MESTRADO EM
MÉTODOS QUANTITATIVOS PARA A DECISÃO ECONÓMICA E EMPRESARIAL

(DOCUMENTO DEFINITIVO)

TRABALHO FINAL DE MESTRADO

PROJETO

PREVISÃO DE DURAÇÕES DE CIRURGIAS – O CASO DO HOSPITAL DO ESPÍRITO SANTO DE ÉVORA

MARIANA FILIPA SALGADO PIRES

ORIENTAÇÃO:

PROF.ª DOUTORA RAQUEL MONTEIRO DE NOBRE COSTA BERNARDINO

PROF. DOUTOR DANIEL REBELO DOS SANTOS

(DOCUMENTO ESPECIALMENTE ELABORADO PARA A OBTENÇÃO DO GRAU DE MESTRE)

OUTUBRO - 2022

AGRADECIMENTOS

No decorrer do percurso académico temos a possibilidade de nos desenvolver e desafiar nos mais diversos níveis. Aprendemos a ter paciência quando algo não corre da maneira que esperamos e a festejar as vitórias por mais pequenas que possam ser. O verdadeiro desafio começa ao tentar conciliar duas obrigações semelhantes, o percurso académico e o profissional, sendo esta uma passagem essencial que permite o desenvolvimento da perseverança, consistência e determinação. Assim, depois de concluída mais uma etapa, é imprescindível o reconhecimento daqueles que de alguma forma ajudaram e incentivaram em todo o processo.

Em primeiro lugar, quero e devo destacar a professora Raquel Bernardino e o professor Daniel Santos. É de conhecimento geral que o desenvolvimento de um TFM não é de todo a tarefa mais fácil, exige muita disciplina e paciência, tal como orientadores empenhados e presentes para que possam guiar os alunos na melhor direção possível. Felizmente tive a sorte de ter dois orientadores que acompanharam o TFM desde início e sempre se mostraram prontos para ajudar e tirar quaisquer dúvidas a qualquer hora do dia, esse acompanhamento foi fulcral nos momentos mais difíceis em que parecia não haver uma solução, tal como toda a motivação e confiança que faziam transparecer em todas as reuniões semanais.

Em segundo lugar, agradeço aos meus pais e à minha avó “Lena” que me deram o privilégio de conseguir estudar numa das melhores faculdades de economia e gestão. À minha mãe que sempre se esforçou para me dar mais oportunidades do que ela teve e, em especial, ao meu pai que, apesar de já não estar fisicamente presente, se tornou uma das minhas maiores fontes de motivação mesmo tornando todas as conquistas um pouco menos felizes. Agradeço também ao meu irmão que à sua maneira ajuda a relativizar qualquer situação e é o verdadeiro exemplo de que a inteligência não se resume a decorar matéria de um livro.

Por último, resta-me agradecer aos meus amigos que depois de tantos “não posso, estou a fazer a tese” ainda persistem em convidar-me para tudo. Ao Tomás que há 18 anos ouve as minhas preocupações sem parecer entediado e ao Alexandre que acompanhou o processo do início até ao fim. À minha colega de trabalhos de grupo de mestrado, a minha Mariana e também a pessoa mais positiva que conheço, e ainda à Rita que cresceu comigo e faz questão de ainda cá estar para me ouvir.

RESUMO

O crescente número de cirurgias programadas ao longo dos anos tem-se mostrado um problema recorrente em Portugal, dado que a consequência deste aumento se traduz em listas de espera extensas e de difícil gestão por parte dos hospitais. Assim, surge a necessidade de se desenvolverem modelos que tenham a capacidade de apoiar o planeamento e o agendamento de cirurgias. No entanto, isto nem sempre se mostra uma tarefa fácil devido a toda a imprevisibilidade que envolve o decorrer de uma cirurgia, havendo fatores externos que por vezes acabam por influenciar a duração da mesma.

Desta forma, o foco deste estudo é criar modelos de previsão para durações de cirurgias, através de variáveis obtidas antes dessas mesmas cirurgias, recorrendo a dois modelos distintos: Regressão Linear Múltipla e XGBoost. O primeiro torna-se útil pela fácil interpretação de resultados, servindo de base de comparação para avaliar o desempenho de modelos de *ensemble* como o XGBoost. Para além da dualidade de modelos, são ainda utilizados dois conjuntos distintos de dados: o primeiro é compostos por todos os dados recolhidos e o segundo corresponde ao primeiro conjunto desagregado por especialidade cirúrgica. Esta divisão em dois conjuntos de dados ocorre dado que, teoricamente, as durações de cirurgias têm a particularidade de estarem dependentes da especialidade na qual se inserem. Posteriormente, são utilizados dois indicadores para efetuar as comparações entre os modelos, nomeadamente o R^2 e o *Root Mean Squared Error*.

Com os dados do Hospital do Espírito Santo de Évora, podemos concluir que a um modelo mais complexo não correspondente necessariamente uma melhor performance, estando isto dependente de características específicas de cada conjunto de dados e do próprio objetivo do estudo. Relativamente aos modelos obtidos para ambos os conjuntos de dados, conclui-se ainda que um modelo construído tendo por base cada especialidade pode de facto apresentar resultados piores do que um modelo agregado, acontecendo particularmente quando o número de observações em cada modelo se mostra insuficiente para produzir previsões precisas.

Palavras-chave: Durações de Cirurgias, Gestão do Bloco Operatório, Previsão, Regressão Linear Múltipla, XGBoost.

ABSTRACT

The growing number of surgeries scheduled over the years has been a recurring problem in Portugal, as the consequence of this increase translates into long waiting lists that are difficult to manage by hospitals. Thus, there is a need to develop models that have the ability to support the planning and scheduling of surgeries. However, this is not always an easy task due to the unpredictability that involves the course of a surgery, with external factors that sometimes end up influencing its duration.

Thus, the focus of this study is to create prediction models for surgery durations, through variables obtained *a priori* from these surgeries, using two different models: Multiple Linear Regression and XGBoost. The first one becomes useful due to the interpretability of its results, serving as a base for comparison to evaluate the performance of ensemble models such as XGBoost. In addition to the the comparison of the two models, two distinct datasets are also used: the first is composed of all the data collected and the second corresponds to the first dataset disaggregated by surgical specialty. This division into two datasets occurs given that, in theory, the surgery durations have the particularity of being dependent on the specialty in which they are inserted. Subsequently, two indicators are used to make comparisons between the models, namely the R^2 and the *Root Mean Squared Error*.

With the data from Hospital do Espírito Santo in Évora, we can conclude that a more complex model does not necessarily correspond to a better performance, this being dependent on the specific characteristics of each data set. Regarding the models obtained for both datasets, we can also conclude that a model built based on each specialty may in fact present worse results than an aggregated model, especially when the number of observations in each specialty proves to be insufficient to produce accurate predictions.

Keywords: Surgery Durations, Operating Room Management, Forecast, Multiple Linear Regression, XGBoost.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

BO – Bloco Operatório

LIC – Lista de Inscritos para Cirurgia

HESE – Hospital do Espírito Santo de Évora

TFM – Trabalho Final de Mestrado

RLM – Regressão Linear Múltipla

ML – *Machine Learning*

MSE – *Mean Squared Error*

RMSE – *Root Mean Squared Error*

ICD – *International Classification of Diseases*

GDH – Grupos de Diagnósticos Homogéneos

ORL – Otorrinolaringologia

VIF – *Variance Inflation Factor*

GB – *Gradient Boosting*

SHAP – *SHapley Additive exPlanations*

CAPÍTULO 1 – INTRODUÇÃO

O ambiente hospitalar é dinâmico, sendo sujeito à incerteza, dificuldade em definir prioridades e à coordenação de recursos escassos (Nawaz Ripon e Henrik Nyman, 2020). No entanto, na maioria dos hospitais, a tarefa de planeamento de cirurgias ainda é feita de forma manual. Assim, não é surpreendente o facto de que ao longo dos últimos anos se tenha observado um aumento constante nos estudos relacionados com a eficiência e automatização de um dos espaços mais importantes num hospital, o bloco operatório (BO). Este assunto surge com tamanha relevância devido à intensificação de um problema que se tornou comum em grande parte do mundo – a lista de inscritos para cirurgia (LIC) é cada vez maior. Esta tendência crescente do número de pacientes em lista de espera deve-se ao aumento da esperança média de vida, bem como a uma redução constante nos recursos hospitalares.

A Figura 1 mostra a evolução do número de inscritos em LIC a nível nacional: a Figura 1.A é obtida com dados nacionais disponibilizados pelo Serviço Nacional de Saúde [1][2] entre o período de 2013 a 2022; a Figura 1.B provém de dados do Hospital do Espírito Santo de Évora (HESE) referentes ao ano de 2017.

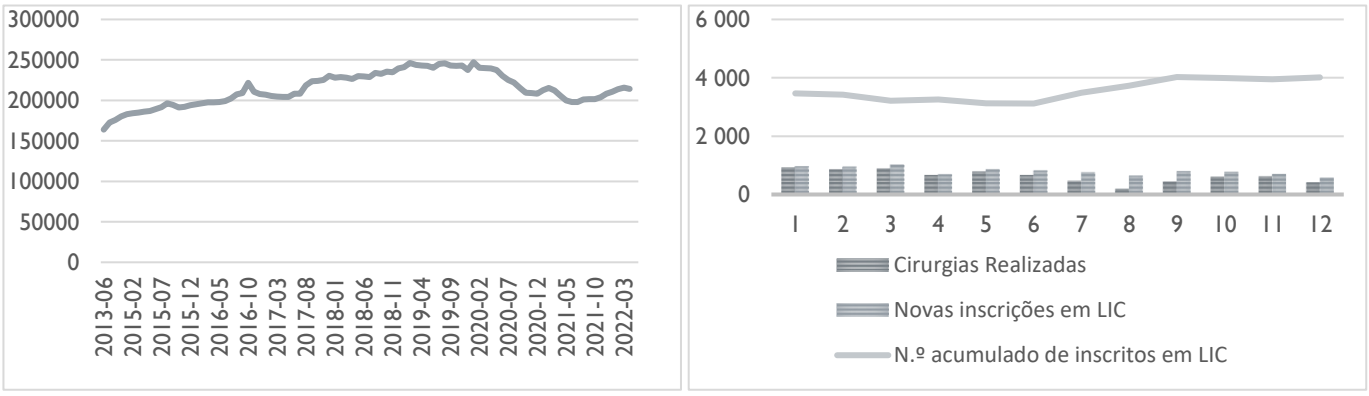


Figura 1.A - N.º de inscritos em LIC.

Figura 1.B - N.º de cirurgias vs. N.º de inscritos em LIC.

Figura 1 - N.º de inscritos em LIC.

Através da Figura 1.A conseguimos ter uma perspetiva geral do comportamento do número acumulado de inscrições em LIC. Nesta pode verificar-se uma tendência crescente ao longo dos anos, havendo uma interrupção a partir de 2020 provocada, provavelmente, pelo aparecimento da pandemia Covid-19 que veio interromper a realização de grande parte das consultas e posteriores inscrições em LIC. Na Figura 1.B, conseguimos observar a discrepância existente entre o número acumulado de doentes em LIC no ano de 2017 e as cirurgias realizadas, bem como as novas entradas de pacientes em LIC em cada um dos meses desse ano. Esta figura reforça a importância do estudo a ser realizado, dado que em nenhum dos meses apresentados se realizaram mais cirurgias do que se contabilizaram novos pacientes a entrar em LIC.

A gestão das listas de espera, procurando equilibrar a oferta e a procura de forma adequada, é, assim, um tópico crítico no contexto da saúde em Portugal. É possível combater o problema de duas formas:

[1] SPMS, 2022. Serviço Nacional de Saúde. Disponível em: https://transparencia.sns.gov.pt/explore/dataset/inscritos-lic-dentro-tmrg/. Consultado a 1 de agosto de 2022.
[2] SPMS, 2022. Serviço Nacional de Saúde. Disponível em: https://transparencia.sns.gov.pt/explore/dataset/inscritos-em-lic-dentro-do-tmrg-180-dias/. Consultado a 1 de agosto de 2022.

1. Aumentar o número de recursos disponíveis (físicos e humanos);
2. Aumentar a eficiência e a eficácia da gestão do BO perante a capacidade instalada.

A segunda opção normalmente é a mais viável para os hospitais, em grande parte pelos custos que a primeira opção acarreta, mas também pelo facto de não ser exequível a curto prazo, dado que muitas das vezes encontrar mais recursos humanos só é possível através da formação de novos profissionais de saúde.

No que diz respeito à segunda opção, uma das fontes de ineficiência prende-se com a incorreta estimativa da duração de uma cirurgia, sendo que a má previsão das durações conduz a diversos problemas. Em muitos hospitais, os métodos utilizados para prever a duração de uma cirurgia são simples e assentam, geralmente, numa estimativa indicada pelo cirurgião ou num valor médio, não fazendo uso de técnicas quantitativas mais avançadas que utilizem informação baseada em dados históricos. Daí surge cada vez mais a necessidade de se desenvolverem modelos estatísticos que tenham a capacidade de proporcionar uma maior qualidade no planeamento e agendamento de cirurgias. Assim, o objetivo deste trabalho final de mestrado (TFM) é desenvolver modelos de previsão de modo a que, através dos mesmos, o planeamento de cirurgias possa ser feito utilizando uma duração o mais aproximada à realidade quanto possível, diminuindo a diferença entre o tempo estimado e o tempo real das cirurgias. Será ainda feita uma comparação entre um modelo clássico de Regressão Linear Múltipla (RLM) e um modelo de *machine learning* (ML), o XGBoost, para determinar se, no caso a ser tratado, o modelo de ML justifica a adicional complexidade computacional.

Os dados obtidos para a previsão de cirurgias são providenciados pelo HESE, sendo esta a principal unidade hospitalar do território alentejano que recebe utentes de praticamente todas as zonas do Alentejo, as quais contabilizam um total de 704.707 habitantes, segundo resultados provisórios dos CENSOS 2021 [3].

Por fim, este TFM divide-se da seguinte forma:

- O Capítulo 2 apresenta uma breve revisão de literatura;
- O Capítulo 3 apresenta um resumo da metodologia utilizada no TFM;
- O Capítulo 4 foca-se no processo de seleção e preparação dos dados;
- O Capítulo 5 apresenta o modelo em RLM, abordando conceitos essenciais como a verificação de pressupostos e seleção de variáveis, e respetivos resultados;
- O Capítulo 6 contém o modelo XGBoost, no qual são apresentados os diversos parâmetros utilizados, a otimização dos mesmos e os resultados obtidos;
- No Capítulo 7 é feita uma comparação dos modelos propostos;
- As conclusões retiradas são apresentadas no Capítulo 8.

[3] CENSOS, 2021. Instituto Nacional de Estatística. Disponível em: https://censos.ine.pt/scripts/db_censos_2021.html. Consultado a 23 de janeiro de 2022.

CAPÍTULO 2 – REVISÃO DE LITERATURA

O planeamento eficaz do BO é uma atividade que contribui para a eficiência e qualidade do serviço, tanto pela qualidade do tratamento que é providenciado aos pacientes como pelo tempo de espera para receberem determinados cuidados (Yuniartha et al., 2021). A previsão da duração de cirurgias tem como objetivo apoiar o planeamento do BO, que é dividido em três níveis de decisão: estratégico, tático e operacional (Cardoen et al., 2010b). Esta segmentação do planeamento pode não garantir um nível de detalhe adequado, pelo que outras taxonomias podem ser propostas consoante características mais especificas dos problemas tratados. Cardoen et al. (2010a) sugerem uma diferenciação mais completa, onde consta a categoria da incerteza. É possível distinguir duas fontes distintas de incerteza: a primeira, a incerteza de chegada, que agrupa problemas relacionados com a imprevisibilidade da chegada de pacientes de emergência ou questões como atrasos dos cirurgiões; e a segunda, a incerteza da duração da cirurgia, que analisa os desvios entre a duração real e a duração prevista da cirurgia usada no planeamento do BO. Assim, a previsão de durações de cirurgias realizada neste TFM enquadra-se no nível de planeamento operacional dado ser fundamental para o eficaz agendamento de cirurgias (Nawaz Ripon e Henrik Nyman, 2020), mais especificamente na categoria da incerteza da duração.

Havendo incerteza na duração da cirurgia, podem ocorrer alterações ao nível do planeamento do BO. Segundo Nawaz Ripon e Henrik Nyman (2020), este problema apresenta uma dificuldade acrescida uma vez que, mesmo que o planeamento se mostre excelente, as expectativas poderão mostrar-se desadequadas na presença de cirurgias atrasadas ou adiantadas. Deste modo, se por um lado a sobrestimação da duração de cirurgias produz uma menor utilização e rendimento do BO até o próximo paciente estar pronto para cirurgia, por outro, a sua subestimação leva ao atraso ou ao cancelamento de cirurgias posteriores e origina custos adicionais derivados de horas extra não planeadas (Tan et al., 2019).

Apesar de uma boa previsão da duração de cirurgia ajudar a evitar os problemas acima mencionados, esta tarefa torna-se especialmente difícil dado que são utilizados dados históricos para a mesma e, tendo em conta que alguns procedimentos realizados são raros, a obtenção desta informação torna-se desafiante (Kayış et al., 2015). Outra questão que dificulta a previsão está relacionada com as variáveis a utilizar. Para além dos fatores que mais contribuem para uma boa previsão variarem consoante as cirurgias e os pacientes, alguns só são conhecidos depois da cirurgia ser terminada e em nada vão contribuir para o modelo de previsão (Devi et al., 2012).

