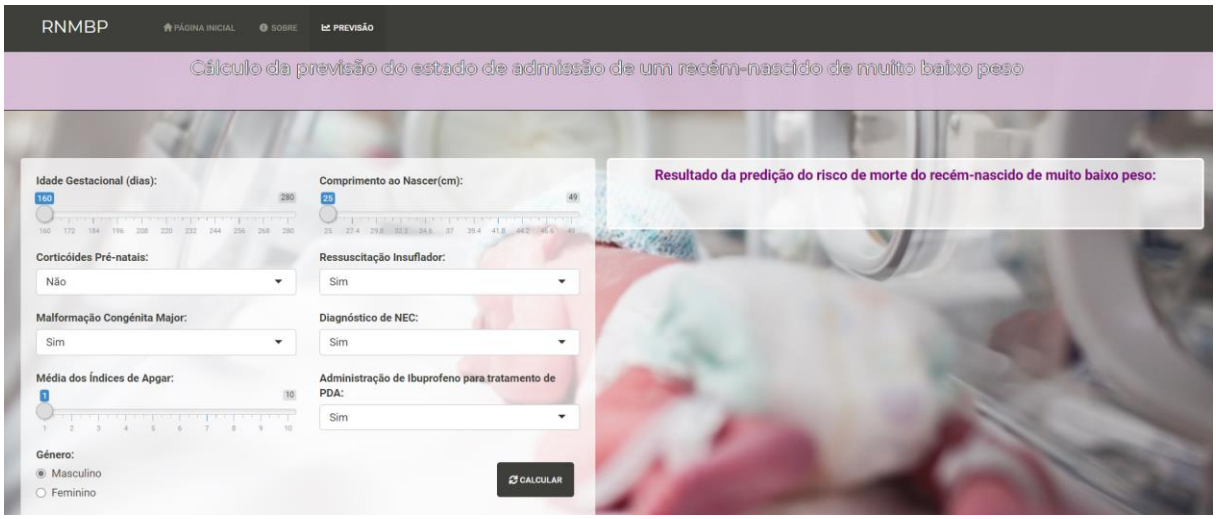

Figura 31 - Aba "Previsão" da aplicação *web*, onde é realizado o cálculo da previsão do estado de admissão de um determinado recém-nascido de muito baixo peso.

Contudo, os valores que o utilizador poderá introduzir em cada *input* encontra-se limitado, sendo que, valores que se encontram fora dos limites presentes na Tabela XXXIII, impossibilita a utilização desta aplicação para o cálculo da previsão.

Tabela XXXIII - Representação dos *inputs* do formulário presente na aba "Previsão" com os respetivos limites de valores que é permitido ser introduzido.

Variável/*Input*	Limite de valores
Idade Gestacional	Entre 160 e 280
Comprimento ao Nascer	Entre 25 e 49
Corticoides Pré-natais	Não/Parcial/Completo
Género	Masculino/Feminino
Média dos Índices de Apgar	Entre 1 e 10
Ressuscitação Insuflador	Sim/Não
Malformação Congénita Major	Sim/Não
Diagnóstico NEC	Sim/Não
Administração de Ibuprofeno para tratamento de PDA	Sim/Não

Assim, quando um utilizador pretender saber qual poderá ser a probabilidade de um determinado recém-nascido de muito baixo peso, que apresente as características presentes na Tabela XXXIII, vir a falecer, basta preencher o formulário e carregar no botão "Calcular", que receberá a resposta de seguida, em tempo real.

De forma a se verificar que de facto a aplicação *Shiny* se encontra funcional, realizou-se alguns ensaios. Na Tabela XXXIV encontra-se representado os valores introduzidos em cada *input* para cada ensaio e na

Tabela XXXV apresenta-se os resultados obtidos para cada ensaio ao fazer o cálculo preditivo através da aplicação desenvolvida.

Tabela XXXIV - Valores introduzidos nos *inputs* do formulário da aplicação para três ensaios diferentes.

Variável	Ensaio 1	Ensaio 2	Ensaio 3
IdadeGestacional	213,0	224,0	200,0
NascimentoComprimento	42,0	36,5	31,0
CorticoidesPrenatais	Completo	Parcial	Completo
Sexo	Masculino	Masculino	Masculino
ApgarMedia	9,0	5,0	6,0
RessuscitacaoInsuflador	Sim	Sim	Não
MalformacaoCongenitaMajor	Não	Sim	Sim
DiagNec	Não	Não	Não
Administração de Ibuprofeno para tratamento de PDA	Não	Não	Não

Tabela XXXV - Resultados obtidos para os três ensaios, utilizando a aplicação *Shiny* desenvolvida.

	Ensaio 1	Ensaio 2	Ensaio 3
Resultado probabilístico (%)	0,838	15,227	67,877
Resultado estado de admissão	Sobreviver	Sobreviver	Falecer

Pela Figura 32 é possível verificar-se que de facto, quando se preenche o formulário e se carrega no botão "Calcular" na aplicação, esta retribui o valor da probabilidade de um recém-nascido de muito baixo peso, com as características introduzidas, vir a falecer. Além disso também indica qual poderá ser o seu estado de admissão final. Os dados introduzidos correspondem ao ensaio 1.

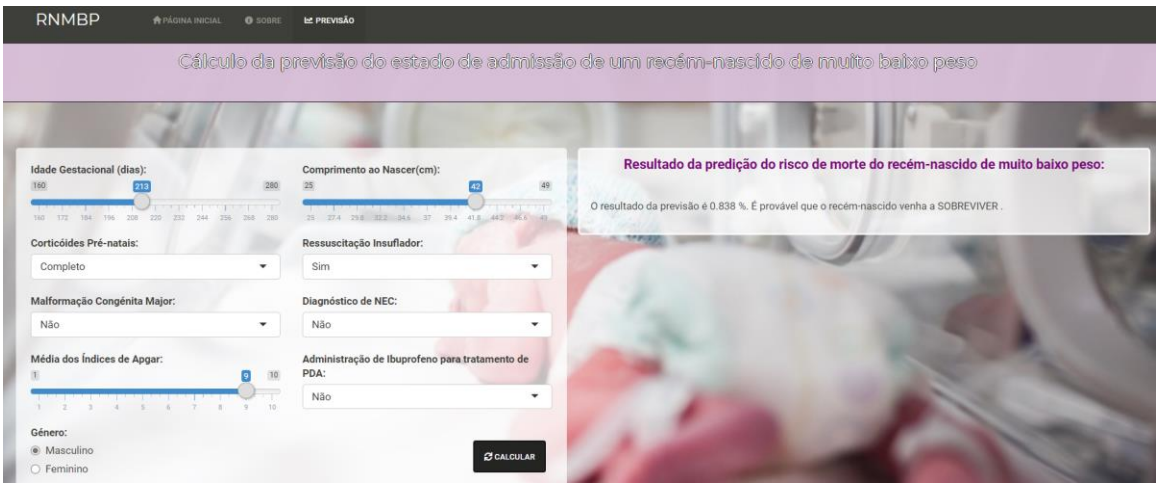

Figura 32 - Resultado obtido, segundo as características do ensaio 1, através da aplicação *Shiny*.