Frequentemente, verificam-se grandes desvios entre a duração real e a duração prevista, uma vez que os métodos utilizados pelos hospitais para efetuar estas previsões passam pelo cálculo de um valor médio utilizando as durações de cirurgias anteriores ou dependem de sistemas de previsão mais subjetivos

e que acabam por decorrer da capacidade do cirurgião fazer uma previsão correta (Dexter et al. 2008). Ao estar dependente de um cirurgião, este tipo de metodologia acaba por falhar em grande parte dos casos porque certos detalhes podem não ser conhecidos ou não ser tidos em consideração quando a cirurgia é planeada, o que origina alterações da duração após planeamento (Dexter et al. 2008). De facto, segundo Laskin et al. (2013), em 100 casos analisados, os cirurgiões conseguiram prever corretamente apenas 26% das durações de cirurgias, enquanto nos restantes 74%, em 42% dos casos houve sobrestimação das durações enquanto em 32% ocorreu subestimação.

Na literatura têm vindo a ser descritos outros métodos computacionais de ML que têm o potencial de dar origem a melhores previsões através de fatores temporais, da cirurgia, do paciente e da equipa médica presente na cirurgia (Tan et al., 2019; Zhao et al. 2019). Em destaque estão os modelos de ML pertencentes à categoria de *ensemble* que demonstram melhorar significativamente os resultados em termos de precisão da previsão (Shahabikargar et al., 2017).

Na Tabela 1 pode-se observar um resumo da literatura consultada, no que diz respeito aos modelos de previsão usados, ao nível de agregação dos dados e aos indicadores de desempenho utilizados para avaliar os modelos.

Tabela 1 - Resumo das características consideradas na literatura.

	Modelos de Previsão	**Nível de Agregação dos Dados**	**Indicador de Desempenho**
Devi et al. (2012)	RLM e Redes Neuronais	Por Procedimento	$RMSE$
Hosseini et al. (2015)	RLM e *StepwiseRegression*	Por Especialidade	R^2, $RMSE$ e MAE
Shahabi Kargar et al. (2017)	RLM, *RandomForest* e Métodos de *Ensemble*	Por Especialidade	$MAPE$
Riekert et al. (2017)	RLM, *RandomForest* e *SupportVectorRegression*	Por Procedimento	R^2
Zhao et al. (2019)	RLM, *RandomForest*, XGBoost e Redes Neuronais	Dados Agregados	$RMSE$
Bartek et al. (2019)	RLM e XGBoost	Por Especialidade	$MAPE$
Martinez et al. (2021)	RLM e Métodos de *Ensemble*	Por Especialidade	$RMSE$ e MSE

Na segunda coluna encontra-se identificado o modelo usado para a previsão da duração de cirurgia, na qual se verifica a utilização da RLM como modelo base de comparação em todos os artigos. O XGBoost mostra-se em destaque por melhorar de forma significativa a precisão dos resultados em comparação com outros modelos. Apesar disto, não é aconselhado um modelo de previsão específico neste tipo de estudos,

uma vez que é impossível saber qual o algoritmo que irá obter o melhor desempenho num determinado conjunto de dados (Kurz et al., 2020).

A terceira coluna da tabela é referente ao facto de que a grande maioria dos estudos sobre o tema da previsão de durações de cirurgias, não realiza apenas uma previsão com a totalidade dos dados, uma vez que como mencionado por Zhu et al. (2019), a duração está dependente da especialidade associada. Segundo Hosseini et al. (2015), cada especialidade tem um determinado conjunto de variáveis que lhe são especificas, ou seja, colocar todas essas variáveis num único modelo vai acabar por desfavorecer determinadas especialidades que poderão ter menos cirurgias. Assim, cada artigo habitualmente dá origem a modelos individuais para cada procedimento ou especialidade.

A quarta coluna identifica os indicadores utilizados para avaliar os modelos, sendo que por norma, é utilizado mais do que um, dado que diferentes indicadores transmitem diferentes informações e utilizar apenas um reduz significativamente a compreensão dos resultados obtidos. Segundo Sutherland et al. (2004) a utilização de medidas de diferença entre médias são as mais apropriadas para a avaliação de um modelo e podem ser obtidas através de indicadores como o *Mean Absolute Error*, *Mean Squared Error* (MSE) e *Root* MSE (RMSE). Desta forma, para além do R^2, que permite avaliar o poder explicativo do modelo, nos artigos analisados são utilizados frequentemente indicadores como o RMSE, que permite determinar o erro médio na unidade da previsão, e o *Mean Absolute Percentage Error*, que permite obter o erro médio em percentagem facilitando a interpretação.

Outra questão analisada na literatura é quais variáveis devem ser utilizadas na previsão. Vários estudos tentaram determinar quais contribuem para uma previsão mais precisa, no entanto, e segundo Iroju et al. (2013), as variáveis disponíveis variam de caso para caso dado que não existe um conjunto de dados uniforme entre hospitais. Nos estudos desenvolvidos por Tan et al. (2019) e Bartek et al. (2019) foi determinado que as variáveis mais importantes consistiam em características do ambiente e da equipa cirúrgica, atribuindo menor utilidade às variáveis relacionadas com o próprio paciente. Por outro lado, em Ng et al., (2017), observamos um conjunto de variáveis significativas que, para além da variável correspondente ao cirurgião que realiza a cirurgia, inclui ainda características do paciente, como o sexo e o estado em que se encontra. Assim, torna-se evidente a constatação feita por Iroju et al. (2013).

Ao longo deste capítulo foram referidos vários artigos que apresentam diferentes metodologias no processo de previsão de durações de cirurgias. Assim, baseado nas boas práticas adotadas na literatura acerca do tema, são desenvolvidos neste TFM modelos de RLM e de XGBoost, tanto com um nível de agregação dos dados por especialidade ou não, de forma a obter a melhor previsão possível. Posteriormente, esses modelos são submetidos a uma avaliação através dos indicadores de desempenho R^2 e RMSE, e comparados entre si.

Apesar da literatura relativa à previsão de durações de cirurgias já incluir alguns estudos, como demonstrado, ainda não é claro em que situações certos modelos poderão ser mais adequados do que outros e, para além disso, que variáveis serão mais importantes para obter previsões mais precisas. Desta forma, este TFM contribui para a literatura de previsão de durações de cirurgias ao comparar e avaliar diferentes modelos num contexto hospitalar real.

CAPÍTULO 3 – METODOLOGIA

Este capítulo apresenta a metodologia utilizada para a previsão das durações de cirurgias, que se encontra representada na Figura 2. A estrutura da metodologia passa por quatro fases distintas que facilitam a compreensão do processo: Seleção de Informação; Preparação de Dados; Modelação; e Resultados.

A Seleção de Informação consiste na seleção dos ficheiros essenciais para a previsão de durações de cirurgias e corresponde à fase de seleção dos dados brutos – Data Set 0 – e posterior organização, que dá origem ao Data Set 1. Esta primeira fase do TFM encontra-se descrita no Capítulo 4, na Subsecção 4.1.

De seguida, é feita a descrição de todo o processo de Preparação de Dados, que inclui a limpeza dos mesmos e a seleção de variáveis que serão potenciais variáveis independentes úteis para os modelos de previsão. Esta fase é descrita nas Subsecções 4.2 e 4.3. O processo descrito culmina num novo conjunto de dados – Data Set 2.

A próxima fase é a Modelação, na qual vamos aplicar os modelos de previsão escolhidos – RLM e XGBoost. Assim, esta fase é descrita individualmente para cada modelo, dado que se aplicam processos diferentes.

No caso da RLM, são mencionados detalhes acerca da validação de pressupostos e da respetiva transformação dos dados necessária, que origina a produção do Data Set RLM Agregado e do Data Set RLM por Especialidade, da seleção de variáveis e da análise do modelo final. No Capítulo 5 encontra-se toda a informação anteriormente mencionada nas Subsecções 5.2, 5.3 e 5.4, respetivamente.

A informação acerca do XGBoost pode ser encontrada no Capítulo 6. É descrita a metodologia adotada para o modelo de XGBoost na Subsecção 6.3, sendo analisados os modelos finais obtidos para os dois conjuntos de dados (Data Set XGB Agregado e Data Set XGB por Especialidade) na Subsecção 6.4.

Por último, os Resultados serão analisados e discutidos no Capítulo 7, no qual é feita a comparação dos resultados obtidos com os modelos de previsão de RLM e XGBoost.

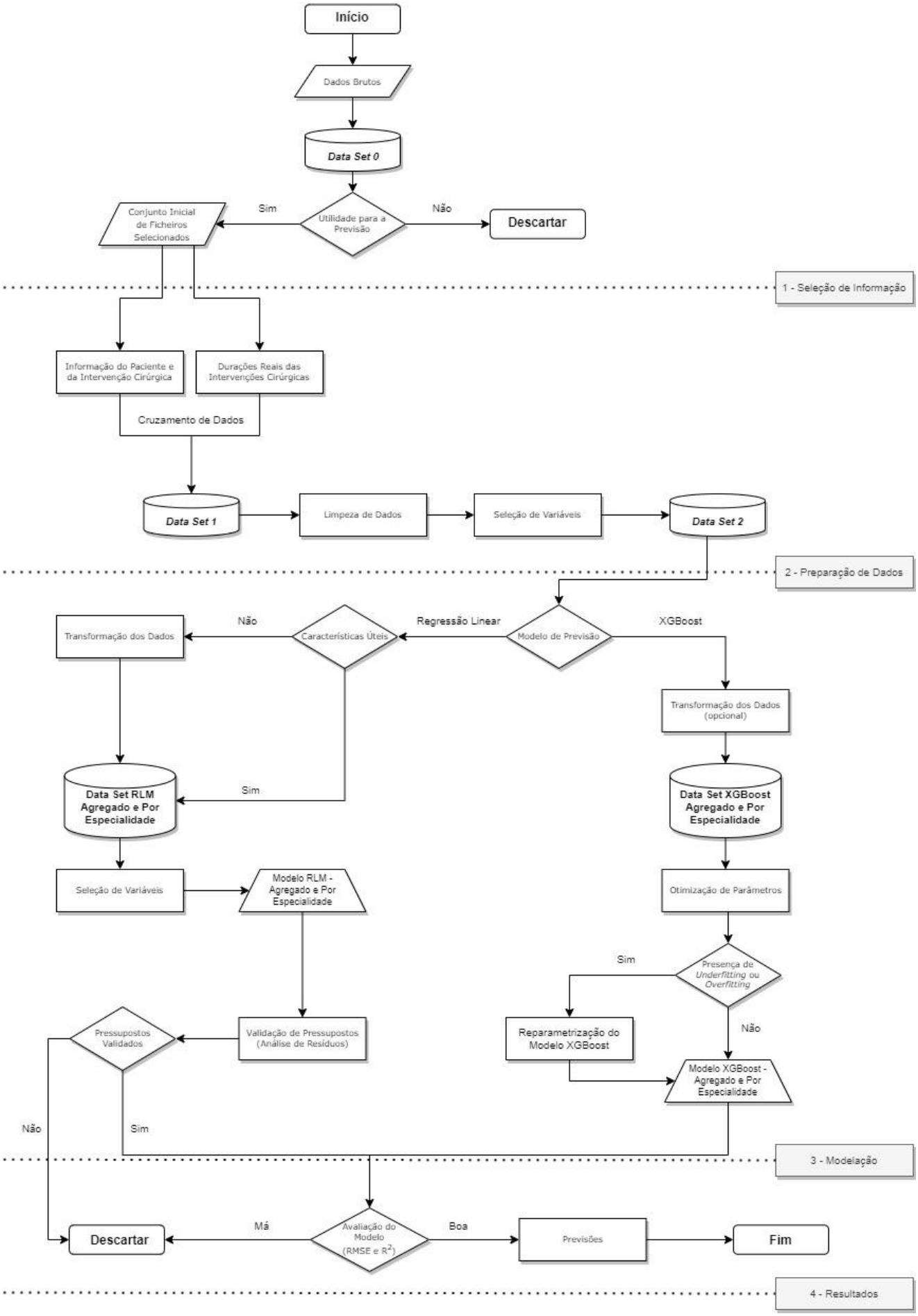


Figura 2 - Esquema da metodologia.

CAPÍTULO 4 – SELEÇÃO E PREPARAÇÃO DE DADOS

4.1 Seleção de Informação e Cruzamento de Dados

Dos ficheiros de dados providenciados em formato *csv* pelo HESE, correspondentes ao Data Set 0, foram selecionados apenas os que continham informação relevante para a previsão de duração de cirurgias. Assim, obtiveram-se dois ficheiros de dados (Ficheiro 1 e Ficheiro 2), que contêm a informação apresentada na Figura 3.

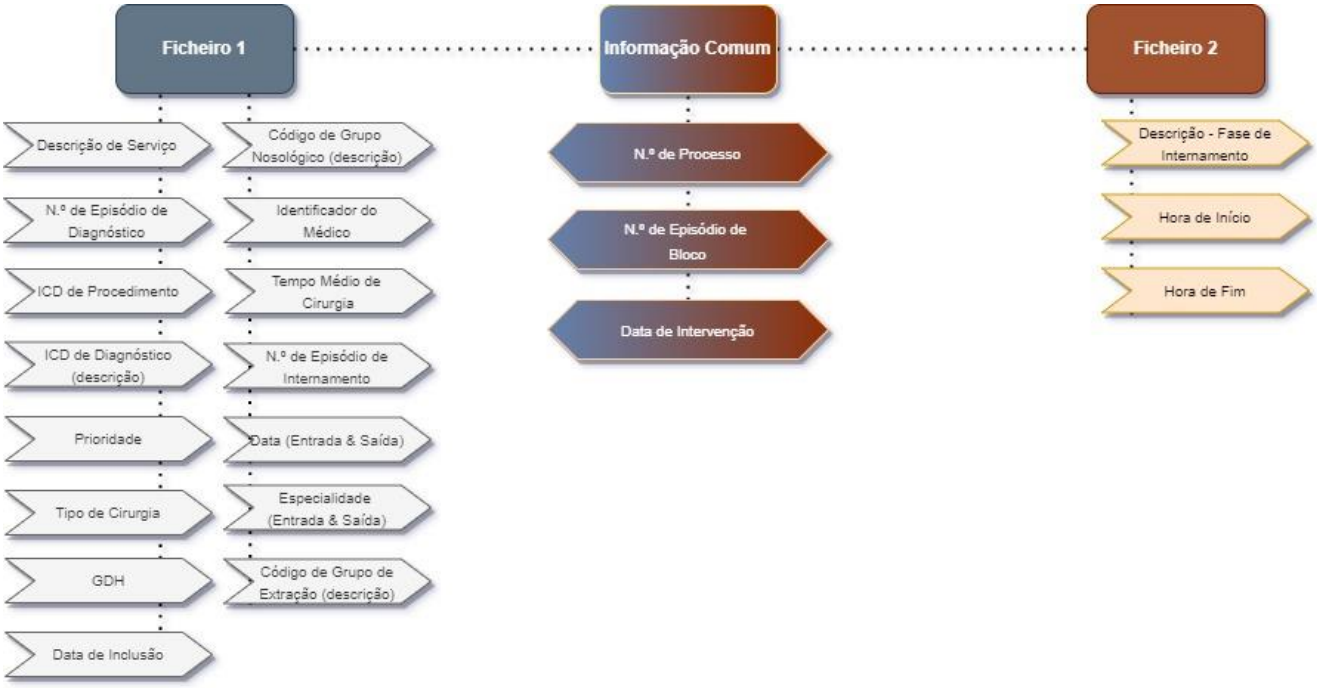


Figura 3 - Resumo da informação existente em cada ficheiro.

O Ficheiro 1 contém informação relativa a diversos fatores relacionados com o procedimento realizado, nomeadamente:

1. O diagnóstico e o procedimento cirúrgico – *International Classification of Diseases* (ICD) de Procedimento, ICD de Diagnóstico, Grupos de Diagnósticos Homogéneos (GDH), Código de Grupo Nosológico e Código de Grupo de Extração e as respetivas descrições do significado de cada um dos códigos mencionados;
1. Características do procedimento – Prioridade, Tipo de Cirurgia (Convencional ou Ambulatório) e Tempo Médio de Cirurgia;
2. Datas - Data de Entrada e Saída do BO e Data de Inclusão em LIC;
3. Informação referente à Especialidade associada a cada cirurgia – Descrição de Serviço, Especialidade de Entrada e Especialidade de Saída;
4. Identificador do médico proponente da cirurgia realizada – Identificador do Médico;

5. Identificadores únicos – N.º de Episódio de Diagnóstico e N.º de Episódio de Internamento (não existente para cirurgias de ambulatório).

O Ficheiro 2 contém a informação relacionada com as durações reais das cirurgias, mais precisamente:

1. A fase do internamento, ou seja, Anestesia, Cirurgia, Bloco ou Sala;
2. Duração real de cada uma das fases de internamento – Hora Início e Hora Fim;

Depois da seleção dos ficheiros segue-se a preparação dos dados, que incorpora o processo de limpeza e de seleção de variáveis que poderão ser úteis para a previsão da duração de cirurgias.

No conjunto inicial dos ficheiros referidos havia informação que poderia ser cruzada de maneira a obter o conjunto primário de variáveis num só ficheiro. Através de identificadores únicos comuns a ambos os ficheiros é possível combinar a informação existente nos dois ficheiros. Para isto, utilizou-se o identificador N.º de Episódio de Bloco, uma vez que podem existir registos duplicados de N.º de Processo, indicando que no ano de 2017 o mesmo paciente teria sido submetido a mais do que uma cirurgia.

Utilizando o N.º de Episódio de Bloco obteve-se a Hora de Início e a Hora de Fim de cada uma das cirurgias e, posteriormente, a duração real das mesmas através da diferença entre a respetiva hora de fim e hora de início. Assim, numa fase inicial obteve-se registos para as quatro durações (duração de anestesia, duração de cirurgia, duração no bloco e duração em sala) obtidas para cada identificador único de entrada no BO.

4.2 Preparação de Dados

A base de dados original, obtida após combinar a informação existente nos Ficheiros 1 e 2 e correspondente ao Data Set 1, contava com 7.696 registos de cirurgias, sendo que 952 foram eliminados dado que se tratavam de registos sem a variável que se encontra em estudo – Duração da Cirurgia – ou registos nos quais a especialidade à qual pertenciam não estava devidamente identificada.

Depois de eliminados os registos anteriormente referidos, analisou-se ainda o número de cirurgias em cada especialidade, onde se verificou que algumas teriam menos de 30 cirurgias. Um número de intervenções reduzido, compromete a fiabilidade da previsão das durações de intervenções cirúrgicas por especialidade. Sendo assim, procedeu-se à exclusão dos registos das especialidades com um número de procedimentos cirúrgicos inferior a 30, nomeadamente, Cardiologia e Estomatologia, como se verifica na Figura 4, o que totalizou uma remoção de 48 registos adicionais.

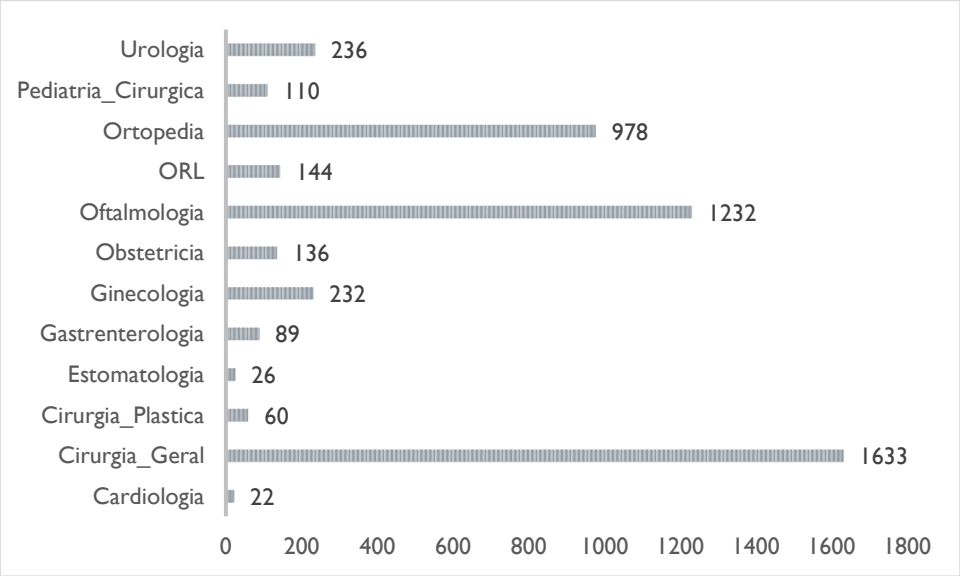


Figura 4 - N.º de intervenções por especialidade.

A Figura 5 mostra a distribuição da duração das cirurgias, na qual tem-se no eixo das ordenadas o número de observações e no das abcissas a duração das cirurgias em minutos.

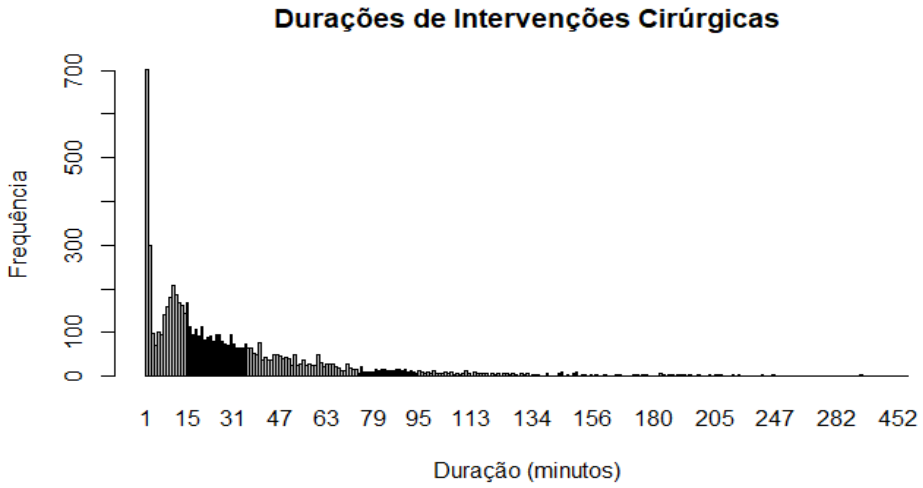


Figura 5 - N.º de intervenções por duração de cirurgia (minutos).

A média da duração das cirurgias é 35,34 minutos, sendo que a mediana é de 20 minutos. Conseguimos observar que cerca de 700 cirurgias têm uma duração de um minuto, o que pode ser indicador de possíveis erros de registo nos dados fornecidos. Assim, determinou-se que apenas seriam incluídos registos de cirurgias com durações superiores a 10 minutos e inferiores a 1.000 minutos (equivalente a 16,67 horas). Esta pré-seleção de registos é comum na literatura, como nos estudos de Ng et al. (2017) e Bartek et al. (2019), e pretende evitar incluir na previsão possíveis erros de registo do próprio hospital que possam influenciar negativamente os modelos de previsão. Desta forma, foram eliminados 1.846 registos de cirurgias.

Cerca de 37% dos registos recebidos relativos ao ano de 2017 foram eliminados durante o processo de Preparação de Dados. É ainda de salientar que os dados foram organizados e tratados de maneira a que não houvessem registos duplicados ou registos que não apresentassem informação válida para a totalidade das variáveis.

Resumidamente, como anteriormente foi descrito, foram eliminados os seguintes registos na fase de Limpeza de Dados:

1. Registos sem a duração real das cirurgias;
2. Registos nos quais havia dúvidas acerca da Especialidade na qual estavam inseridos;
3. Registos de especialidades com menos de 30 cirurgias;
4. Registos de cirurgias com menos de 10 minutos e com mais de 1.000 minutos de duração.

Desta forma, partimos de uma base de dados com um total de 4.850 cirurgias realizadas no ano de 2017 em 4.041 pacientes e em 10 especialidades distintas: Cirurgia Geral; Ortopedia; Oftalmologia; Ginecologia; Obstetrícia; Urologia; Cirurgia Plástica; Otorrinolaringologia (ORL); Pediatria Cirúrgica e Gastroenterologia.

4.3 Seleção de Variáveis

Após a preparação de dados, foram analisadas todas as possíveis variáveis independentes e mantidas apenas as que tivessem relevância para a previsão da duração das cirurgias.

Numa primeira fase foram retiradas as variáveis que não teriam qualquer influência na duração da cirurgia e que apenas faziam parte do conjunto de dados recebidos. Estas são os identificadores únicos, N.º de Episódio de Diagnóstico e N.º de Episódio de Internamento, que apenas identificam a entrada no processo de diagnóstico e internamento, respetivamente, ou as descrições de alguns códigos essenciais para a identificação universal de determinados procedimentos e diagnósticos. Assim, eliminaram-se as variáveis apresentadas na Tabela 2.

Tabela 2 - Lista de variáveis excluídas na primeira fase de seleção de variáveis.

Variáveis Excluídas
N.º de Processo
N.º de Episódio
N.º de Episódio de Bloco
N.º de Episódio de Internamento
Descrição Procedimento
Descrição Diagnóstico
Descrição Código de Extração
Descrição Grupo Nosológico

Adicionalmente, apenas se teve em conta variáveis que pudessem ser determinadas antes da cirurgia dado que, caso só fossem conhecidas depois de realizada a cirurgia, o modelo obtido com essas variáveis não seria útil. Desta forma, as variáveis Duração de Anestesia, Duração em Sala e Duração em Bloco foram excluídas do conjunto de dados.

Acrescentaram-se três médias de duração das cirurgias que poderiam ter influência na previsão a realizar: Média por Médico; Média por Prioridade; e Média por ICD de Procedimento. De seguida, também se considerou a variável Tempo de Espera - calculada através da diferença entre a Data de Inclusão (data de entrada em LIC) e a Data de Cirurgia (data de realização da cirurgia) – tal como a variável binária que identifica a existência de internamentos anteriores no ano de 2017, podendo evidenciar complicações cirúrgicas e, consequentemente, durações cirúrgicas maiores. Também foi adicionada a variável binária que identifica se a cirurgia foi realizada em dia útil ou não. Por fim, existem comorbilidades que têm uma relação direta com a duração da cirurgia, como a obesidade (Larsson, 2013). Esta justificar possíveis complicações não previstas durante a cirurgia, aumentando a duração esperada inicialmente. Como os dados recebidos continham informação acerca de quais pacientes têm obesidade, foi então criada a variável binária OBESIDADE_C.

Excluíram-se também variáveis usadas para a produção de outras, pois não são independentes entre si, como seria o caso das seguintes:

- Data de Cirurgia e Data de Inclusão;
- Início Cirurgia, Fim Cirurgia, Início Anestesia, Fim Anestesia, Início Bloco, Fim Bloco, Início Sala e Fim Sala;
- Especialidade Entrada e Especialidade Saída.

Após os passos referidos obteve-se o Data Set 2, e os nomes das variáveis foram alterados de maneira a simplificar o processo de análise dos modelos obtidos. Assim, a primeira listagem das variáveis importadas em RStudio está representada na Tabela 3, para se dar início à etapa de seleção de variáveis. É importante ainda salientar que nos próximos capítulos do TFM as variáveis serão denominadas pelo nome indicado na coluna Nome RStudio de maneira a facilitar a leitura. O conjunto final das variáveis selecionadas são classificadas e apresentadas também na Tabela 3, em duas categorias, nomeadamente:

- Categoria C - variáveis categóricas;
- Categoria N - variáveis numéricas.

Tabela 3 - Lista das variáveis selecionadas e respetiva conversão do Nome Original para o Nome em RStudio.

	Nome RStudio	Tipo de Variável	N.º de Categorias
Internamento Anterior	INT_ANTERIOR	C	2
ICD de Procedimento	ICD_PROC	C	357
ICD de Diagnóstico	ICD_DIAG	C	504
Prioridade	C_PRIORIDADE	C	4
Tempo de Espera (dias)	T_ESPERA	N	N.A.
Semana	SEMANA_C	C	7
Mês	MES_C	C	12
Fim de Semana ou Feriado	NUTEIS_C	C	2
Tipo de Cirurgia	C_CIRURGIA	C	2
GDH	GDH	C	70
Código de Extração	COD_EXTRACAO	C	60
Código do Grupo Nosológico	COD_GNOSOLOGICO	C	27
Obesidade	OBESIDADE_C	C	2
Especialidade	ESPECIALIDADE_C	C	11
Duração de Cirurgia	D_CIRURGIA	N	N.A.
Média por Médico	MEAN_MEDICO	N	N.A.
Média por Prioridade	MEAN_PRIORIDADE	N	N.A.
Média por ICD de Procedimento	MEAN_ICDPROC	N	N.A.

N.A.: Não Aplicável

Para cada variável classificada como categórica foi indicado o número de categorias presentes. Como é evidente, para variáveis numéricas, e dado que as mesmas não apresentam qualquer tipo de categorização nos seus valores, foi considerada a terceira coluna como Não Aplicável (N.A.). Posto isto, temos 18 variáveis, das quais 13 são categóricas e as restantes cinco são numéricas. No caso das variáveis categóricas, sabemos ainda que o número de categorias varia entre duas, correspondente às variáveis independentes binárias (INT_ANTERIOR, NUTEIS_C, C_CIRURGIA, OBESIDADE_C), e 504 categorias da variável independente ICD_DIAG. Temos ainda a variável D_CIRURGIA que representa a variável dependente em estudo.

Tal como já foi referido, no tema das previsões de durações de cirurgias o nível de agregação dos dados é um elemento importante, sendo frequente que as mesmas sejam feitas em dois cenários distintos: um primeiro conjunto de dados agregados; e um segundo conjunto de dados desagregados por categorias que poderão ser úteis na previsão ao serem agrupadas em modelos independentes, como por procedimento ou por especialidade. Apesar de na literatura serem realizadas previsões por procedimento (Hosseini et al., 2015; Riekert et al., 2017), no caso deste TFM tal não se torna possível, dado que existem muitos tipos de procedimentos distintos e cada um com poucas observações. Logo, como alternativa, é feita uma análise de modelos com dados desagregados por especialidade. Assim, a aplicação dos modelos de RLM e XGBoost vai ser efetuada em dois conjuntos de dados:

- Um conjunto de dados constituído pela totalidade dos mesmos;
- 10 conjuntos de dados, cada um associado a uma especialidade, que contêm apenas as observações dessa mesma especialidade.

CAPÍTULO 5 – REGRESSÃO LINEAR MÚLTIPLA

5.1 Introdução à RLM

5.1.1 RLM

A aplicação de modelos de RLM permite obter resultados mais interpretáveis do que os modelos de ML. Assim, a RLM será o modelo base com a qual o modelo de ML será comparado, sendo o ponto de partida do TFM.

A RLM é uma extensão da regressão linear simples com diversas variáveis independentes x_k $(k = 1, 2, \dots, p)$ e apenas uma variável dependente y_i $(x_{i1}, x_{i2}, \dots, x_{ip}, y_i)$ (Eberly, 2007), sendo representada pela seguinte expressão:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i \text{ , com } i = 1, 2, \dots, n.$$

O n representa a dimensão da amostra, ε_i é uma variável aleatória que representa o termo do erro e, por fim, os β_k , com $k = 0, 1, 2, \dots, p$, os parâmetros do modelo.

Os modelos de RLM admitem como válidos o seguinte conjunto de pressupostos:

- Distribuição normal de ε – Os erros aleatórios devem seguir uma distribuição normal com valor médio 0 e variância constante, ou seja,

$$\varepsilon_i \cap N(0, \sigma^2) \text{ com } i = 1, \dots, n.$$

- Independência – Os erros aleatórios devem mostrar-se independentes entre si, isto é,

$$Cov\left(\varepsilon_i, \varepsilon_j\right) = 0 \text{ com } i, j = 1, \dots, n;\ i \neq j.$$

5.1.2 Validação de Pressupostos

A validação dos pressupostos, através da análise de resíduos anteriormente mencionada, apenas é realizada após a obtenção do modelo final de RLM. No entanto, existem outras características que podem ser garantidas, antes do processo de seleção de variáveis, e que acabam por ser úteis para a correta utilização dos modelos de RLM:

1. Distribuição Normal da variável dependente;
    - A normalidade da variável aleatória ε implica que a variável dependente y também tenha distribuição Normal. Assim, caso a normalidade da variável dependente não se verifique, a distribuição dos erros aleatórios também não será Normal e, portanto, um dos pressupostos da RLM não é verificado.
    - Pode recorrer-se à representação gráfica da variável aleatória dependente através de um histograma e verificar-se se é semelhante à função densidade de probabilidade da Normal.
2. Multicolinearidade;

- A não verificação desta característica conduz a uma redução na qualidade das estimativas dos mínimos quadrados obtidas pois aumenta a variância entre parâmetros (Belsley, 1982).
- Pode ser verificada através da análise do *Variance Inflation Factor* (VIF). Temos $VIF = \frac{1}{(1- R_j^2)}$, no qual R_j^2 é o R^2 obtido quando se faz uma RLM das variáveis independentes x_i com $i = 1, \dots, j-1, j+1, \dots, n$ em função da variável independente j, sendo que esta passa a ser a variável dependente do novo modelo. Belsley (1982) sugere que não existe um valor de VIF pré-definido para o qual é comprovada a existência de multicolinearidade, variando de modelo para modelo. Contudo, este não deve ser muito elevado.

Depois de ajustado o modelo de RLM ao conjunto de dados, estamos em condições de fazer a análise de resíduos, que vai evidenciar a variabilidade da variável dependente que não foi explicada pelo modelo. Assim, devem ser validados os seguintes pressupostos:

3. Distribuição Normal dos erros aleatórios;
    - Para avaliar este pressuposto, por um lado, pode recorrer-se à representação dos resíduos estandardizados numa comparação da função empírica de distribuição cumulativa dos dados (o chamado *ppplot*), por outro podemos analisar a representação de um gráfico de quartis da distribuição Normal(0,1) (*qqplot*).
4. Resíduos homocedásticos;
    - Estes pressupostos podem ser verificados ao avaliar um gráfico de dispersão dos resíduos.
5. Independência.
    - Devemos verificar a correlação entre os valores dos resíduos, dado que a correlação entre variáveis se vai refletir nos mesmos. Para isto vai ser utilizado o teste de Durbin-Watson, que testa as seguintes hipóteses:

H0: Não existe correlação entre os erros aleatórios

vs.

H1: Existe correlação entre os erros aleatórios

5.1.3 Seleção de Variáveis

5.1.3.1 ANOVA

Relativamente ao processo de seleção de variáveis, o teste da ANOVA é usado quando a variável dependente é quantitativa e se verifica a existência de variáveis independentes categóricas. Segundo

Tabachnick & Fidell (2007), este teste é fundamental na análise de variáveis categóricas pois permite determinar se as médias entre as populações, que são definidas pelas várias categorias que a variável pode tomar, diferem entre si. Mostra-se ainda equivalente a uma regressão linear na qual todas as variáveis independentes são variáveis categóricas.