7. CONCLUSÕES E TRABALHO FUTURO

O valor da taxa de mortalidade em recém-nascidos de muito baixo peso é um dos principais problemas da saúde pública que mais preocupação tem suscitado a nível nacional, assim como, a nível mundial. Neste sentido, surgiu o tema desta dissertação no qual se pretendeu desenvolver um modelo preditivo que fosse capaz de prever o risco de morte em recém-nascidos de muito baixo peso. Este modelo servirá para ajudar os profissionais de saúde a perceberem quais os que têm um maior risco de morte, e assim, auxiliá-los na tomada de decisão de como devem atuar perante um recém-nascido desta população.

Um dos modelos desenvolvidos ao longo deste trabalho e que apresenta uma boa capacidade preditiva, trata-se de um modelo de regressão logística capaz de prever com 0,926 de certeza o estado de admissão de um recém-nascido de muito baixo peso, ou seja, se irá sobreviver ou falecer. Este modelo tem em conta 9 características que são medidas nas unidades hospitalares a estes recém-nascidos, nomeadamente a idade gestacional (em dias), o comprimento ao nascer (em gramas), a administração ou não de corticoides pré-natais, o sexo do recém-nascido, a média dos três índices apgar (1º,5º e 10º minutos), a ocorrência ou não de reanimação inicial com insuflador, se apresenta ou não alguma malformação congénita major, se apresenta ter ou não enterocolite necrotizante (NEC), e se foi administrado Ibuprofeno ao recém-nascido para tratamento de persistência de ductos arteriosos (PDA).

Relativamente aos valores apresentados por este modelo para as medidas de qualidade de ajuste e de capacidade preditiva, são resultados bastante satisfatórios, tendo em vista que, os valores de área sob a curva ROC foram de 0,891 e 0,797 para a validação interna e externa, respetivamente e que o modelo se encontra bem ajustado com valores de 373,498, 428,279 para o AIC e BIC, respetivamente, assim como, valores de 0,314, 0,139 e 0,367 para os pseudo R^2 de McFadden, Cox e Snell e Nagelkerke, respetivamente. Porém, é de realçar que todos os modelos desenvolvidos neste trabalho apresentaram um bom ajuste aos dados e uma boa capacidade preditiva, sendo que, qualquer um deles seria um bom candidato ao desenvolvimento de um classificador.

Por fim, com este trabalho também foi possível concluir que este classificador poderá ser uma ferramenta alternativa a usar nos centros hospitalares, uma vez que, este apresentou melhores resultados preditivos que os indicadores CRIB e SNAPPE II, que são muito utilizados atualmente. Assim, para um conjunto de teste com 350 registos, o classificador desenvolvido apresentou um valor de AUC de 0,909 e os indicadores CRIB e SNAPPE II um valor de 0,830 e 0,794, respetivamente.

Numa perspetiva de trabalho futuro, podia-se construir um outro tipo de modelo preditivo, nomeadamente um modelo de *machine learning,* como por exemplo, *Random Forest, Support Vector*

Machine, Artificial Neural Networks, entre outros, uma vez que, atualmente se tem feito muitos estudos nessa área e têm apresentado bons resultados. Para além disso, seria oportuno construir um modelo que fosse capaz predizer uma morbilidade que estivesse associada a esses recém-nascidos, como por exemplo, a sépsis, e comparar o mesmo com índices que já foram admitidas nos serviços de cuidados intensivos de neonatologia.

Relativamente à aplicação *web* que suporta o modelo preditivo, seria interessante explorar mais recursos que se poderiam integrar na aplicação *Shiny*, como por exemplo, o uso de JavaScript*,* para tornar a aplicação mais dinâmica, e de Jquery. Uma outra alternativa, poderia ser explorar e recorrer a outras ferramentas para desenvolvimento da aplicação, como por exemplo, recorrer à linguagem Python que fornece uma biblioteca que permite a criação de aplicações *Dash*.

BIBLIOGRAFIA

Acock, A. C. (2005). Working with missing values. *Journal of Marriage and Family*, *67*(4), 1012–1028. https://doi.org/10.1111/j.1741-3737.2005.00191.x

Acuña, E., & Rodriguez, C. (2004). The treatment of missing values and its effect on classifier accuracy. In *Classification, Clustering, and Data Mining Applications* (pp. 639–647). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-17103-1_60

Alemayehu, D., & Zou, K. H. (2012). Applications of ROC Analysis in Medical Research: Recent Developments and Future Directions. *Academic Radiology*, *19*(12), 1457–1464. https://doi.org/10.1016/J.ACRA.2012.09.006

Alonzo, T. A., & Pepe, M. S. (2007). Development and evaluation of classifiers. In W. T. Ambrosius (Ed.), *Methods in Molecular Biology, vol. 404: Topics in Biostatistics* (pp. 89–116). Humana Press.

Ambalavanan, N., & Carlo, W. A. (2001). Comparison of the prediction of extremely low birth weight neonatal mortality by regression analysis and by neural networks. *Early Human Development*, *65*(2), 123–137.

Asoglu, M. R., Bears, B., Turan, S., Harman, C., & Turan, O. M. (2020). The factors associated with mode of delivery in fetuses with congenital heart defects. *Journal of Maternal-Fetal & Neonatal Medicine*, *33*(5), 816–824. https://doi.org/10.1080/14767058.2018.1505855

Attali, D. (2020). *shinyjs: easily improve the user experience of your shiny apps in seconds*. http://https//cran.r-project.org/web/packages/shinyjs/index.html

Austin, P. C., & Tu, J. V. (2004). Automated variable selection methods for logistic regression produced unstable models for predicting acute myocardial infarction mortality. *Journal of Clinical Epidemiology*, *57*(11), 1138–1146. https://doi.org/10.1016/J.JCLINEPI.2004.04.003

Ballot, D. E., Chirwa, T. F., & Cooper, P. A. (2010). Determinants of survival in very low birth weight neonates in a public sector hospital in Johannesburg. *BMC Pediatrics*, *10*(30), 1–11. http://www.biomedcentral.com/1471-2431/10/30%5Cnhttp://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed9&NEWS=N&AN=2010358492

Basu, S., Rathore, P., & Bhatia, B. D. (2008). Predictors of mortality in very low birth weight neonates in India. *Singapore Medical Journal*, *49*(7), 556–560.

Batista, G. E. A. P. A., & Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, *17*(5–6), 519–533.

https://doi.org/10.1080/713827181

Beeley, C. (2013). Web application development with R using Shiny. In *Packt Publishing* (Vol. 2, Issues 2–3). https://doi.org/10.1017/CBO9781107415324.004

Bernstein, I. M., Horbar, J. D., Badger, G. J., Ohlsson, A., & Golan, A. (2000). Morbidity and mortality among very-low-birth-weight neonates with intrauterine growth restriction. *American Journal of Obstetrics and Gynecology*, *182*(1), 198–206. https://doi.org/https://doi.org/10.1016/S0002-9378(00)70513-8

Bielecki, M. J. V., & White, E. D. (2005). Estimating cost growth from schedule changes: A regression. *The AACE International Journal of Cost Estimation, Cost/Schedule Control, and Project Management*, *47*(8), 28–34.

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling*, *15*(4), 651–75.

Braga, A. C., & Carneiro, P. (2016). Development and validation of a logistic regression model to estimate the risk of WMSDs in Portuguese home care nurses. In O. Gervasi, B. Murgante, S. Misra, A. M. A. C. Rocha, C. M. Torre, D. Taniar, B. O. Apduhan, E. Stankova, & S. Wang (Eds.), *Computational Science and Its Applications – ICCSA 2016* (Vol. 9786, pp. 97–109). Springer, Cham. https://doi.org/10.1007/978-3-319-42085-1_8

Braga, A. C. da S. (2000). *Curvas ROC: Aspectos funcionais e aplicações*. Universidade do Minho.

Braga, A. C. da S., & 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. https://doi.org/10.1108/09526860310479677

Brito, A. S. J. de, Matsuo, T., Gonzalez, M. R. C., Carvalho, A. B. R. de, & Ferrari, L. S. L. (2003). CRIB score, birth weight and gestational age in neonatal mortality risk evaluation. *Revista de Saúde Pública*, *37*(5), 597–602. https://doi.org/10.1590/s0034-89102003000500008

Carneiro, J. A., Vieira, M. M., Reis, T. C., & Caldeira, A. P. (2012). Risk factors for mortality of very low birth weight newborns at a neonatal intensive care unit. *Revista Paulista de Pediatria*, *30*(3), 369–376. https://doi.org/10.1590/S0103-05822012000300010

Carter, J. V., Pan, J., Rai, S. N., & Galandiuk, S. (2016). ROC-ing along: Evaluation and interpretation of receiver operating characteristic curves. *Surgery*, *159*(6), 1638–1645. https://doi.org/10.1016/J.SURG.2015.12.029

Chang, W., Ribeiro, B. B., RStudio, Studio, A., & Incorporated, A. S. (2018). *shinydashboard: create dashboards with "shiny."* http://https//cran.r-

project.org/web/packages/shinydashboard/index.html

Chang, W., RStudio, Park, T., Dziedzic, L., Willis, N., Corporation, G., McInerney, M., Incorporated, Systems, A., & Ltd, C. (2018). *shinythemes: themes for shiny*. http://https//cran.r-project.org/web/packages/shinythemes/index.html

Commenges, D., Sayyareh, A., Letenneur, L., Guedj, J., & Bar-Hen, A. (2008). Estimating a difference of Kullback-Leibler risks using a normalized difference of AIC. *Annals of Applied Statistics*, *2*(3), 1123–1142. https://doi.org/10.1214/08-AOAS176

Considine, G., & Zappalà, G. (2002). The influence of social and economic disadvantage in the academic performance of school students in Australia. *Journal of Sociology*, *38*(2), 129–148. https://doi.org/https://doi.org/10.1177/144078302128756543

Costa, L. P. F. da C. (2018). *Modelação de dados dos serviços de urgência no Hospital de Braga*. Universidade do Minho.

Criscuolo, N. G., & Angelini, C. (2020). Structurly: A novel shiny app to produce comprehensive, detailed and interactive plots for population genetic analysis. *PLoS ONE*, *15*(2), 1–12. https://doi.org/10.1371/journal.pone.0229330

Cunha, M., Cadete, A., Virella, D., & Grupo do Registo Nacional de Muito Baixo Peso. (2010). Acompanhamento dos recém-nascidos de muito baixo peso em Portugal. *Acta Pediátrica Portuguesa*, *41*(4), 155–161.

Curley, C., Krause, R. M., Feiock, R., & Hawkins, C. V. (2019). Dealing with missing data: a comparative exploration of approaches using the integrated city sustainability database. *Urban Affairs Review*, *55*(2), 591–615. https://doi.org/10.1177/1078087417726394

Cutland, C. L., Lackritz, E. M., Mallett-Moore, T., Bardají, A., Chandrasekaran, R., Lahariya, C., Nisar, M. I., Tapia, M. D., Pathirana, J., Kochhar, S., Muñoz, F. M., & Group, T. brighton collaboration low birth weight working. (2017). Low birth weight: Case definition & guidelines for data collection, analysis, and presentation of maternal immunization safety data. *Vaccine*, *35*(48), 6492–6500. https://doi.org/10.1016/j.vaccine.2017.01.049

da Silva, A. C. (2011). *Análise Estatística de Inquéritos online*. Univerisdade do Minho.

De Castro, E. C. M., Leite, Á. J. M., & Guinsburg, R. (2016). Mortality in the first 24h of very low birth weight preterm infants in the Northeast of Brazil. *Revista Paulista de Pediatria*, *34*(1), 106–113. https://doi.org/10.1016/j.rppede.2015.12.008

de Vonderweid, U., Carta, A., Chiandotto, V., Chiappe, F., Colarizi, S., Colarizi, P., Corchia, C., De Luca, T., Didato, M., & Gioeli, R. (1991). Italian Multicenter Study on Very Low Birth Weight Babies. *Annali*

Dell'istituto Superiore Di Sanita, *27*(4), 633–650.