Posto isto, o teste da ANOVA aplicado às variáveis categóricas tem como hipóteses:

H0: As médias das durações das cirurgias são iguais para todas as categorias

vs.

H1: Existem pelo menos duas categorias tais que as médias das durações das cirurgias diferem entre si

Caso o *p-value* observado seja inferior a um determinado nível de significância existe evidência estatística para afirmar que existem diferenças significativas nas durações médias de cirurgia entre as diversas categorias, ou seja, a variável é estatisticamente significativa para a previsão destas durações e, portanto, faz sentido incluí-la no modelo. Caso não fosse rejeitada a hipótese nula, não existiria evidência estatística para afirmar que existam diferenças significativas nas durações de cirurgias entre as categorias e a variável pode ser excluída do modelo.

5.1.3.2 Backward Stepwise Regression

Determinadas as variáveis categóricas a utilizar, verifica-se a significância estatística das variáveis independentes binárias e numéricas através da *Backward Stepwise Regression*. Segundo Thayer (2002), a *Backward Stepwise Regression* forma um modelo de previsão "de cima para baixo" de acordo com o seguinte algoritmo:

Passo 1: Construir o modelo de RLM com p variáveis independentes.

Passo 2: Determinar a variável com o maior *p-value*.

Se: *p-value* superior ao α_s pré-determinado, ir para Passo 3.

C.c: Fim.

Passo 3: Remover a variável com o maior *p-value*. Fazer $p = p - 1$ e voltar ao Passo 1.

Este algoritmo acaba assim que for respeitado o critério de paragem para as restantes variáveis, ou seja, quando todas as variáveis no modelo forem significativas tendo em conta um nível de significância escolhido de forma a adequar-se aos objetivos do estudo. De acordo com Pasha (2002), o mais habitual é considerar $\alpha_s = 0{,}05$.

5.2 Metodologia

Como referido anteriormente, partindo do Data Set 2 vamos obter um conjunto de dados agregados (Data Set RLM Agregado) e um segundo conjunto de dados, que é composto por vários conjuntos de dados,

para cada uma das especialidades (Data Set RLM por Especialidade). Para ambos será aplicado o processo descrito na Figura 6.

Figura 6 - Metodologia adotada nos modelos de RLM.

Após a verificação das características úteis mencionadas, nomeadamente a distribuição normal de y e a ausência de multicolinearidade entre variáveis, estamos em condições de iniciar o processo de seleção de variáveis. É importante acrescentar que no conjunto de dados Data Set RLM por Especialidade, para cada especialidade é feita uma nova seleção de variáveis, uma vez que não podemos garantir que para cada especialidade exista a mesma significância estatística entre variáveis.

Para verificar se as variáveis categóricas são estatisticamente significativas ou não, é utilizado o teste da ANOVA (Martinez et al., 2021). A utilização deste teste torna-se imprescindível em situações nas quais se está perante variáveis categóricas com um elevado número de categorias, como mencionado anteriormente. O elevado número de categorias não permite uma análise da sua significância através do comando *lm* do RStudio, dado que este, por omissão, divide cada uma das categorias em *dummys*, logo, obtemos um *p-value* para as diversas categorias em oposição a um único *p-value* associado à variável categórica e é colmatado o problema inicial. Neste caso, o teste foi aplicado a todas as variáveis independentes categóricas e consideram-se excluídas as variáveis com *p-value* maior ou igual a 5%.

Após a aplicação da ANOVA é necessário analisar se as variáveis categóricas anteriormente testadas têm fortes correlações entre si tendo em conta que existem várias variáveis que contém informação semelhante. Para isso, observa-se o R^2 dos modelos com todas as variáveis e compara-se com o R^2 do modelo obtido ao excluir as variáveis que poderão estar correlacionadas. Desta forma, podemos determinar se as variáveis estão correlacionadas, pois caso estejam a sua exclusão do modelo pouco irá alterar o valor do R^2.

Analogamente, são também excluídas as variáveis categóricas que são linearmente dependentes de outras e ainda variáveis que apresentam a mesma categoria para todas as observações. No entanto, isto apenas se verifica nos modelos de RLM para o Data Set RLM por Especialidade, dado o reduzido número de observações em comparação com o Data Set RLM Agregado.

Depois da exclusão das variáveis correlacionadas, com coeficientes "NA" e com apenas uma categoria, segue-se a etapa da *Backward Stepwise Regression*. Este método de seleção de variáveis tem início com um modelo com as variáveis numéricas, binárias e categóricas mantidas no passo anterior. De seguida, é aplicado o processo descrito na Subsecção 5.1.3.2.

Utilizando o modelo final de RLM obtido vai ser feita uma análise de resíduos de forma a verificar o segundo conjunto de pressupostos, e assim pode-se garantir que as inferências feitas acerca do modelo são válidas, caso contrário, o modelo é descartado.

5.3 Construção do Modelo

5.3.1 Data Set RLM Agregado

5.3.1.1 Características Úteis

A Figura 7, apresenta um histograma da duração de cirurgia, consistindo numa representação gráfica de uma distribuição de frequências.

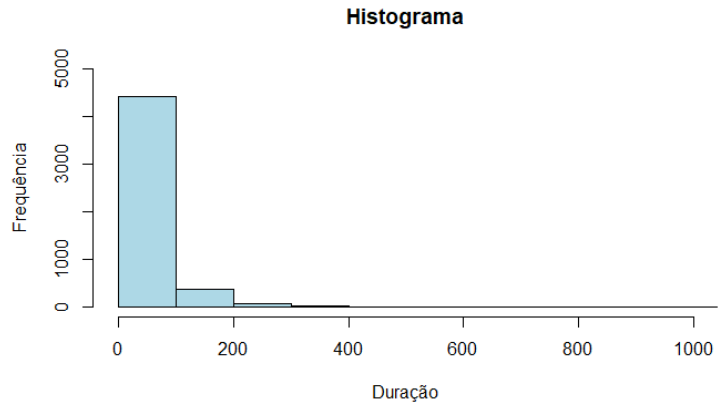


Figura 7 - Histograma da distribuição da variável dependente.

Conseguimos observar que a variável dependente apresenta uma distribuição com enviesamento à esquerda. Isto não é recomendado tendo em conta o pressuposto de normalidade dos erros aleatórios, que é essencial para a validade dos testes de hipóteses a realizar posteriormente.

A maioria dos artigos revistos acerca do problema de planeamento de cirurgias assume que as durações de cirurgias seguem uma distribuição log-normal (Nawaz Ripon e Henrik Nyman, 2020). Strum et al. (2000) realizaram diversos estudos de modo a "legitimar as transformações log-normais como ferramenta usada para a exploração de durações de intervenções cirúrgicas". Segundo Strum et al. (2000), testes de normalidade, como o teste de *Shapiro-Wilks*, podem rejeitar de forma inapropriada a hipótese de normalidade uma vez que este teste foi desenhado para ser usado conjuntamente com métodos de análise gráfica, justificando então o porquê de não ter sido aplicado um teste de normalidade após a transformação logarítmica.

Assim, as principais conclusões retiradas dos estudos de Strum et al. (2000) foram as seguintes:

1. As durações de cirurgias adequam-se mais a distribuições log-normais do que a distribuições normais;

2. O teste de normalidade de *Shapiro-Wilk* necessita de ser confirmado usando os gráficos da distribuição normal antes da rejeição da normalidade de uma amostra de grandes dimensões;
3. É seguro afirmar que após a transformação logarítmica das durações é possível aplicar análises de variância e modelos de regressão.

Desta forma, foi realizada uma transformação logarítmica das durações de cirurgias e, verificou-se que, após a mesma, a variável dependente apresenta uma distribuição aproximadamente normal como ilustrado no histograma da Figura 8. É importante salientar que não existem observações no intervalo $[0,2[$ de $\log(Duração)$ uma vez que, como descrito na Subsecção 4.2, foram eliminados os registos com durações inferiores a dez minutos.

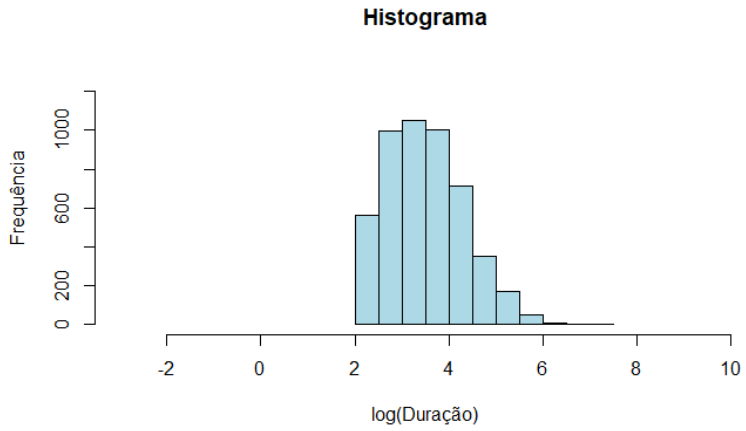


Figura 8 - Histograma da distribuição da variável dependente após transformação.

Para verificar a existência de multicolinearidade recorreu-se ao cálculo dos valores de VIF. Pela interpretação da fórmula do VIF, percebemos que, no mínimo, o seu valor será 1. Considera-se um valor elevado acima de 10, ou seja, quando o R_j^2 iguala 0,9, sendo evidência da existência de multicolinearidade.

Não é aconselhado analisar a multicolinearidade das variáveis categóricas através de valores VIF, pois segundo Allen (1997), este cálculo falha nos casos em que existem variáveis categóricas com mais do que uma categoria. Em Wissmann & Toutenburg (2007), de modo a garantir que não é elevada a existência de multicolinearidade, é importante a escolha de uma *dummy* de referência. Se a proporção de casos na categoria de referência for menor do que nas restantes, obtêm-se valores de VIF altos, mesmo que não exista multicolinearidade. Pelas razões apresentadas anteriormente, avaliam-se apenas os valores VIF calculados para as variáveis numéricas e binárias, que são apresentados na Tabela 4.

Tabela 4 - Resultados VIF das variáveis numéricas e binárias.

Variável Independente	VIF
INT_ANTERIOR	1,0637
T_ESPERA	1,5094
OBESIDADE_C	1,3066
NUTEIS_C	1,0508
C_CIRURGIA	1,7167
MEAN_MEDICO	1,8097
MEAN_PRIORIDADE	1,3129
MEAN_ICD	1,6242

O valor mais alto é dado pela variável independente MEAN_MEDICO com um VIF de aproximadamente 1,81, que é significativamente inferior a 10, pelo que podemos afirmar que as variáveis testadas não apresentam multicolinearidade e, consequentemente, serão mantidas.

5.3.1.2 Seleção de Variáveis

Verificados os pressupostos necessários, iniciamos o processo de seleção de variáveis recorrendo ao teste da ANOVA, no qual foi determinada a relevância de todas as variáveis categóricas testadas para a previsão das durações de cirurgias. Em todos os testes obteve-se um *p-value* de aproximadamente zero, o que originou a rejeição da hipótese nula e, consequentemente, validou-se que todas apresentam diferenças significativas entre as categorias e demonstram-se úteis para a previsão em questão (ver Tabela A 2.1 no Anexo 2).

De seguida, dado que nenhuma variável categórica foi excluída, há que ter em conta as variáveis que poderão estar correlacionadas entre si considerando a informação que contêm. Assim, estas variáveis são agrupadas em duas categorias tendo em conta a sua informação:

1. ICD_PROC, GDH, COD_EXTRACAO e COD_GNOSOLOGICO – contêm informação relativa ao procedimento efetuado no paciente.
2. ICD_DIAG e ESPECIALIDADE_C – relacionadas com determinada especialidade.

Ao aplicar o modelo, observando o R^2, conseguimos perceber que, excluir simultaneamente ICD_PROC e ICD_DIAG, não é sustentável dado que o modelo perde cerca de 4% do seu poder explicativo inicial com a totalidade das variáveis.

Desta forma, analisaram-se as restantes variáveis que poderão estar correlacionadas com os ICD_PROC e ICD_DIAG:

- Como o ICD_PROC e o GDH contêm a mesma informação, procedeu-se ao estudo do R^2 ao eliminar individualmente cada uma dessas variáveis. Observou-se que eliminar o ICD_PROC resulta numa redução do R^2 em 0,40% relativamente ao modelo inicial. Excluindo o GDH esta redução é apenas de 0,13%, por isso excluiu-se o GDH do modelo.

- Eliminar individualmente COD_EXTRACAO e COD_GNOSOLOGICO, não resulta em diferenças no R^2. Verificou-se então se seria possível excluir ambas as variáveis e é obtido um R^2 com menos 0,10% do poder explicativo inicial, ou seja, ambas as variáveis podem ser excluídas do modelo.
- A eliminação da variável categórica ESPECIALIDADE_C reduz em apenas 0,01% o poder explicativo do modelo. Assim, tendo em conta que a exclusão de ICDs resulta numa perda muito maior do poder explicativo em todos os casos, pode optar-se pela exclusão da variável ESPECIALIDADE_C em relação à ICD_DIAG.

Por fim, vamos proceder à *Backward Stepwise Regression*. Primeiramente, excluiu-se a variável T_ESPERA com um *p-value* de 0,7682. A próxima a ser excluída foi a variável numérica MEAN_PRIORIDADE, relativa à média de durações por prioridade das cirurgias, com *p-value* = 0,6226. A terceira variável a ser excluída foi NUTEIS_C (*p-value* = 0**,**2316), a variável binária referente ao dia da semana em que a intervenção ocorreu. Por último, excluiu-se a variável MEAN_ICDPROC, relativa à média de durações de intervenção por ICD de procedimento, com um *p-value* de 0,0998. Os resultados apresentados podem ser verificados também no Anexo 2 e Tabela A 2.2.

O modelo final apresenta todas as variáveis como significativas pois verifica-se o critério de paragem estabelecido – todas as variáveis têm um *p-value* inferior a 5%.

5.3.1.3 Análise de Resíduos

A análise de resíduos tem início na validação do primeiro pressuposto referente à distribuição normal dos mesmos. Assim, recorreu-se à análise gráfica do Q-Q *plot*, apresentado na Figura 9, que representa a distribuição dos resíduos contrastando-a com a esperada normal.

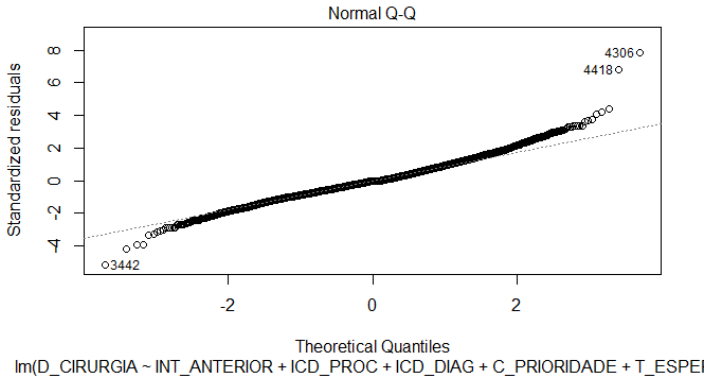


Figura 9 - Q-Q *Plot* dos resíduos.

Podemos verificar que os resíduos estandardizados encontram-se maioritariamente sob a reta a tracejado, à exceção de alguns pontos em ambas as extremidades, o que indica que parecem seguir uma

distribuição normal. Tratando-se de dados reais, os resultados obtidos são satisfatórios uma vez que podem ter registos incorretos.

Na Figura 10 observa-se um *p-value* de 0,1320 no teste de Durbin-Watson, logo, pode-se afirmar que para o modelo do Data Set RLM Agregado não se rejeita a hipótese nula e existe evidência estatística para afirmar que os resíduos no modelo de regressão não estão correlacionados.

```
lag Autocorrelation D-W Statistic p-value
  1      0.01847621      1.963045   0.132
Alternative hypothesis: rho != 0
```

Figura 10 - Teste de Durbin-Watson.

O próximo passo na análise de resíduos consiste em obter um gráfico de dispersão, no qual se consideram os resíduos no eixo de y e os valores ajustados no eixo de x, representado na Figura 11.

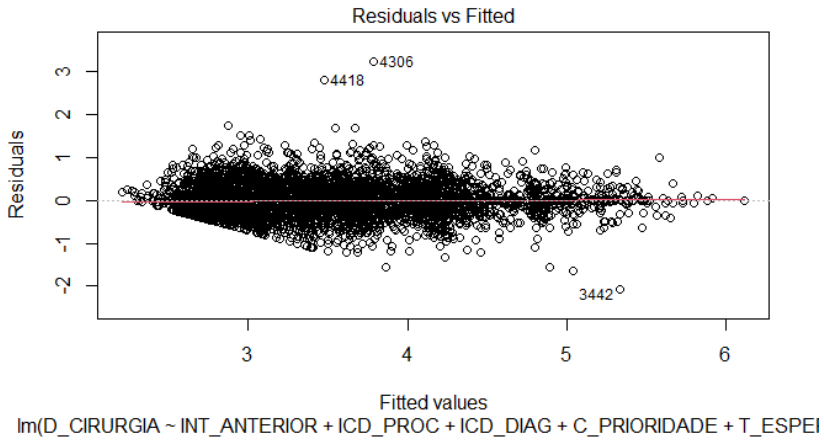


Figura 11 - Gráfico de dispersão dos resíduos.