Dobbin, K. K., & Simon, R. M. (2011). Optimally splitting cases for training and testing high dimensional classifiers. *BMC Medical Genomics*, *4*(31), 1–8. https://doi.org/10.1186/1755-8794-4-31

Dorling, J. S., Field, D. J., & Manktelow, B. (2005). Neonatal diseases severity scoring systems. *Archives of Disease in Childhood: Fetal and Neonatal Edition*, *90*(1), 11–16. https://doi.org/10.1136/adc.2003.048488

Dreiseitl, S., & Ohno-Machado, L. (2002). Logistic regression and artificial neural network classification models: A methodology review. *Journal of Biomedical Informatics*, *35*(5–6), 352–359. https://doi.org/10.1016/S1532-0464(03)00034-0

Fan, J., Upadhye, S., & Worster, A. (2006). Understanding receiver operating characteristic (ROC) curves. *CJEM*, *8*(01), 19–20. https://doi.org/10.1017/S1481803500013336

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

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., & 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), 419–422. https://doi.org/10.1136/adc.2003.031286

Gandrud, C. (2013). *Reproducible research with R and R Studio* (C. Press (ed.)). Taylor & Francis. https://books.google.com.br/books?id=u-nuzKGvoZwC

Gera, T., & Ramji, S. (2001). Early predictors of mortality in very low birth weight neonates. *Indian Pediatrics*, *38*(6), 596–604.

Goksuluk, D., Korkmaz, S., Zararsiz, G., & Karaagaoglu, A. E. (2016). EasyROC: An interactive web-tool for roc curve analysis using R language environment. *The R Journal*, *8*(2), 213–230. https://doi.org/10.32614/rj-2016-042

Gooden, M., Younger, N., & Trotman, H. (2014). What is the best predictor of mortality in a very low birth weight infant population with a high mortality rate in a medical setting with limited resources? *American Journal of Perinatology*, *31*(6), 441–446. https://doi.org/10.1055/s-0033-1351658

Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, *8*(3), 206–213. https://doi.org/10.1007/s11121-007-0070-9

Gray, J. E., Richardson, D. K. McCormick, M. C., & Workman-Daniels, K. Goldmann, D. A. (1992).

Neonatal therapeutic intervention scoring system: a therapy-based severity-of-illness index. *Pediatrics*, *90*(4), 561–567.

Gulliver, K., & Yoder, B. A. (2018). Bronchopulmonary dysplasia: effect of altitude correction and role for the Neonatal Research Network Prediction Algorithm. *Journal of Perinatology*, *38*(8), 1046–1050. https://doi.org/https://doi.org/10.1038/s41372-018-0113-z

Harsha, S. S., & Archana, B. R. (2015). SNAPPE-II (score for neonatal acute physiology with perinatal extension-II) in predicting mortality and morbidity in NICU. *Journal of Clinical and Diagnostic Research*, *9*(10), 10–12. https://doi.org/10.7860/JCDR/2015/14848.6677

Hauck, W. W., & Donner, A. (1977). Wald's Test as Applied to Hypotheses in Logit Analysis. *Journal of the American Statistical Association*, *72*(360), 851–853. https://doi.org/doi:10.2307/2286473

Henry, A. J., Hevelone, N. D., Lipsitz, S., & Nguyen, L. L. (2013). Comparative methods for handling missing data in large databases. *Journal of Vascular Surgery*, *58*(5), 1353-1359.e6. https://doi.org/10.1016/j.jvs.2013.05.008

Hosmer, D.W, & Lemeshow, S. (1989). Applied logistic regression. In *John Wiley & sons*.

Hosmer, David W., & Lemeshow, S. (2000). *Applied Logistic Regression* (Second Edi). John Wiley & Sons, Inc.

Hu, B., Shao, J., & Palta, M. (2006). PSEUDO-R 2 in logistic regression model. *Statistica Sinica*, *16*(3), 847–860.

Hu, Z., & Lo, C. P. (2007). Modeling urban growth in Atlanta using logistic regression. *Computers, Environment and Urban Systems*, *31*(6), 667–688. https://doi.org/10.1016/J.COMPENVURBSYS.2006.11.001

Hull, M. A., Fisher, J. G., Gutierrez, I. M., Jones, B. A., Kang, K. H., Kenny, M., Zurakowski, D., Modi, B. P., Horbar, J. D., & Jaksic, T. (2014). Mortality and Management of Surgical Necrotizing Enterocolitis in Very Low Birth Weight Neonates: A Prospective Cohort Study. *Journal of the American College of Surgeons*, *218*(6), 1148–1155. https://doi.org/https://doi.org/10.1016/j.jamcollsurg.2013.11.015

Jafrasteh, A., Baharvand, P., & Karami, F. (2017). Clinical risk index for neonates II score for the prediction of mortality risk in premature neonates with very low birth weight. *World Family Medicine Journal/Middle East Journal of Family Medicine*, *15*(8), 183–187. https://doi.org/10.5742/mewfm.2017.93074

Jašić, M., Dessardo, N. S., Dessardo, S., & Rukavina, K. M. (2016). CRIB II score versus gestational age and birth weight in preterm infant mortality prediction: Who will win the bet? *Signa Vitae*, *11*(1),

172–181. https://doi.org/10.22514/SV111.052016.12

Jeschke, E., Biermann, A., Günster, C., Böhler, T., Heller, G., Hummler, H. D., & Bührer, C. (2016). Mortality and major morbidity of very-low-birth-weight infants in Germany 2008-2012: A report based on administrative data. *Frontiers in Pediatrics*, *4*, 1–8. https://doi.org/10.3389/fped.2016.00023

Kang, H. (2013). The prevention and handling of the missing data. *Korean Journal of Anesthesiology*, *64*(5), 402–406. https://doi.org/10.4097/kjae.2013.64.5.402

Kardum, D., Filipović-Grčić, B., Müller, A., & Dessardo, S. (2019). Survival until discharge of very-low-birth-weight infants in two croatian perinatal care regions: a retrospective cohort study of time and cause of death. *Acta Clinica Croatica*, *58*(3), 446–454. https://doi.org/10.20471/acc.2019.58.03.07

Konishi, S., & Kitagawa, G. (2008). *Information Criteria and Statistical Modeling*. Springer Science + Business Media.

Kuhn, M. (2019). *The caret Package*. http://topepo.github.io/caret/

Lasko, T. A., Bhagwat, J. G., Zou, K. H., & Ohno-Machado, L. (2005). The use of receiver operating characteristic curves in biomedical informatics. *Journal of Biomedical Informatics*, *38*(5), 404–415. https://doi.org/10.1016/j.jbi.2005.02.008

Lee, H. C., & Gould, J. B. . (2006). Survival advantage associated with cesarean delivery in very low birth weight vertex neonates. *Obstet Gynecology*, *107*(1), 97–105. https://doi.org/doi:10.1097/01.AOG.0000192400.31757.a6

Lee, K. I., & Koval, J. J. (1997). Determination of the best significance level in forward stepwise logistic regression. *Communications in Statistics - Simulation and Computation*, *26*(2), 559–575. https://doi.org/10.1080/03610919708813397

Liu, W., Bretz, F., Srimaneekarn, N., Peng, J., & Hayter, A. J. (2019). Confidence Sets for Statistical Classification. *Stats*, *2*(3), 332–346. https://doi.org/10.3390/stats2030024

Liu, Y., Zhang, H. H., & Wu, Y. (2011). Hard or soft classification? large-margin unified machines. *Journal of the American Statistical Association*, *106*(493), 166–177. https://doi.org/https://doi.org/10.1198/jasa.2011.tm10319

Maier, R. F., Caspar-Karweck, U. E., Grauel, L. E., Bassir, C., Metze, B. C., & Obladen, M. (2002). A comparison of two mortality risk scores for very low birthweight infants: clinical risk index for babies and Berlin score. *Intensive Care Medicine*, *28*(9), 1332–1335. https://doi.org/10.1007/s00134-002-1403-6

Maroco, J., Silva, D., Rodrigues, A., Guerreiro, M., Santana, I., & De Mendonça, A. (2011). Data mining

methods in the prediction of Dementia: A real-data comparison of the accuracy, sensitivity and specificity of linear discriminant analysis, logistic regression, neural networks, support vector machines, classification trees and random forests. *BMC Research Notes*, *4*(1), 299. https://doi.org/10.1186/1756-0500-4-299

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., & 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. https://doi.org/10.1038/sj.jp.7211362

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. *Frontiers in Econometrics*, *33*(8), 105–142. https://doi.org/10.1080/07373937.2014.997882

Medlock, S., Ravelli, A. C. J., Tamminga, P., Mol, B. W. M., & Abu-Hanna, A. (2011). Prediction of mortality in very premature infants: A systematic review of prediction models. *PLoS ONE*, *6*(9), 1–9. https://doi.org/10.1371/journal.pone.0023441

Medvedev, M. M., Brotherton, H., Gai, A., Tann, C., Gale, C., Waiswa, P., Elbourne, D., Lawn, J. E., & Allen, E. (2020). Development and validation of a simplified score to predict neonatal mortality risk among neonates weighing 2000 g or less (NMR-2000): an analysis using data from the UK and The Gambia. *The Lancet Child and Adolescent Health*, *4*(4), 299–311. https://doi.org/10.1016/S2352-4642(20)30021-3

Menard, S. (2000). Coefficients of determination for multiple logistic regression analysis. *The American Statistician*, *54*(1), 17–24. https://doi.org/10.1080/00031305.2000.10474502

Metz, C.E. (1986). Special articles roc methodology in radiologic imaging. *Investigative Radiology*, *21*(9), 720–733. https://doi.org/10.1097/00004424-198609000-00009

Metz, Charles E. (2008). ROC analysis in medical imaging: a tutorial review of the literature. *Radiological Physics and Technology*, *1*(1), 2–12. https://doi.org/10.1007/s12194-007-0002-1

Mia, R. A., Etika, R., Harianto, A., Indarso, F., & Damanik, S. M. (2005). The use of score for neonatal acute physiology perinatal extention II (SNAPPE II) in predicting neonatal outcome in neonatal intensive care unit. *Paediatrica Indonesiana*, *45*(11–12), 241–245. https://doi.org/10.14238/pi45.6.2005.241-5

Mittlböck, M., & Schemper, M. (1996). Explained variation for logistic regression. *Statistics in Medicine*, *15*(19), 1987–1997. https://doi.org/10.1002/(SICI)1097-0258(19961015)15:19<1987::AID-SIM318>3.0.CO;2-9

Mourão, M. F., & Braga, A. C. da S. (2012). Evaluation of the CRIB as an indicator of the performance of neonatal intensive care units using the software ROCNPA. *12th International Conference on Computational Science and Its Applications*, *1*, 151–154. https://doi.org/10.1109/ICCSA.2012.37

Mourão, M. F., Braga, A. C. da S., Almeida, A., Mimoso, G., & Oliveira, P. N. (2015). Adjusting covariates in CRIB score index using ROC regression analysis. In O. Gervasi, B. Murgante, S. Misra, M. L. Gavrilova, A. M. A. C. Rocha, C. Torre, D. Taniar, & B. O. Apduhan (Eds.), *Computational Science and Its Applications – ICCSA 2015* (pp. 157–171). Springer, Cham. https://doi.org/https://doi.org/10.1007/978-3-319-21407-8_12

Mourão, M. F., Braga, A. C. da S., & Oliveira, P. N. (2014). Accommodating maternal age in CRIB scale: Quantifying the effect on the classification. *In International Conference on Computational Science and Its Applications*, 566–579. https://doi.org/10.1007/978-3-319-09150-1_41

Mourão, M. F. T. G. F. (2016). *Aplicação da metodologia ROC na avaliação de desempenho de índices de gravidade clínica em Unidades de Neonatologia de Portugal*. Universidade do Minho.