Através do mesmo, conseguimos analisar a existência de homocedasticidade e se o valor médio dos resíduos é zero. Assim, podemos observar que não existe uma tendência na dispersão dos resíduos ao longo do eixo de x, ou seja, a variância do erro não depende dos valores ajustados. Através do gráfico de dispersão é ainda possível verificar que os resíduos se encontram aleatoriamente dispersos em torno de zero, o que indica que o seu valor médio é de facto zero. Devemos ainda ter novamente em consideração possíveis erros de registo, sendo que tal pode explicar os valores mais afastados da reta (observações 4418, 4306 e 3442).

Validados todos os pressupostos do modelo de RLM podem-se analisar os resultados do modelo e tecer conclusões acerca da precisão da previsão.

5.3.2 *Data Set* RLM por Especialidade

No caso do Data Set RLM por Especialidade, não são analisados os resultados para validação de pressupostos dos modelos. No entanto, como já foi referido, o mesmo processo efetuado no Data Set RLM Agregado foi replicado e foram também validados todos os pressupostos para cada uma das especialidades.

Nos Anexos 1, Figura A 1.1 e Tabela A 1.2, podem observar-se os histogramas da variável dependente e os valores de VIF, respetivamente. Durante este processo foram eliminadas as variáveis OBESIDADE_C, MEAN_MEDICO e MEAN_ICDPROC na especialidade de Gastrenterologia e as variáveis INT_ANTERIOR e MEAN_ICDPROC na especialidade de Obstetrícia, pois apresentaram valores de VIF muito superiores a dez.

Este conjunto de dados, para além dos passos anteriormente mencionados, passou ainda pela identificação e remoção de variáveis independentes categóricas e binárias cujo valor seria igual em todos os registos e, por isso, não seriam úteis para a previsão. A variável OBESIDADE_C apresenta-se com o valor "0" para todos os registos em todas as especialidades, à exceção de Gastrenterologia. O mesmo acontece com a variável ICD_DIAG que apresenta a mesma categoria em todos os registos da especialidade de Gastrenterologia, tal como a variável COD_GNOSOLOGICO na especialidade de Obstetrícia. Assim, estas variáveis categóricas, não serão utilizadas nos modelos de previsão para as especialidades indicadas.

Na função *coef* do RStudio, é devolvido o coeficiente da variável no modelo de RLM, incluindo as variáveis cujos coeficientes assumem um valor de "NA". Nestes casos, as variáveis são excluídas do modelo (R Documentation). Verificamos que com todos os modelos no *Data Set* RLM Por Especialidade para a variável MEAN_ICDPROC os seus coeficientes são "NA", logo, a variável foi excluída e não voltará a ser mencionada na análise seguinte.

Para verificar se as variáveis categóricas são significativas aplicou-se o teste da ANOVA, como foi feito para o Data Set RLM Agregado. Assim, foram excluídas as variáveis que apresentavam *p-value* superior a 5% e que se encontram representadas na Tabela 5, sendo que os resultados discriminados encontram-se no Anexo 2, Tabela A 2.1. Na Tabela 5 observam-se as especialidades identificadas nas linhas e cada uma das variáveis dependentes a ser testada (nas colunas), sendo que as variáveis eliminadas são identificadas pelo símbolo "x".

Tabela 5 - Variáveis excluídas pelo teste da ANOVA.

Especialidade	INT_ANTERIOR	ICD_PROC	ICD_DIAG	C_PRIORIDADE	MES_C	SEMANA_C	OBESIDADE_C	C_CIRURGIA	GDH	COD_EXTRACAO	COD_GNOSOLOGICO	NUTEIS_C
Cirurgia Geral	x											
Cirurgia Plástica	x			x	x			x				
Gastrenterologia	x				x							x
Ginecologia				x	x					x	x	x
Obstetrícia			x	x	x	x						x
Oftalmologia											x	
ORL	x				x	x				x	x	x
Ortopedia												
Pediatria Cirúrgica	x							x				x
Urologia												x

Relativamente à análise das variáveis categóricas que poderiam estar correlacionadas, a mesma foi feita de forma semelhante ao Data Set RLM Agregado, e os resultados mostraram-se iguais para todas as especialidades exceto Urologia. À semelhança do Data Set RLM Agregado:

- A exclusão do ICD_PROC e ICD_DIAG originou uma perda demasiado grande do poder explicativo do modelo e, sendo assim, optou-se pela exclusão do GDH.
- A exclusão dos COD_EXTRACAO e COD_GNOSOLOGICO não resultou nenhuma diferença no poder explicativo do modelo e, portanto, foram ambas eliminadas.
- No caso da especialidade de Urologia, através do R^2 foi possível concluir que neste caso ao analisar os dois conjuntos de variáveis que podem ter elevadas correlações, ao contrário do que foi observado anteriormente, seria preferível excluir ICD_DIAG em oposição ao GDH.

Para cada uma das especialidades, as variáveis excluídas estão divididas em três grupos e são apresentadas na Tabela 6:

1. Variáveis Correlacionadas – pares de variáveis independentes categóricas que dão informação semelhante;
2. Variáveis com Coeficientes "NA" – variáveis independentes com coeficiente "NA";
3. Variáveis com Uma Categoria – variáveis independentes categóricas que apresentam a mesma categoria para todas as observações em determinada especialidade.

Tabela 6 - Variáveis excluídas: Variáveis Correlacionadas; Variáveis c/ Coeficientes "NA"; Variáveis c/ Uma Categoria.

	Variáveis Correlacionadas	Variáveis c/ Coeficientes "NA"	Variáveis c/ Uma Categoria
Cirurgia Geral	GDH COD_EXTRACAO COD_GNOSOLOGICO	MEAN_ICDPROC MEAN_PRIORIDADE NUTEIS_C	OBESIDADE_C
Cirurgia Plástica	GDH COD_EXTRACAO COD_GNOSOLOGICO	MEAN_ICDPROC OBESIDADE_C NUTEIS_C	OBESIDADE_C
Gastrenterologia	GDH COD_EXTRACAO COD_GNOSOLOGICO	MEAN_PRIORIDADE	ICD_DIAG
Ginecologia	GDH COD_EXTRACAO COD_GNOSOLOGICO	MEAN_ICDPROC	OBESIDADE_C
Obstetrícia	GDH COD_EXTRACAO COD_GNOSOLOGICO	MEAN_ICDPROC C_CIRURGIA	OBESIDADE_C COD_GNOSOLOGICO
Oftalmologia	GDH COD_EXTRACAO COD_GNOSOLOGICO	MEAN_ICDPROC MEAN_PRIORIDADE NUTEIS_C	OBESIDADE_C
ORL	GDH COD_EXTRACAO COD_GNOSOLOGICO	MEAN_ICDPROC MEAN_PRIORIDADE	OBESIDADE_C
Ortopedia	GDH COD_EXTRACAO COD_GNOSOLOGICO	MEAN_ICDPROC MEAN_PRIORIDADE	OBESIDADE_C
Pediatria Cirúrgica	GDH COD_EXTRACAO COD_GNOSOLOGICO	MEAN_ICDPROC MEAN_PRIORIDADE	OBESIDADE_C
Urologia	ICD_DIAG COD_EXTRACAO COD_GNOSOLOGICO	MEAN_ICDPROC MEAN_PRIORIDADE	OBESIDADE_C

Na etapa de *Backward Stepwise Regression* foi respeitado o critério de paragem no qual se exige um *p-value* máximo de 5% para cada uma das variáveis independentes do modelo. Assim, na Tabela 7 encontram-se representadas as variáveis independentes excluídas e os respetivos *p-value* de exclusão (ver Anexo 2, Tabela A 2.3 até Tabela A 2.7).

Tabela 7 - Variáveis excluídas durante a *Backward Stepwise Regression*.

	Variável Independente Excluída	*p-value*
Cirurgia Geral	T_ESPERA	0,5964
Cirurgia Plástica	MEAN_PRIORIDADE	0,7370
	MEAN_MEDICO	0,3652
	T_ESPERA	0,3195
Gastrenterologia	C_CIRURGIA	0,7763
Ginecologia	INT_ANTERIOR	0,3633
	C_CIRURGIA	0,3397
	MEAN_PRIORIDADE	0,1137
	T_ESPERA	0,0538
Obstetrícia	T_ESPERA	0,1815
Oftalmologia	T_ESPERA	0,2560
	C_CIRURGIA	0,1856
	INT_ANTERIOR	0,0560
ORL	MEAN_MEDICO	0,7769
	T_ESPERA	0,6830
Ortopedia	T_ESPERA	0,0729
	NUTEIS_C	0,0714
	INT_ANTERIOR	0,0541
Pediatria Cirúrgica	MEAN_MEDICO	0,4302
	T_ESPERA	0,2319
Urologia	MEAN_MEDICO	0,7262
	T_ESPERA	0,6395
	C_CIRURGIA	0,5960
	INT_ANTERIOR	0,2403

Por último, a análise de resíduos para cada uma das especialidades pode ser consultada no Anexo 3 e na Figura A 3.1, Tabela A 3.2 e Figura A 3.3.

5.4 Modelos Finais

Na Tabela 8 são apresentados os resultados obtidos, nos modelos de RLM, para os indicadores de desempenho RMSE e R^2.

Tabela 8 - Resultados dos indicadores de desempenho dos modelos finais de RLM.

Modelo RLM	**RMSE**	R^2
Data Set RLM Agregado	0,3772	0,7750
Data Set RLM por Especialidade		
Cirurgia Geral	0,3919	0,7769
Cirurgia Plástica	0,2405	0,8559
Gastrenterologia	0,3309	0,8882
Ginecologia	0,3553	0,7829
Obstetrícia	0,2478	0,2873
Oftalmologia	0,3245	0,5912
ORL	0,3687	0,7012
Ortopedia	0,3537	0,7854
Pediatria Cirúrgica	0,2125	0,8615
Urologia	0,3999	0,7151

Todos os resultados apresentados na Tabela 8 foram obtidos através do RStudio (versão 1.4.1106). A leitura e análise dos dados foi feita com auxílio do pacote *readxl* e *graphics* (utilizado para obter a distribuição das durações das cirurgias). Para a fase de validação de pressupostos foram utilizados os pacotes *base* e *car*, obtendo com os mesmos os *outputs* que permitiram analisar os resultados do VIF e o comportamento dos resíduos. No que diz respeito à seleção de variáveis foi utilizado o pacote *stats* que se demonstra duplamente útil uma vez que: permite a utilização do comando *lm,* responsável pela criação do modelo de RLM; e do comando *aov* que é utilizado para aplicar o teste da ANOVA. Por fim, o indicador de desempenho RMSE foi obtido através do pacote *Metrics*.

5.4.1 Data Set RLM Agregado

Após a execução da metodologia descrita, resultou um conjunto de nove variáveis das quais oito são variáveis categóricas e apenas uma é variável independente numérica. Na Tabela 9 encontram-se as variáveis que compõem o modelo tal como o respetivo valor estimado do parâmetro β. Estes valores não são apresentados para as variáveis categóricas que contêm um elevado número de categorias.

Tabela 9 - Lista das variáveis do modelo de RLM para o Data Set RLM Agregado e respetivos β.

	β
β_0	2,0832
INT_ANTERIOR	0,0611
ICD_PROC	–
ICD_DIAG	–
C_PRIORIDADE	–
SEMANA_C	–
MES_C	–
C_CIRURGIA	0,2374
OBESIDADE_C	4,1262
MEAN_MEDICO	0,0055

O modelo final apresenta um R^2 de 0,7750 (Tabela 8), ou seja, as variáveis selecionadas para o modelo explicam 77,50% da variável dependente. O seu RMSE é de 0,3772.

5.4.2 Data Set RLM por Especialidade

No conjunto Data Set RLM por Especialidade, observa-se, através da Tabela 8, que oito de dez especialidades apresentam um R^2 superior a 70%. Destacam-se as especialidades de Gastrenterologia e Pediatria Cirúrgica com valores de 0,8882 e 0,8615, respetivamente.

Na especialidade de Gastrenterologia, as variáveis que constituem o modelo são: ICD_PROC; C_PRIORIDADE; T_ESPERA e SEMANA_C. No caso de Pediatria Cirúrgica, observamos um conjunto de variáveis semelhante: ICD_PROC; ICD_DIAG; C_PRIORIDADE; SEMANA_C e MES_C.

Os modelos com piores resultados de R^2 correspondem às especialidades de Obstetrícia e Oftalmologia, com valores de 0,2873 e 0,5912, respetivamente. Apesar disto, no caso de Obstetrícia percebemos que o seu valor de RMSE não é tão elevado como em alguns modelos com melhor R^2.

O modelo da especialidade de Obstetrícia conta com o seguinte conjunto de variáveis independentes: ICD_PROC; MEAN_MEDICO e MEAN_PRIORIDADE. Por sua vez, o modelo correspondente à especialidade de Oftalmologia conta com seis variáveis independentes, das quais cinco são variáveis categóricas: ICD_PROC; ICD_DIAG; C_PRIORIDADE; SEMANA_C; MES_C e MEAN_MEDICO.

CAPÍTULO 6 – XGBOOST

6.1 Introdução ao XGBoost

Atualmente, a ML é um tópico muito estudado dado que pode ser utilizado nas mais diversas áreas. Este não só se encontra aplicado a probabilidades e estatística, como também a áreas de engenharia, informação e aspetos das ciências sociais (el Naqa & Murphy, 2015).

Segundo Shinde & Shah (2018), o termo ML surgiu em meados de 1950, introduzindo ao mundo uma área na qual "as máquinas iriam tentar ser mais inteligentes do que os humanos". O primeiro algoritmo de ML surgiu em 1952, sendo que o mesmo foi desenvolvido com o objetivo de executar um jogo de xadrez.

Eventualmente estes estudos progrediram de maneira a preencher diversas necessidades. Assim, em 1995 surgiu às mãos de Freund and Schapire um algoritmo chamado *Adaboost*.

Tendo por base o algoritmo de Freund and Schapire, Breiman, em 2001, desenvolveu uma nova versão do algoritmo, chamado de *Gradient Boosting* (GB), que contém diversas árvores de decisão criadas por um subconjunto aleatório dos dados (Shinde & Shah, 2018). Assim, "em cada iteração uma subamostra é retirada de forma aleatória (sem reposição) do conjunto de dados de treino" (Friedman, 2002). No método de GB existe uma aprendizagem iterativa, ou seja, o GB acaba por melhorar a sua precisão ao desenvolver diversos modelos em sequência e colocando mais ênfase nos casos de treino mais difíceis de estimar. Neste método, não existe a mesma probabilidade de seleção para as subamostras em todo o conjunto de dados, uma vez que, dada a informação referida anteriormente, é mais provável serem selecionados casos incorretamente estimados (Zhang & Haghani, 2015). Assim, cada modelo tem como objetivo a correção dos erros existentes em modelos anteriores.

É importante referir que, apesar de com os modelos de GB ser possível obter previsões de elevada precisão, a interpretação do modelo deixa de ser intuitiva como no caso da RLM (James et al., 2013), sendo necessário ter em conta o objetivo da previsão. Esta pode ter como finalidade um de dois objetivos: precisão ou inferência. No primeiro, supondo dado um conjunto X e um conjunto Y que não é facilmente obtido, pode-se dizer que Y é obtido através de $\hat{Y} = \hat{f}(X)$ onde "$\hat{f}$ é frequentemente tratado como uma caixa preta, no sentido em que normalmente não existe a preocupação pela forma exata de $\hat{f}$, desde que produza previsões precisas de $\hat{Y}$" (James et al., 2013). Por outro lado, se o objetivo for a inferência existe a preocupação de entender em que medida determinadas variáveis independentes vão influenciar a variável dependente. Tendo em conta que o principal objetivo deste TFM é obter modelos precisos, sendo a análise de como determinadas variáveis independentes poderão influenciar a duração da cirurgia um objetivo secundário, o modelo de ML escolhido para a previsão de durações de cirurgias foi o XGBoost. Este é

mencionado na literatura como sendo o modelo que habitualmente apresenta resultados mais favoráveis na previsão.

O XGBoost insere-se na categoria de aprendizagem supervisionada (*supervised machine learning*) e no grupo de algoritmos chamados *ensemble*. Segundo Kotsiantis (2007) a aprendizagem supervisionada remete para a procura de algoritmos que têm como objetivo produzir hipóteses a partir de instâncias externas, onde posteriormente serão feitas previsões para instâncias ainda não observadas. Por sua vez, segundo James et al. (2013) algoritmos de *ensemble* combinam vários modelos mais simples, conhecidos como *weak learners*, de forma a obter um modelo que irá ser potencialmente mais preciso na previsão. Sabe-se ainda que o método original seria o *Bayesian Averaging*, no entanto, mais recentemente acabaram por ser desenvolvidos métodos como o *bagging* e o *boosting*, onde se inclui o XGBoost (Dietterich, 2000).

Desta forma, o XGBoost é um modelo de otimização que vem combinar árvores de decisão através de um modelo de *boosting* (Yu et al., 2019). Este algoritmo é baseado em árvores de decisão, quando se trata de problemas de classificação, e em árvores de regressão para problemas de regressão (Dong et al., 2020).