Naskar, N., Swain, A., Das, K. N., & Patnayak, A. K. (2014). Maternal risk factors, complications and outcome of very low birth weight babies: prospective Cohort study from a tertiary care centre in Odisha. *Journal of Neonatal Biology*, *3*(3), 1–7. https://doi.org/10.4172/2167-0897.1000142

Nayeri, F., Emami, Z., Mohammadzadeh, Y., Shariat, M., Sagheb, S., & Sahebi, L. (2019). Mortality and Morbidity Patterns of Very Low Birth Weight Newborns in Eastern Mediterranean Region: A Meta-Analysis Study. *Journal of Pediatrics Review*, *7*(2), 67–76. https://doi.org/10.32598/jpr.7.2.67

NCSS. (n.d.). *Stepwise Regression*. NCSS Statistical Software. https://doi.org/10.4135/9781412950589.n974

Noghrehchi, F., Stoklosa, J., & Penev, S. (2020). Multiple imputation and functional methods in the presence of measurement error and missingness in explanatory variables. *Computational Statistics*, 1–27. https://doi.org/10.1007/s00180-020-00976-2

Pallmann, P., Wan, F., Mander, A. P., Wheeler, G. M., Yap, C., Clive, S., Hampson, L. V., & Jaki, T. (2019). Designing and evaluating dose-escalation studies made easy: The MoDEsT web app. *Clinical Trials*, *1*, 1–10. https://doi.org/10.1177/1740774519890146

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

Peng, G., Tang, Y., Cowan, T. M., Enns, G. M., Zhao, H., & Scharfe, C. (2020). Reducing false-positive results in newborn screening using machine learning. *International Journal of Neonatal Screening*,

6(16), 1–12. https://doi.org/10.3390/ijns6010016

Pepe, M. S. (2003). *The statistical evalution of medical tests for classification and prediction*. Oxford University Press.

Pepe, M. S. (2005). Evaluating technologies for classification and prediction in medicine. *Statistics in Medicine*, *24*(24), 3687–3696. https://doi.org/10.1002/sim.2431

Rachuri, S., Paul, S., & D., J. M. (2019). SNAPPE II score: predictor of mortality in NICU. *International Journal of Contemporary Pediatrics*, *6*(2), 422–426. https://doi.org/10.18203/2349-3291.ijcp20190544

Raja, M., K, S. K., & Deneshkumar, V. (2017). Regression Modeling for Maternal Determinants of low birth weight. *International Journal of Statistics and Systems*, *12*(3), 585–591.

Royston, P., & White, I. R. (2011). Multiple imputation by chained equations (MICE): Implementation in Stata.". *Journal of Statistical Software*, *45*(4), 1–20.

RStudio. (2020). *Share your apps*. https://shiny.rstudio.com/tutorial/written-tutorial/lesson7/

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, *63*(3), 581–592. https://doi.org/10.1093/biomet/63.3.581

Salgado, C. M., Azevedo, C., Proença, H., & Vieira, S. M. (2016). Missing Data. In *Secondary Analysis of Electronic Health Records* (pp. 163–183). Springer, Cham. https://doi.org/10.1007/978-3-319-43742-2_13

Sarquis, A. L. F., Miyaki, M., & Cat, M. N. L. (2002). Aplicação do escore CRIB para avaliar o risco de mortalidade neonatal. *Jornal de Pediatria*, *78*(3), 225–229.

Sathar, A., Shanavas, A., Girijadevi, P. S., Jasmin, L. B., Kumar, S. S., & Pillai, R. K. (2018). Risk factors of retinopathy of prematurity in a tertiary care hospital in South India. *Clinical Epidemiology and Global Health*, *6*(1), 44–49. https://doi.org/10.1016/j.cegh.2017.02.002

Scheffer, J. (2002). Dealing with missing data. *Research Letters in the Information and Mathematical Sciences*, *3*(1), 153–160.

Seal, A., & Wild, D. J. (2018). Netpredictor: R and Shiny package to perform drug-target network analysis and prediction of missing links. *BMC Bioinformatics*, *19*(265), 1–10. https://doi.org/10.1186/s12859-018-2254-7

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2015). *ROCR*. http://rocr.bioinf.mpi-sb.mpg.de/

Sivaprakasam, B., & Sadagopan, P. (2019). Development of an Interactive Web Application "Shiny App for Frequency Analysis on Homo sapiens Genome (SAFA-HsG)." *Interdisciplinary Sciences: Computational Life Sciences*, *11*(4), 723–729. https://doi.org/10.1007/s12539-019-00340-z

Spackman, K. A. (1989). Signal detection theory: valuable tools for evaluating inductive learning. *Proceedings of the Sixth International Workshop on Machine Learning*, 160–163. https://doi.org/10.1016/B978-1-55860-036-2.50047-3

Stare, J., & Maucort-Boulch, D. (2016). Odds ratio, hazard ratio and relative risk. *Metodoloski Zvezki*, *13*(1), 59–67.

Steyerberg, E. W., Eijkemans, M. J. C., Jr, F. E. H., & Habbema, J. D. F. (2000). Prognostic modelling with logistic regression analysis : a comparison of selection and estimation methods in small data sets. *Statistics in Medicine*, *19*(8), 1059–1079.

Stylianou, C., Pickles, A., & Roberts, S. A. (2013). Using Bonferroni, BIC and AIC to assess evidence for alternative biological pathways: Covariate selection for the multilevel Embryo-Uterus model. *BMC Medical Research Methodology*, *13*(1), 1–13. https://doi.org/10.1186/1471-2288-13-73

Sundaram, V., Dutta, S., Ahluwalia, J., & Narang, A. (2009). Score for neonatal acute physiology II predicts mortality and persistent organ dysfunction in neonates with severe septicemia. *Indian Pediatrics*, *46*(9), 775–780.

Swets, J. A. (1996). *Signal detection theory and roc analysis in psychology and diagnostics* (1st Editio). Scientific Psychology Series.

Tsai, C.-F., & Chang, F.-Y. (2016). Combining instance selection for better missing value imputation. *Journal of Systems and Software*, *122*, 63–71. https://doi.org/10.1016/J.JSS.2016.08.093

van Ginkel, J. R., Linting, M., Rippe, R. C. A., & van der Voort, A. (2019). Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *Journal of Personality Assessment*, *102*(3), 297–308. https://doi.org/10.1080/00223891.2018.1530680

Vincer, M. J., Armson, B. A., Allen, V. M., Allen, A. C., Stinson, D. A., Whyte, R., & Dodds, L. (2015). An Algorithm for Predicting Neonatal Mortality in Threatened Very Preterm Birth. *Journal of Obstetrics and Gynaecology Canada*, *37*(11), 958–965. https://doi.org/10.1016/S1701-2163(16)30045-7

Walker, D. A., & Smith, T. J. (2016). Nine pseudo R2 indices for binary logistic regression models. *Journal of Modern Applied Statistical Methods*, *15*(1), 848–854. https://doi.org/10.22237/jmasm/1462078200

Wickham, H. (2015). ggplot2: elegant Graphics for Data Analysis. In *Springer*. https://doi.org/10.1007/978-0-387-98141-3

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis* (2nd ed.). Springer International Publishing. https://doi.org/10.1007/978-3-319-24277-4

Wickham, H., & Grolemund, G. (2016). *R for Data Science: import, tidy, transform, visualize and model*

data (M. Beaugureau & M. Loukides (eds.)). O'Reilly. https://doi.org/10.1016/j.envsoft.2017.01.023

Wu, P. L., Lee, W.-T., Lee, P.-L., & Chen, H.-L. (2015). Predictive power of serial neonatal therapeutic intervention scoring system scores for short-term mortality in very-low-birth-weight infants. *Pediatrics and Neonatology*, *56*(2), 108–113. https://doi.org/10.1016/j.pedneo.2014.06.005

Yanagihara, H., Kamo, K., Imori, S., & Satoh, K. (2012). Bias-corrected AIC for selecting variables in multinomial logistic regression models. *Linear Algebra and Its Applications*, *436*(11), 4329–4341. https://doi.org/10.1016/J.LAA.2012.01.018

Zaghdoudi, T. (2013). Bank failure prediction with logistic regression. *International Journal of Economics and Financial Issues*, *3*(2), 537–543.

Zellner, D., Keller, F., & Zellner, G. E. (2004). Variable selection in logistic regression models. *Communications in Statistics - Simulation and Computation*, *33*(3), 787–805. https://doi.org/10.1081/SAC-200033363

Zhang, Z. (2016). Variable selection with stepwise and best subset approaches. *Annals of Translational Medicine*, *4*(7), 1–6. https://doi.org/10.21037/atm.2016.03.35

Zile, I., Ebela, I., & Rozenfelde, I. R. (2017). Risk factors associated with neonatal deaths among very low birth weight infants in Latvia. *Current Pediatric Research*, *21*(1), 64–68.

Zou, K. H., O'Malley, A. J., & Mauri, L. (2007). Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, *115*(5), 654–657. https://doi.org/10.1161/CIRCULATIONAHA.105.594929

APÊNDICE I – BASE DE DADOS

Variável	Descrição	Codificação
DataNascimento	Data em que o recém-nascido de muito baixo peso nasceu	----
MaeIdade	Idade em anos da mãe do recém-nascido de muito baixo peso	----
Distrito	Distrito em que a mãe do recém-nascido reside	----
IdadeGestacional	Idade gestacional em dias do recém-nascido, registada no dia de nascimento	----
NascimentoPeso	Peso do recém-nascido ao nascer, em gramas	----
NascimentoComprimento	Comprimento do recém-nascido ao nascer, em centímetros	----
NascimentoPerimetroCefalico	Perímetro cefálico do recém-nascido ao nascer, em centímetros	----
NascimentoOutborn	Local onde o recém-nascido nasceu	**1 -** Recém-nascido nasceu fora do hospital de registo sendo para ele transferido ("Outborn") **2 –** Recém-nascido nasceu no hospital responsável pelo registo ("Inborn")
NascimentoTipoLocal	Tipo de hospital onde o recém-nascido nasceu	**1 –** Hospital de Apoio Perinatal Diferenciado **2 –** Hospital de Apoio Perinatal

Variável	Descrição	Codificação
		3 – Instituição de Saúde sem Apoio Perinatal **4 –** Local extra Hospitalar **5 –** Hospital Privado
CuidadeosPrenatais	Indica se a mãe recebeu ou não cuidados obstétricos pré-natais	**0 -** Não **1-** Sim
ConcepcaoAssistida	Indica se a conceção foi ou não medicamente assistida	**0 -** Não **1-** Sim
CorticoidesPrenatais	Indica se houve ou não administração de corticoides antes do nascimento do recém-nascido	**1 –** Não **2 –** Parcial, ou seja, o nascimento ocorreu menos de 24 horas após a primeira dose de corticoides, ou mais de uma semana após a última dose **3 –** Completo, ou seja, o nascimento ocorreu mais de 24 horas e menos de uma semana após pelo menos uma dose administrada de corticoides
PatologiasNaGravidez	Indica se a gravidez decorreu com ou sem patologia materna	**0 -** A gravidez decorreu sem patologia materna **1 -** Foi detetada alguma patologia materna durante a gravidez
TipoDeParto	Indica o tipo de parto	**1 -** Parto vaginal **2 -** Parto cesariana
MotivoDoParto	Indica qual foi o motivo para ter ocorrido o parto	**1 -** Espontâneo **2 -** Patologia materna

Variável	Descrição	Codificação
		3 - Patologia fetal **4 -** Interrupção voluntária da gravidez (IVG)
Sexo	Indica o sexo do recém-nascido	1 - Masculino 2 - Feminino
ApgarMedia	Média dos valores dos indicadores de apgar ao 1º, 5º e 10º minuto	- - - -
RessuscitacaoOxigenio	Indica se ocorreu ou não reanimação inicial com oxigénio	**0 -** Não **1-** Sim
RessuscitacaoInsuflador	Indica se ocorreu ou não reanimação inicial com insuflador	**0 -** Não **1-** Sim
RessuscitacaoEntubacaoEt	Indica se ocorreu ou não reanimação inicial do recém-nascido, através de submissão de ventilação assistida por um tubo endotraqueal (TET)	**0 -** Não **1-** Sim
RessuscitacaoCompressaoCardiaca	Indica se ocorreu ou não reanimação inicial do recém-nascido, através de compressão cardíaca	**0 -** Não **1-** Sim
RessuscitacaoAdrenalina	Indica se ocorreu ou não reanimação inicial do recém-nascido, através da administração de adrenalina por qualquer via	**0 -** Não **1-** Sim
SROxigenio	Indica se foi administrado ou não suplemento de oxigénio	**0 -** Não **1-** Sim