6.2 Problema Conhecido - *Overfitting*

O XGBoost tem um amplo conjunto de parâmetros que ajudam a melhorar a aprendizagem e a evitar o *overfitting* (Shi et al., 2019). Este problema traduz-se no uso de modelos que incluem termos desnecessários ou são mais complexos do que o que seria necessário para obter um modelo "ótimo" consoante um conjunto de dados (Hawkins, 2004). O modelo começa a assimilar o ruído e flutuações aleatórias, acabando por considerar ambos úteis para a construção do mesmo (Parsa et al., 2020). Em termos práticos, isto verifica-se quando a amostra de teste obtém resultados no indicador escolhido para avaliação do modelo muito piores que os obtidos pela amostra de treino.

Uma das formas utilizadas para combater este problema, para além da otimização dos diversos parâmetros fornecidos pelo XGBoost, é a utilização da Validação Cruzada.

6.2.1 Otimização de Parâmetros

O XGBoost contém vários parâmetros que têm finalidades distintas e que podem ser divididos em três categorias, que se encontram representadas na Tabela 10, e são:

1. Parâmetros gerais, que são responsáveis por determinar as funcionalidades gerais do XGBoost;
2. Parâmetros do *booster,* que permitem guiar o *booster* para o modelo de XGBoost escolhido nos parâmetros anteriores;

3. Parâmetros de aprendizagem, que têm como finalidade controlar a otimização do modelo.

Tabela 10 - Parâmetros do XGBoost.

Parâmetros	Nome	*Default*	*Range*	Definição
Gerais	booster	gbtree	"gbtree; gblinear"	Tipo de modelo de XGBoost.
	silent	0		Define se as mensagens da corrida são visíveis.
Booster	eta	0,3	[0;1]	Diminui o peso de novas variáveis tornando o modelo menos suscetível a inseri-las. Usado para prevenir o *overfitting*.
	gamma	0	[0;∞[	Quanto maior o valor de gamma maior o valor da redução mínima de perda que se deve obter para haver um novo ramo na árvore.
	max_depth	6	]0;∞[	Profundidade máxima da árvore. Quanto maior o valor, maior a probabilidade de apresentar *overfitting*.
	min_child_weigth	1	[0;∞[	Valores maiores são usados para prevenir o *overfitting*, tornando o modelo menos sensível à modificação dos parâmetros.
	subsample	1	]0,1]	Percentagem da amostra que vai ser usada como amostra de treino em cada iteração do algoritmo.
	colsample_bytree	1	]0,1]	Proporção de subamostra ao construir cada árvore.
Aprendizagem	objective	reg:squarederror		Tem o propósito de definir a função de perda a ser minimizada.
	eval_metric	RMSE		Indicador utilizado nos dados de validação.

6.2.2 Validação Cruzada

Uma parte essencial na utilização de qualquer modelo de ML é a validação e o treino desse mesmo modelo, pois estes são algoritmos tão automatizados e focados na precisão que podem assimilar ruídos e flutuações aleatórias que poderão enviesar a precisão do resultado das previsões. Para colmatar isto poderia dividir-se a amostra em duas subamostras, uma de treino e outra de teste, onde uma tem $\gamma\%$ das observações e a outra $(100-\gamma)\%$. No entanto, no caso em questão isto iria traduzir-se numa perda significativa de informação tendo em conta o número reduzido de observações. Para além disso, os modelos baseados em árvores de decisão têm elevada variância, o que significa que, se se dividir a amostra de treino em dois conjuntos diferentes selecionados aleatoriamente e se aplicar árvores de decisão a ambos, os resultados obtidos em cada um dos modelos poderão ser muito diferentes (James et al., 2013). Assim, recorre-se a outras técnicas de ML, como a Validação Cruzada, que permitem validar o modelo sem comprometer a informação disponibilizada ao submeter uma amostra limitada a um processo de reamostragem.

A Validação Cruzada é então considerada uma técnica que divide a amostra em dois conjuntos: o primeiro é usado para treinar o modelo enquanto o segundo é usado para a validação do modelo. Na forma mais básica desta técnica – Validação Cruzada k-*folds* – "a amostra é repartida em k segmentos de igual tamanho (ou aproximadamente igual)". Seguidamente, são realizadas k iterações de treino e de validação de forma a que, em cada iteração, um subconjunto de dados diferente é tido em conta para a validação do modelo enquanto os restantes $k-1$ subconjuntos são usados para o treino do modelo (Refaeilzadeh et al., 2016). Desta forma, cada subconjunto vai ter a oportunidade de pertencer ao grupo de validação do modelo uma vez e pertencer ao grupo de treino $k-1$ vezes, sendo que este processo é repetido k vezes nas quais se obtêm k valores para o indicador de desempenho do modelo. Posto isto, este indicador vai ser a média dos k valores obtidos.

6.3 Metodologia

O primeiro passo para a obtenção dos modelos de XGBoost foi fazer uma transformação logarítmica das durações de cirurgias que, apesar de não ser necessária neste modelo, facilita o processo de comparação entre os resultados obtidos com os da RLM. É ainda importante mencionar que, ao contrário da RLM, não foi feita nenhuma seleção de variáveis.

De seguida, utilizou-se o *One Hot Encoding* através do pacote *fastDummies*, transformando as diversas categorias das variáveis categóricas não binárias em diversas variáveis *dummy*. Foi escolhido este método dado que não parece existir uma relação ordinal entre as diversas categorias. Assim, para cada uma delas é criada uma variável binária independente. Por exemplo, usando a variável categórica C_PRIORIDADE, que contém quatro categorias, na Figura 12 a codificação obtida através do método de *One Hot Encoding é* a que se apresenta.

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

Normal
Prioritário
Muito Prioritário
Urgência Diferida

Figura 12 - Matriz *One Hot Encoding, variável C_PRIORIDADE .*

O conjunto total de dados foi dividido posteriormente numa amostra de treino e numa amostra de teste que é composta por 20% da totalidade dos dados. Foi necessário manipular os dados de forma a separar a variável dependente y, correspondente à duração de cirurgias, das restantes variáveis independentes x com auxílio do pacote *dplyr*. De seguida, estes dois conjuntos de dados – { y amostra de treino, x amostra de treino } e { y amostra de teste, x amostra de teste } – foram ainda sujeitos a uma nova

transformação utilizando o comando *xgb.DMatrix*, uma vez que o pacote *xgboost*, mais especificamente a função *xgb.train*, utilizado pelo RStudio, necessita que os dados estejam no formato *xgb.DMatrix*. Este é um formato de estruturação de dados que foi desenvolvido especificamente para modelos gerados através do XGBoost, sendo responsável pela otimização da gestão da memória e da velocidade de treino do modelo.

O primeiro procedimento é executado com auxílio do comando *xgb.cv*, que incorpora a Validação Cruzada. O segundo, através do pacote *caret*, que executa uma série de comandos que permite fazer diversas combinações dos parâmetros com valores pré-determinados para os mesmos. Dentro do comando *train*, usa-se o *expand.grid*, correspondente à *Grid Search*.

A *Grid Search*, é considerada uma das principais ferramentas de otimização de parâmetros usadas em modelos de ML. No caso do XGBoost, é determinado o conjunto (*grid*) de valores a testar para cada um dos parâmetros, procurando posteriormente a combinação que minimiza o erro do modelo de previsão para um determinado conjunto de dados. O estudo de Yu et al. (2019) corrobora o uso desta técnica afirmando que "o uso da *Grid Search* para a otimização de parâmetros em alguns problemas de otimização mostrou-se como a melhor escolha para o algoritmo de XGBoost".

Com a melhor combinação de parâmetros obtida, constrói-se o modelo final com o comando *xgb.train*, através do qual podemos obter diversos *outputs*, como por exemplo:

- Gráficos que evidenciam a existência de *overfitting* ou *underfitting*;
- Identificação das variáveis mais significativas para o modelo final e valores *SHapley Additive exPlanations* (SHAP).

Através do primeiro *output*, caso sejam identificados indícios de *overfitting*, vão ser alterados os valores dos parâmetros obtidos através do *Grid Search* para o modelo final. Os parâmetros reguladores do modelo que mostraram ter mais influência na prevenção do problema em questão foram o *eta*, o *gamma* e o *colsample_bytree*.

Em suma, o modelo final foi obtido realizando os seguintes passos:

1. Determinar, com o comando *xgb.cv*, quais os números de iterações que originam os três melhores (mais baixos) valores de RMSE.
2. Para cada número de iterações obtido no passo anterior, testar várias combinações de parâmetros através da *Grid Search* e determinar as que minimizam a função de perda MSE.
3. Testar o modelo final com a combinação de parâmetros que originou melhores valores de RMSE no Passo 2.
4. Caso o modelo final com os parâmetros selecionados apresente evidências de *overfitting*, alterar os valores dos parâmetros usados para tentar impedir tal problema voltando ao Passo 2.

6.4 Modelos Finais

Como referido anteriormente, no caso do modelo de RLM, todos os resultados para o modelo de XGBoost foram obtidos através do RStudio (versão 1.4.1106). Para que estes fossem reproduzíveis utilizou-se uma *seed*.

Depois de seguida a metodologia anteriormente apresentada, obtemos os resultados presentes na Tabela 11 para o modelo de XGBoost nos dois conjuntos de dados – Data Set XGBoost Agregado e Data Set XGBoost por Especialidade.

Tabela 11 - Resultados XGBoost (Data Set XGBoost Agregado e Data Set XGBoost por Especialidade).

Modelo XGBoost	**RMSE**	R^2
Data Set XGBoost Agregado	0,3939	0,7547
Data Set XGBoost por Especialidade		
Cirurgia Geral	0,4130	0,7523
Cirurgia Plástica	0,2252	0,8736
Gastrenterologia	0,2569	0,9326
Ginecologia	0,2940	0,8514
Obstetrícia	0,2936	-0,0001
Oftalmologia	0,3198	0,6031
ORL	0,4959	0,4597
Ortopedia	0,3989	0,7269
Pediatria Cirúrgica	0,1782	0,8299
Urologia	0,3807	0,7417

6.2.1 Data Set XGBoost Agregado

O Data Set XGBoost Agregado obteve um RMSE de aproximadamente 0,3939 após parametrização do modelo. Em termos de R^2, o modelo tem um valor de 0,7547.

Na Figura 13, observam-se os valores de RMSE por iteração. No eixo das ordenadas é registado o valor de RMSE e no eixo das abcissas a respetiva iteração, sendo que a linha a vermelho corresponde ao conjunto de dados de teste enquanto os pontos azuis correspondem aos dados de treino.

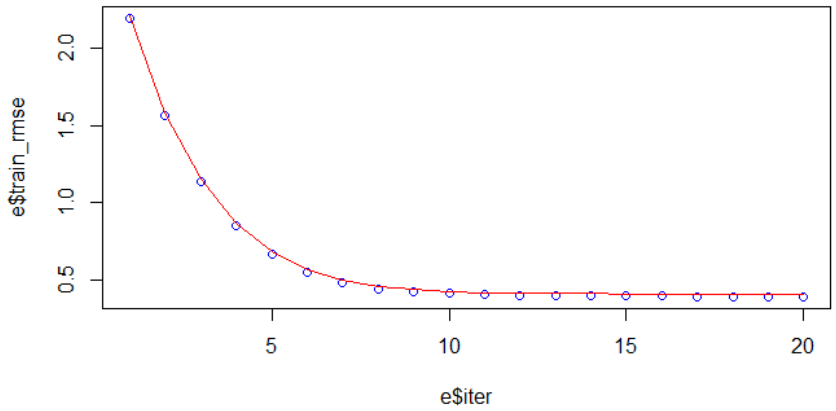


Figura 13 - Resultados RMSE por iteração (dados de treino e dados de teste).

Verifica-se que não existe diferença significativa entre os resultados, em termos de RMSE, de ambos os conjuntos de dados. Logo o modelo de XGBoost não apresenta evidências de *overfitting*.

Uma das funcionalidades do pacote *xgboost* RStudio permite identificar as variáveis que mostram ser significativas para o modelo. A Figura 14 regista as cinco variáveis mais importantes para a previsão, sendo que no eixo das abcissas consta a "contribuição percentual de cada variável para o modelo" (R Documentation). Quanto maior esta percentagem maior será a importância da variável para a previsão.

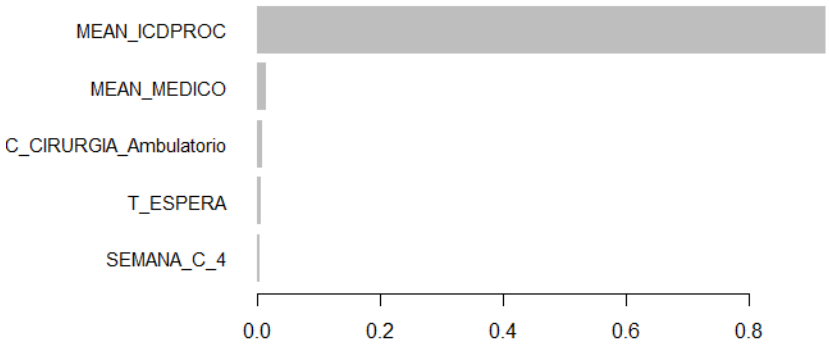


Figura 14 – Matriz de importância das variáveis, Data Set XGBoost Agregado.

Desta forma, no presente modelo existe um grande destaque para a variável MEAN_ICDPROC correspondente à média da duração por ICD de Procedimento, ultrapassando os 80%. As restantes variáveis já são de menor importância, uma vez que apresentam uma contribuição percentual inferior a 2%.

Uma das formas para colmatar a falta de interpretabilidade que existe em modelos como o XGBoost é o recurso aos valores SHAP. Neste caso, o valor SHAP traduz o impacto da mudança num valor de uma variável independente na duração da cirurgia, dado que quanto maior o SHAP maior a duração da mesma.

Na Figura 15, apresenta-se o valor SHAP no eixo das ordenadas e o valor original da variável no eixo das abcissas, tendo em conta que cada ponto azul é uma cirurgia registada ou observação.

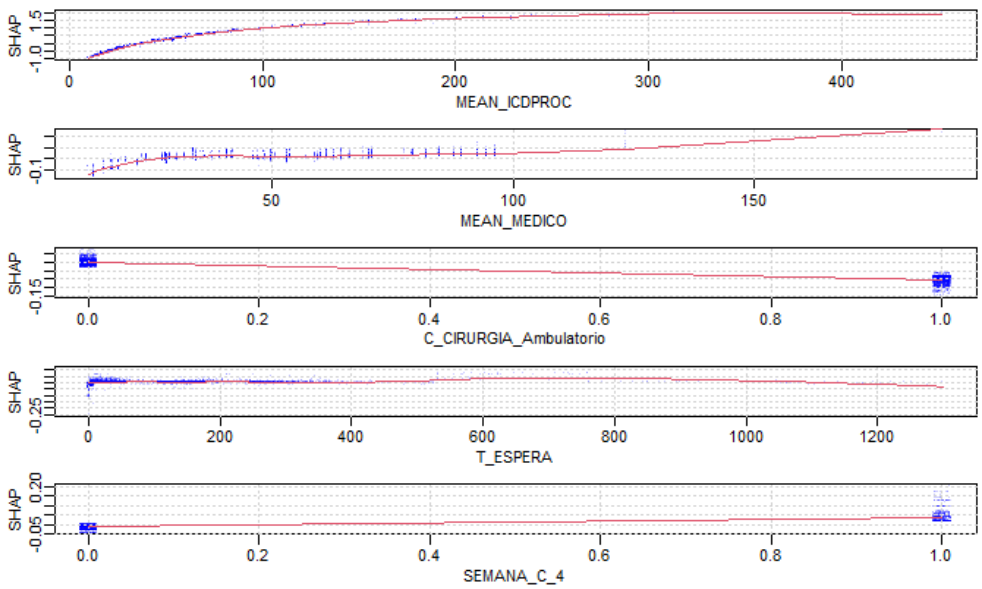


Figura 15 - Valores SHAP das cinco variáveis mais importantes.

No caso da variável MEAN_ICDPROC facilmente se observa que quanto menor é o valor da duração média por ICD de procedimento, menor vai ser a duração da cirurgia. Relativamente à variável MEAN_MEDICO podemos estar perante uma variável que se torna importante para a previsão mediante interação com outras variáveis uma vez que um determinado valor de duração de cirurgia por médico pode dar origem a um grande intervalo de valores SHAP (formação de linhas azuis verticais em diversos valores da variável). No caso seguinte, como seria de esperar, observa-se que quando a cirurgia é do tipo Ambulatório (não requer internamento), ou seja, quando a variável binária é igual a um, existe uma tendência para que a duração da cirurgia seja menor uma vez que os valores SHAP são menores comparativamente aos valores obtidos quando a variável binária assume o valor de zero. Para a variável T_ESPERA verificar-se que, de facto, validando a informação da Figura 14, o tempo de espera não exerce grande influência sobre a duração de cirurgia uma vez que os valores SHAP mantém-se praticamente no mesmo intervalo, apesar do aumento no valor da variável. O mesmo se pode dizer acerca da variável SEMANA_C_4, variável *dummy* associada ao quarto dia da semana (Quinta), na qual não existe uma relação clara entre esta categoria e a duração de cirurgia.

6.2.2 Data Set XGBoost por Especialidade

No Data Set XGBoost por Especialidade, a média de RMSE para as 10 especialidades foi de aproximadamente 0,3256 (Tabela 11).

O conjunto de dados que melhor se adaptou ao modelo e, consequentemente, obteve o melhor R^2 de 0,9326, foi a especialidade de Gastrenterologia.

Contrariamente, o pior modelo conta com um R^2 de -0,0001 e corresponde à especialidade de Obstetrícia. Tendo em conta a fórmula utilizada no cálculo do R^2, $1 - \frac{SS_{res}}{SS_{tot}} = 1 - \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$, sabe-se que, no

caso da RLM, os valores estarão sempre no intervalo $[0, 1]$, uma vez que a mesma vai sempre dar origem a modelos que minimizam os resíduos, ou seja, a soma total dos quadrados (SS_{tot}) vai ser sempre superior à soma dos quadrados dos resíduos (SS_{res}). Este valor negativo de R^2 é consequência de não serem verificadas as condições de normalidade, que no caso do XGBoost não são exigidas, logo é possível que o modelo de previsão seja pior do que o correspondente a uma linha reta representativa da média das observações. Assim, nestes modelos a média passa a ser melhor do que a própria previsão, consequentemente $SS_{res} > SS_{tot}$, e obtem-se um valor fora do intervalo $[0, 1]$.

No caso da especialidade de Gastrenterologia vê-se, através da Figura 16, que o modelo obtido com a combinação de parâmetros que originou o menor RMSE não apresenta *overfitting*.

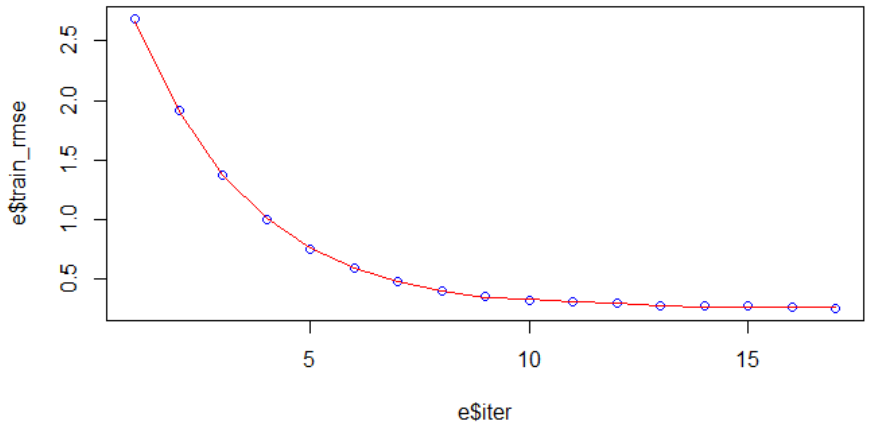


Figura 16 - Resultados RMSE por iteração (dados de treino e dados de teste), para a especialidade de Gastrenterologia.

Na Figura 17, constam as cinco variáveis mais importantes para o modelo de Gastrenterologia.

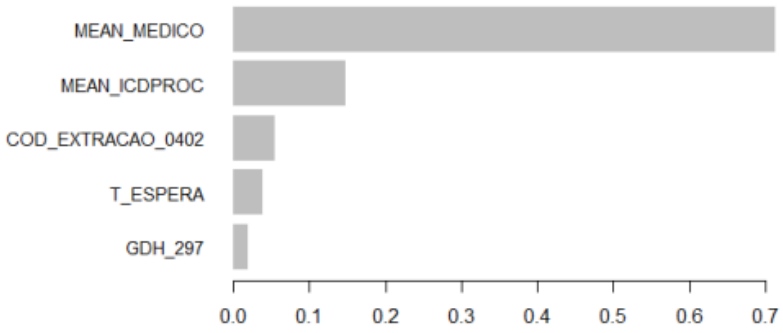


Figura 17 - Matriz de importância das variáveis, especialidade de Gastrenterologia.

A variável MEAN_MEDICO, representativa da média da duração de cirurgia por médico, tem um papel fulcral na previsão de durações de cirurgias relativas a Gastrenterologia. É ainda de destacar a variável MEAN_ICDPROC cuja importância é de 15%, enquanto as restantes se encontram abaixo deste valor, sendo menos importantes para a previsão. É importante referir que apesar de se obter um modelo melhorado, o aumento dos parâmetros que influenciam a construção do modelo origina um modelo mais conservador pois, vai ignorar relações menos significativas, acabando por incluir um menor número de variáveis para a

previsão. Um exemplo deste comportamento é a especialidade de ORL, na qual após as modificações nos valores testados pela *Grid Search,* a mesma conta apenas com três variáveis importantes para o modelo (ver Figura A 4.3 no Anexo 4).

Os valores SHAP para os modelos por especialidade são apresentados no Anexo 4 pois o número de observações é significativamente reduzido fazendo com que os pontos azuis não sejam percetíveis, logo, não é fazível a análise dos mesmos.

No caso da especialidade de Obstetrícia observa-se na Figura 18, entre a segunda e a décima iteração, a existência de *underfitting* (situação oposta ao *overfitting*, isto é, o modelo é incapaz de detetar uma relação entre as variáveis).

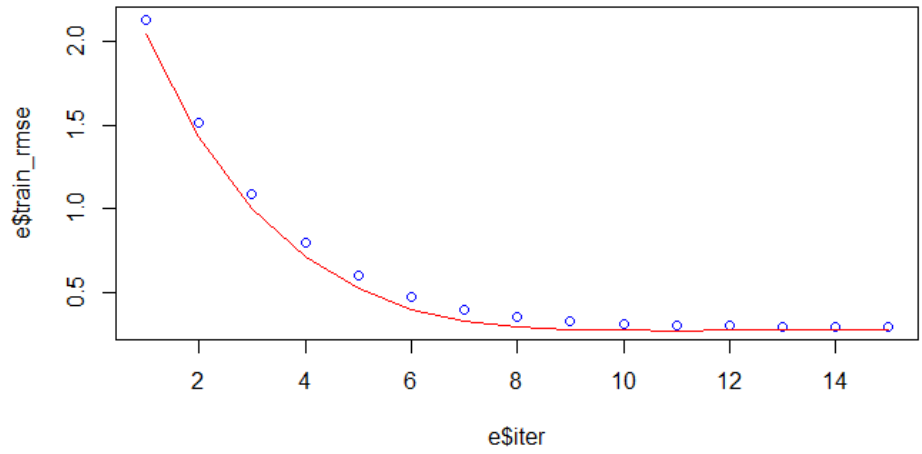


Figura 18 - Resultados RMSE por iteração (dados de treino e dados de teste), para a especialidade de Obstetrícia.

Foram alterados os intervalos de valores dos parâmetros de forma inversa à aplicada para anular o *overfitting*, acrescentando inclusive um maior número de ramificações (*max_depth*), e obteve-se o modelo que deu origem aos resultados presentes na Figura 19.

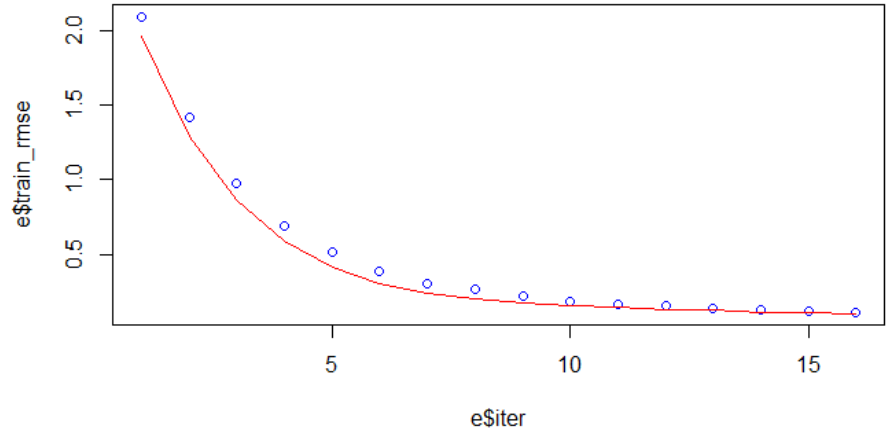


Figura 19 - Resultados RMSE por iteração (dados de treino e dados de teste), para a especialidade de Obstetrícia após reparametrização.

Após alteração dos valores dos parâmetros observa-se que não se obtém um modelo sem *underfitting*. Nestes casos, por vezes, é sugerida a adição de novas variáveis. No entanto, para efeitos de comparação com os anteriores modelos de RLM tal não se torna possível, logo, mantém-se o modelo obtido com os parâmetros selecionados pela *Grid Search* e correspondente à Figura 19.

Um dos modelos que apresentou *overfitting* foi o modelo correspondente à especialidade de Cirurgia Plástica. Neste caso, a Figura 20 mostra a existência de *overfitting* para o modelo obtido através de combinações de parâmetros com intervalos de valores próximos aos seus valores por omissão.

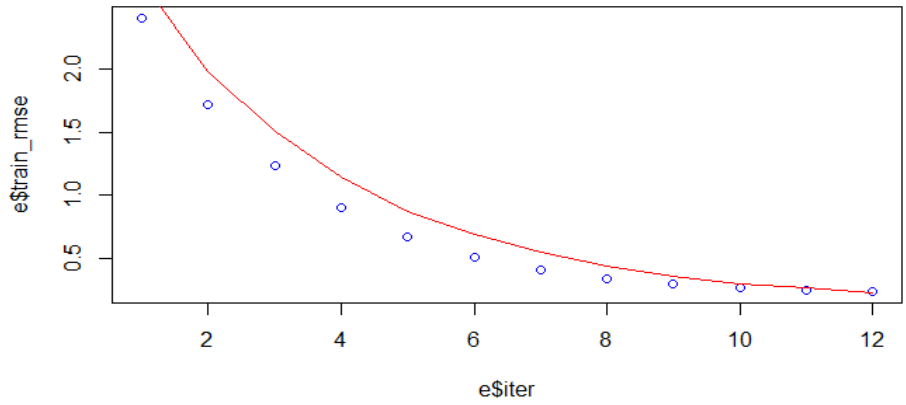


Figura 20 - Resultados RMSE por iteração (dados de treino e dados de teste), para a especialidade de Cirurgia Plástica.

Verificando-se o problema em questão, foram alterados os intervalos de valores para os parâmetros na *Grid Search* que são responsáveis por controlar este tipo de situação de forma a albergarem intervalos de valores mais amplos nos parâmetros: *eta*, *gamma* e *colsample_bytree*. Assim, temos um novo modelo, representado na Figura 21, que tem menos *overfitting* comparativamente ao modelo inicial e, por isso, será o usado para obter os resultados finais.

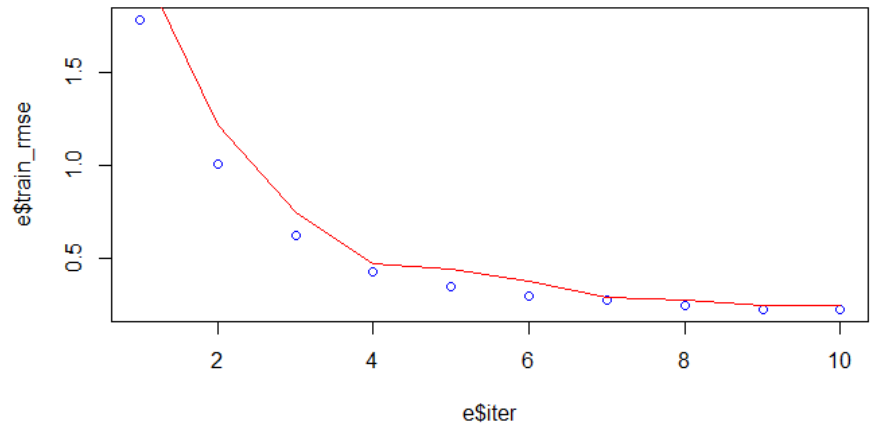


Figura 21 - Resultados RMSE por iteração (dados de treino e dados de teste), para a especialidade de Cirurgia Plástica, após reparametrização.

O mesmo procedimento foi aplicado para as restantes especialidades sendo que, para além das anteriormente analisadas, os modelos que apresentaram evidências de *overfitting* ou *underfitting* foram os modelos correspondentes às especialidades de ORL (*overfitting*) e Pediatria Cirúrgica (*underfitting*). Nestes foram também alterados os valores dos parâmetros e obteve-se um novo modelo com uma menor evidência de *overfitting* e *underfitting*, respetivamente, como se pode observar no Anexo 4, Figura A 4.1.

CAPÍTULO 7 – RESULTADOS

7.1 Indicadores de Desempenho

O R^2 é um indicador comummente utilizado na avaliação de modelos de previsão que demonstra através de uma percentagem quanto da variável dependente é explicada pelas variáveis independentes do modelo. Segundo Healy (1984), o R^2 é enganador quando usado para comparar modelos com um número diferente de variáveis independentes, dado que a soma dos quadrados não diminui quando novas variáveis são adicionadas ao modelo, logo, mesmo que as variáveis adicionadas não sejam significativas, o R^2 aumenta sempre. É também por esta razão que muitas vezes é usado o *Adjusted* R^2 em oposição ao R^2. No entanto, no caso deste TFM, como se pretende comparar modelos de RLM e XGBoost, foi utilizado o R^2 tendo em conta que o cálculo do *Adjusted* R^2 não seria possível nos modelos de XGBoost, cujo *output* não apresenta o número de variáveis independentes em utilização.

O RMSE é calculado através da seguinte fórmula: $RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$. Este indicador é ainda característico por atribuir um elevado peso a erros grandes, logo, torna-se útil quando o objetivo é detetar a presença dos mesmos. Assim, o RMSE é essencial para a avaliação e comparação de modelos na medida em que providencia informação acerca do tamanho do erro produzido pelo modelo na unidade das observações, fazendo com que seja um indicador de fácil interpretação (Willmott, 1981).

Ambos os indicadores permitem uma interpretação melhorada sobre como o modelo se adequa aos dados e da sua capacidade preditiva.

7.2 Comparação entre RLM e XGBoost

Na Tabela 12 encontram-se os valores de R^2 e de RMSE, para os modelos de RLM e XGBoost e para os conjuntos de dados Data Set Agregado e por Especialidade, o que permite ter uma visão geral facilitando a análise e comparação dos modelos de previsão ao resumir os resultados anteriormente apresentados nos Capítulos 5 e 6.

Tabela 12 - Indicadores de desempenho dos modelos RLM e XGBoost.

Modelo	RMSE **RLM**	R^2 **RLM**	RMSE **XGBoost**	R^2 **XGBoost**
Data Set Agregado	0,3772	0,7750	0,3939	0,7547
Data Set por Especialidade				
Cirurgia Geral	0,3919	0,7769	0,4130	0,7523
Cirurgia Plástica	0,2405	0,8559	0,2252	0,8736
Gastrenterologia	0,3309	0,8882	0,2569	0,9326
Ginecologia	0,3553	0,7829	0,2940	0,8514
Obstetrícia	0,2478	0,2873	0,2936	-0,0001
Oftalmologia	0,3245	0,5912	0,3198	0,6031
ORL	0,3687	0,7012	0,4959	0,4597
Ortopedia	0,3537	0,7854	0,3989	0,7269
Pediatria Cirúrgica	0,2125	0,8615	0,1782	0,8299
Urologia	0,3999	0,7151	0,3807	0,7417

7.2.1 Data Set Agregado

No modelo desenvolvido para o Data Set Agregado, a RLM mostra-se superior tendo em conta que o seu R^2 supera em 2,03% o modelo de previsão em XGBoost. Ao observar o RMSE de cada um destes modelos percebemos que esta diferença no R^2 se traduz também num menor RMSE por parte do modelo de RLM, com menos 0,0246 $\log(Duração\ de\ Cirurgia)$.

No que diz respeito às variáveis significativas, apenas se podem comparar as numéricas ou binárias, dado que não é possível identificar a significância das restantes como uma variável só. Assim, no modelo de RLM as duas variáveis mais significativas são MEAN_MEDICO e C_CIRURGIA, enquanto no XGBoost a variável mais importante é a MEAN_ICDPROC, seguindo-se da MEAN_MEDICO. As restantes apresentam percentagens muito baixas de importância, como já foi referido. Apesar da variável MEAN_ICDPROC ser a mais importante para o XGBoost, foi eliminada no processo de *Backward Stepwise Regression* (quarta iteração) da RLM.

Na Figura 22 apresentada, a vermelho representam-se os valores reais de duração de cirurgia e a azul os valores previstos com os modelos de RLM e XGBoost, respetivamente. É percetível a dificuldade dos modelos em prever corretamente valores de $\log(Duração\ de\ Cirurgia)$ abaixo de $y = 2{,}5$, ou seja, durações abaixo de 12 minutos. Por outro lado, a RLM prevê com mais precisão valores longe do habitual como podemos ver em $x = 2000$ e imediatamente abaixo de $x = 5000$.

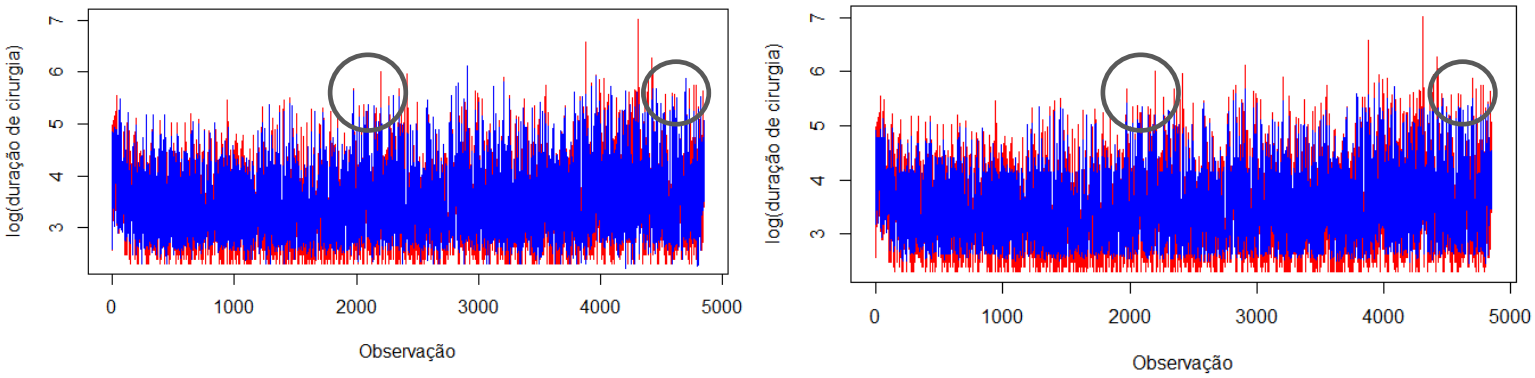


Figura 22 - Valores reais vs. valores previstos, modelo de RLM (esquerda) e modelo XGBoost (direita).