Variável	Descrição	Codificação
	ao recém-nascido durante o internamento	
SRCpap	Indica se o recém-nascido recebeu ou não CPAP (*Continuous Positive Airway Pressure*) nasal durante o internamento	**0 -** Não **1-** Sim
SRVppni	Indica se o recém-nascido recebeu ou não qualquer tipo de ventilação não invasiva por pressão positiva, sem entubação endotraqueal durante o internamento	**0 -** Não **1-** Sim
SRVentilacaoOxidoNitrico	Indica se o recém-nascido recebeu ou não qualquer tipo de ventilação com uso de óxido nítrico, durante o internamento	**0 -** Não **1-** Sim
SRVaf	Indica se o recém-nascido recebeu ou não qualquer tipo de ventilação de alta frequência, via tubo endotraqueal, durante o internamento	**0 -** Não **1-** Sim
SRVafni	Indica se o recém-nascido recebeu ou não qualquer tipo de ventilação de alta frequência não invasiva, via tubo endotraqueal, durante o internamento	**0 -** Não **1-** Sim

Variável	Descrição	Codificação
SRDiasVentilacao	Indica o número de dias em que o recém-nascido tem recebido ou não qualquer tipo de ventilação	- - - -
MalformacaoCongenitaMajor	Indica se foi diagnosticado ou não alguma malformação congénita major ao recém-nascido	0 - Não 1- Sim
SepsisMeningiteTardia	Indica se o recém-nascido foi ou não diagnosticado com sépsis tardia ou com meningite	0 - Não 1- Sim
SurfactanteInicial	Indica se foi administrado ou não ao recém-nascido surfactante exógeno, durante o processo de ventilação	0 - Não 1- Sim
SurfactantePosterior	Indica se foi administrado ou não ao recém-nascido surfactante exógeno, durante o internamento	0 - Não 1- Sim
Crib	Valor obtido pelo cálculo do índice CRIB	- - - -
Snappe2	Valor obtido pelo cálculo do índice SNAPPE II	- - - -
OxigenioDia28	Indica se o recém-nascido se encontrava ou não a receber qualquer suplemento de oxigénio no 28º dia de vida	0 - Não 1- Sim
OxigenioSemana36	Indica se o recém-nascido se encontrava ou não a receber qualquer suplemento de	0 - Não 1- Sim

Variável	Descrição	Codificação
	oxigénio às 36 semanas de vida	
CorticoidesDPC	Indica se foi ou não administrado corticoides ao recém-nascido, depois deste ter nascido, para tratar ou prevenir doenças pulmonares crónicas	0 - Não 1- Sim
DiagSdr	Indica se o recém-nascido apresentava ou não a síndrome de dificuldade respiratória (SDR)	0 - Não 1- Sim
DiagPneumotorax	Indica se o recém-nascido apresentava ou não ar extrapleural, diagnosticado por radiografia ou drenagem pleural	0 - Não 1- Sim
DiagPda	Indica se o recém-nascido apresentava ou não persistência de ductos arteriosos (PDA)	0 - Não 1- Sim
DiagNec	Indica se o recém-nascido apresentava ter ou não Enterocolite Necrotizante (NEC)	0 - Não 1- Sim
DiagPerfuracaoGi	Indica se o recém-nascido teve ou não uma perfuração gastrointestinal focal isolada independente de NEC	0 - Não 1- Sim
PdaProfilatico	Indica se foi ou não administrado ao recém-	0 - Não 1- Sim

Variável	Descrição	Codificação
	nascido indometacina ou ibuprofeno, após o seu nascimento, para profilaxia de PDA	
PdaTerapeutico	Indica se foi ou não administrado ao recém-nascido indometacina ou ibuprofeno, após o seu nascimento, para o tratamento de PDA	**0 -** Não **1-** Sim
CirurgiaPda	Indica se foi ou não realiza laqueação cirúrgica do canal arterial do recém-nascido	**0 -** Não **1-** Sim
CirurgiaNec	Indica se foi ou não realizada alguma intervenção para o tratamento de enterocolite necrotizante (NEC)	**0 -** Não **1-** Sim
CirurgiaMajorOutra	Indica se foi ou não realiza outro tipo de cirurgia major, para além das mencionadas anteriormente	**0 -** Não **1-** Sim
ImagiologiaCerebralDia28	Indica se o recém-nascido realizou ou não algum exame de imagem cerebral até completar 28 dias de vida	**0 -** Não **1-** Sim
EcografiaTf	Indica o número de ecografias transfontanelar ou cerebral (TF) que o recém-nascido realizou	**1 -** 1 ecografia TF **2 -** 2 ecografias TF **3 -** 3 ecografias TF **10 -** mais de 4 ecografias TF
EcografiaTfIdadeUltimaSemana	Indica a idade, em semanas, da ecografia TF que o recém-	- - - -

Variável	Descrição	Codificação
	nascido tenha realizado mais próximo da última semana	
Hpiv	Indica o grau mais grave de hemorragia peri ou intraventricular (HIV) que o recém-nascido possa apresentar	**0 –** Não apresenta ter HIV **1 –** HIV com menos de 10% da área ventricular **2 –** HIV com 10-50% da área ventricular **3 –** Com mais de 50% de HIV na área ventricular
Evhp	Indica se o recém-nascido teve ou não enfarte venoso hemorrágico periventricular (EVHP) associado à HIV	**0 -** Não **1-** Sim
DilatacaoVentricularPh	Indica se um recém-nascido teve ou não dilatação ventricular pós-hemorrágica	**0 -** Não **1-** Sim
LpvGrau	Indica o registo do grau mais grave de leucomalácia periventricular (LPV)	**0 –** Não apresenta LPV **1 –** Hiperecogenicidade periventricular transitória persistindo = 7 dias **2 –** Hiperecogenicidade periventricular transitória que evoluiu para pequenos quistos fronto-parientais localizados **3 –** Hiperecogenicidade periventricular transitória que evoluiu para lesõe quísticas periventriculares extensas

Variável	Descrição	Codificação
		4 – Hiperecogenicidade que atingiu a substância branca profunda, e que evoluiu para lesões quísticas extensas
SepsisMeningitePrecoce	Indica se um recém-nascido apresenta ou não um diagnóstico compatível com sépsis ou meningite precoce	**0 -** Não **1-** Sim
ExameOftalmologico	Indica se o recém-nascido foi ou não submetido a um exame oftalmológico	**0 -** Não **1-** Sim
IdadeDataDestinoEmDias	Idade, em dias, quando o recém-nascido deixou de estar internado	- - - -
Obito	Indica se um recém-nascido faleceu ou não, durante o período de internamento.	**1 –** Faleceu **0 -** Sobreviveu