Apesar da diferença entre os modelos não ser significativa, tendo em conta que o XGBoost é um modelo mais complexo, tanto em termos de otimização de parâmetros como da complexidade computacional (tempo gasto ao correr o modelo do XGBoost), é razoável dizer que é preferível a utilização da RLM. Esta produz resultados mais aproximados aos valores reais e é um modelo de fácil interpretação, uma vez que o desempenho do XGBoost está dependente da correta afinação de parâmetros complexos. Adicionalmente, a RLM é muito menos dispendiosa em termos de tempo gasto em cada corrida.

7.2.2 Data Set por Especialidade

O Data Set por Especialidade divide-se em dois grupos, cinco modelos nos quais a RLM se apresenta com uma maior precisão tendo em conta o indicador R^2 e os restantes cinco modelos nos quais o XGBoost supera a RLM.

No modelo de Cirurgia Geral, comparando os resultados do indicador R^2, o modelo de RLM explica uma maior percentagem da variável dependente, dado que $0{,}7769 > 0{,}7523$. Consequentemente, o RMSE é superior no modelo de XGBoost e, assim, constata-se que a RLM produz melhores resultados para esta especialidade. A mesma situação pode ser observada no modelo correspondente à especialidade de Ortopedia. No caso de Obstetrícia, observamos um R^2 negativo para o XGBoost, como referido na análise de resultados do capítulo anterior, demonstrando uma concordância com o modelo de RLM no facto de que os modelos nesta especialidade não se ajustam bem aos dados. Isto pode acontecer por falta de variáveis que expliquem as durações de cirurgias em Obstetrícia, ou seja, pela própria imprevisibilidade inerente aos procedimentos executados nesta especialidade. O modelo de ORL apresenta um R^2 muito inferior no XGBoost, provavelmente devido ao *overfitting* presente, fazendo com que, mesmo após a reparametrização, o modelo continue a efetuar previsões desadequadas ao ser testado em novos dados. Analogamente, na Pediatria Cirúrgica, após reparametrização, ainda há indícios de *underfitting*, logo, é esperado que, apesar do R^2 no XGBoost ser inferior, o seu RMSE apresente um resultado mais favorável em comparação com o modelo de RLM.

Nas restantes cinco especialidades ainda não mencionadas o XGBoost apresentou tanto um melhor poder explicativo (R^2) como um melhor valor de RMSE. O modelo no qual esta melhoria foi mais notória foi o modelo correspondente à especialidade de Ginecologia, apresentando um aumento de 6,85% no valor do indicador R^2 e diminuindo 0,0613 no indicador RMSE.

Analisando todos os modelos, o que obteve o melhor RMSE foi o modelo de Pediatria Cirúrgica, tanto na RLM como no XGBoost. Por outro lado, o pior modelo em termos de RMSE, no caso da RLM corresponde à Oftalmologia enquanto no XGBoost corresponde à ORL, consequência do problema acima referido.

Relativamente às variáveis numéricas e binárias, nos modelos desagregados de XGBoost verifica-se que a variável mais significativa em todos eles, à exceção do modelo correspondente à especialidade de Gastrenterologia, é a variável MEAN_ICDPROC (ver Anexo 4, Figura A 4.3). Contrariamente, e como referido no Capítulo 5, nos modelos de RLM esta variável é sempre excluída em todas as especialidades provavelmente por ter dependências com outra variável, como a ICD_PROC. Posto isto, o XGBoost não tem em consideração a verificação de qualquer pressuposto e assume esta variável como essencial para o modelo final. A variável MEAN_MEDICO apresenta-se como uma das cinco variáveis mais importantes em oito dos dez modelos de XGBoost, à semelhança dos modelos de RLM, nos quais se apresenta como a variável independente numérica que mais frequentemente é incluída nos modelos finais. A variável C_CIRURGIA demonstra ser a segunda variável que mais vezes é significativa para os modelos finais de RLM, sendo que no XGBoost se encontra incluída nas cinco variáveis mais importantes em cinco modelos. As restantes variáveis, INT_ANTERIOR, NUTEIS_C e MEAN_PRIORIDADE, nunca são consideradas variáveis importantes para os modelos de XGBoost, à exceção da variável NUTEIS_C considerada no modelo da especialidade de Ortopedia. No caso dos modelos em RLM, as variáveis INT_ANTERIOR e NUTEIS_C mencionadas também não se demonstram como significativas, não sendo utilizadas em nenhum modelo, sendo que a variável MEAN_PRIORIDADE é considerada apenas uma vez.

Nos modelos de RLM, quando comparados os modelos obtidos para ambos os conjuntos de dados em relação ao R^2, observa-se que em quatro das dez especialidades é preferível utilizar o modelo agregado, nomeadamente nas especialidades de Obstetrícia, Oftalmologia, ORL e Urologia. No entanto, quando a comparação é feita tendo por base o RMSE, em oito das dez especialidades é preferível utilizar os modelos obtidos através do Data Set RLM por Especialidade. No caso dos modelos em XGBoost, observa-se que em seis especialidades o R^2 obtido não justifica a utilização do modelo por especialidade. Relativamente à análise do RMSE, nos modelos de XGBoost, sete de dez especialidades obtêm melhores resultados de RMSE nos modelos do Data Set XGBoost por Especialidade do que no modelo agregado.

Com o objetivo de fazer uma comparação entre o comportamento dos modelos para as especialidades com poucas observações, nomeadamente Cirurgia Plástica e ORL, fizeram-se dois gráficos que representam os valores reais e os valores previstos.

Na Figura 23 encontram-se representados os gráficos associados à especialidade de Cirurgia Plástica.

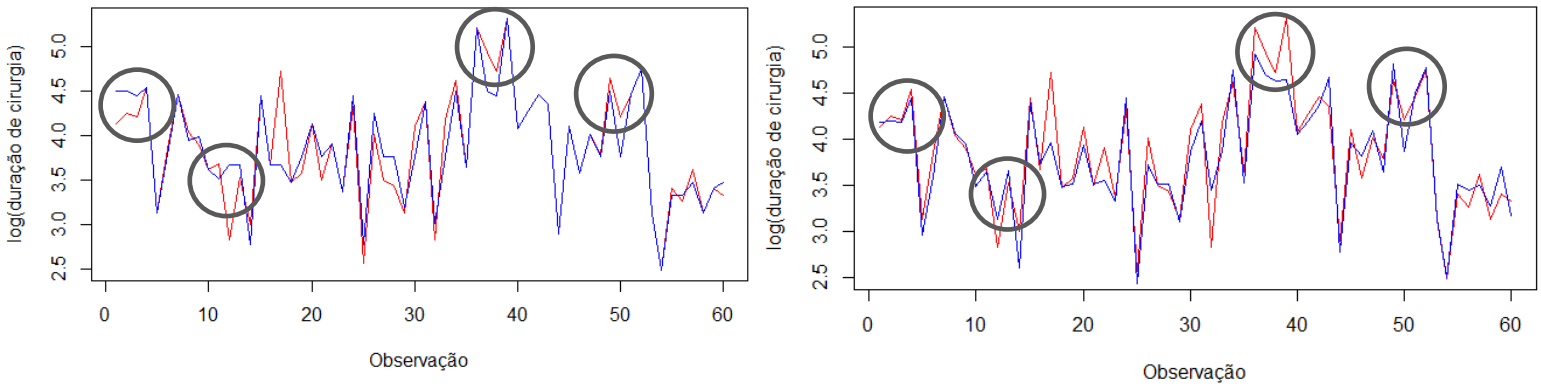


Figura 23 - Cirurgia Plástica: valores reais vs. valores previstos, modelo de RLM (esquerda) e modelo XGBoost (direita).

Consegue-se perceber que o XGBoost se mostra mais preciso no intervalo entre $\log(Duração\ de\ Cirurgia) = 3$ e $\log(Duração\ de\ Cirurgia) = 4{,}5$, no entanto, ainda que o R^2 deste modelo seja superior ao da RLM, a RLM mostra-se mais útil quando as durações atingem valores não contidos no intervalo anteriormente mencionado.

Numa outra perspetiva observa-se agora o segundo conjunto de modelos, correspondente à especialidade de ORL, através da Figura 24.

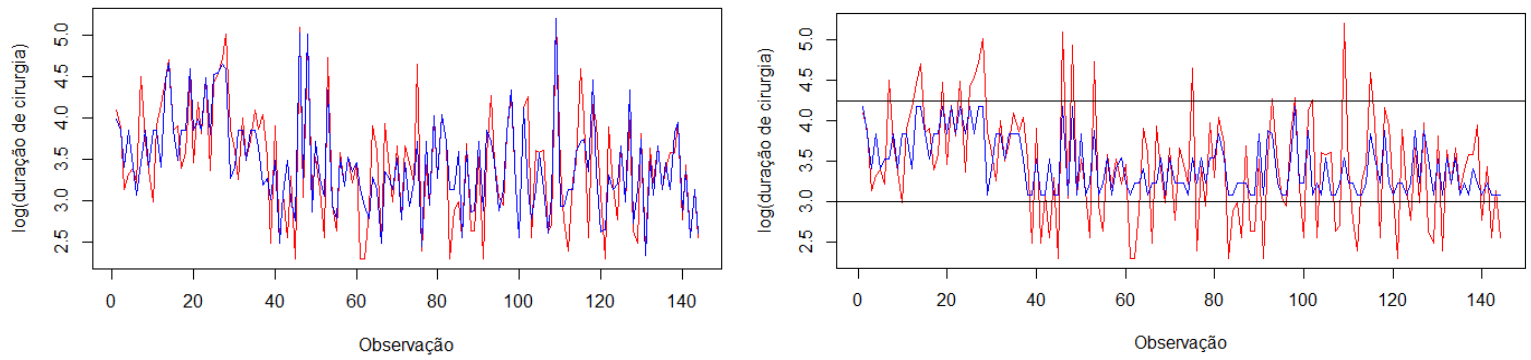


Figura 24 - ORL: valores reais vs. valores previstos, modelo de RLM (esquerda) e modelo XGBoost (direita).

Verifica-se que o XGBoost falha em termos de precisão em $\log(Duração\ de\ Cirurgia) > 4{,}25$ e $\log(Duração\ de\ Cirurgia) < 3$, havendo um pequeno intervalo no qual as suas previsões são aceitáveis em comparação com os valores reais. Tal vai ao encontro da diminuição de 2,45% no R^2 do modelo em XGBoost e ao problema do *overfitting*, comprovando ainda que o modelo não tem uma boa adaptação quando testado em novos dados. Esta fraca precisão em determinados intervalos de durações aparenta ser

um padrão no XGBoost, tendo em conta que a mesma situação foi detetada nos gráficos anteriormente analisados, que pode ser causada pela reduzida quantidade de observações disponíveis em cada modelo.

Os gráficos referentes às especialidades não mencionadas neste Capítulo podem ser consultados no Anexo 4.

CAPÍTULO 8 - CONCLUSÃO

O principal objetivo deste TFM era a elaboração de modelos de previsão de durações de cirurgias que efetuassem previsões o mais precisas quanto possível. Para cumprir tal objetivo foi necessário estabelecer uma metodologia que fosse adequada tendo em conta os modelos escolhidos (RLM e XGBoost) e replicável em diferentes conjuntos de dados (Data Set Agregado e Data Set por Especialidade). Os modelos obtidos têm como variável dependente a duração da cirurgia e como variáveis independentes um conjunto diverso de variáveis numéricas, binárias e categóricas.

Ao longo da construção dos modelos de previsão foram-se observando algumas dificuldades. No caso da RLM, o facto de haver modelos com variáveis independentes numéricas e categóricas dificultou o processo de seleção de variáveis, tendo em conta o elevado número de diferentes categorias presentes nas variáveis independentes categóricas. Por outro lado, os modelos em XGBoost, visto que o modelo de previsão é uma caixa preta, apresentaram uma dificuldade acrescida em definir uma metodologia que permitisse otimizar ao máximo a escolha de valores para cada um dos parâmetros do modelo.

Considerando os resultados obtidos, uma das conclusões que se pode tirar é que, de facto, a maioria dos modelos de RLM e XGBoost apresentam uma performance bastante satisfatória, sendo que em alguns casos o XGBoost mostrou uma melhor performance do que a RLM. Apesar disto, a diferença entre modelos não é significativa, pelo que se pode concluir que, para este conjunto de dados, o modelo de RLM mostra-se mais vantajoso, tendo em conta o tempo despendido na afinação de parâmetros de forma a obter um modelo de XGBoost satisfatório.

Ao analisar o comportamento de ambos os conjuntos de dados percebe-se que os modelos por especialidade nem sempre têm uma melhor performance do que os modelos agregados, como seria de esperar uma vez que, como referido, a duração da cirurgia está dependente da especialidade associada. Isto leva a uma das limitações deste TFM, nomeadamente o número de observações para cada uma das especialidades que pode não ser suficiente para se obter um modelo por especialidade com melhor performance do que um modelo agregado. Outro fator limitante é a falta de informação essencial que pode condicionar a duração da cirurgia, sendo a idade e o género do paciente dois exemplos frequentemente usados em estudos semelhantes e que não foram disponibilizados. Por fim, e tendo em consideração que é necessário um conhecimento médico para determinar quais fatores poderão ter uma maior influência na duração de cirurgias, apenas são determinadas as variáveis estatisticamente significativas para o modelo, não havendo uma análise crítica acerca dos resultados obtidos.

A título de investigação futura, assumindo que existe uma maior disponibilidade de informação por parte do HESE, seria interessante fazer o mesmo estudo, com uma amostra de maior dimensão e com

variáveis importantes, e neste caso não disponibilizadas, para verificar se de facto um modelo de previsão por especialidade se torna mais preciso do que um modelo agregado. Adicionalmente, os resultados deviam ser discutidos com a gestão hospitalar do HESE de forma a perceber a interpretação que dão aos resultados e que ajustes devem ser efetuados aos modelos obtidos.

BIBLIOGRAFIA

Allen, M. P. (1997). The problem of multicollinearity. In *Understanding Regression Analysis* (pp. 176–180). Springer New York, NY. https://doi.org/https://doi.org/10.1007/b102242

Bartek, M. A., Saxena, R. C., Solomon, S., Fong, C. T., Behara, L. D., Venigandla, R., Velagapudi, K., Lang, J. D., & Nair, B. G. (2019). Improving Operating Room Efficiency: Machine Learning Approach to Predict Case-Time Duration. *Journal of the American College of Surgeons*, *229*(4), 346–354. https://doi.org/10.1016/j.jamcollsurg.2019.05.029

Belsley, D. A. (1982). ASSESSING THE PRESENCE OF HARMFUL COLLINEARITY AND OTHER FORMS OF WEAK DATA THROUGH A TEST FOR SIGNAL-TO-NOISE. *Journal of Econometrics*, *20*, 211–253.

Cardoen, B., Demeulemeester, E., & Beliën, J. (2010). Operating room planning and scheduling: A literature review. *European Journal of Operational Research*, *201*(3), 921–932. https://doi.org/10.1016/j.ejor.2009.04.011

Devi, s & Rao, K & Sangeetha, S. (2010). Prediction of Surgery Times and Scheduling of Operation Theaters in Optholmology Department. Journal of medical systems. 36. 415-30. 10.1007/s10916-010-9486-z.

Dexter F, Macario A, Traub RD, Hopwood M, Lubarsky DA. An operating room scheduling strategy to maximize the use of operating room block time: computer simulation of patient scheduling and survey of patients' preferences for surgical waiting time. Anesth Analg. 1999 Jul;89(1):7-20. doi: 10.1097/00000539-199907000-00003.

Dietterich, T. G. (2000). Ensemble Methods in Machine Learning. In *Lecture Notes in Computer Science* (Vol. 1857, pp. 1–15). Springer, Berlin, Heidelberg. https://doi.org/https://doi.org/10.1007/3-540-45014-9_1

Dong, W., Huang, Y., Lehane, B., & Ma, G. (2020). XGBoost algorithm-based prediction of concrete electrical resistivity for structural health monitoring. *Automation in Construction*, *114*. https://doi.org/10.1016/j.autcon.2020.103155

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., Marquéz, J. R. G., Gruber, B., Lafourcade, B., Leitão, P. J., Münkemüller, T., Mcclean, C., Osborne, P. E., Reineking, B., Schröder,

B., Skidmore, A. K., Zurell, D., & Lautenbach, S. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Eberly, L. E. (2007). Multiple Linear Regression. In *Methods in Molecular Biology* (Vol. 404, pp. 165–187). https://doi.org/10.1007/978-1-59745-530-5_9

el Naqa, I., & Murphy, M. J. (2015). What Is Machine Learning? In *Machine Learning in Radiation Oncology* (pp. 3–11). Springer International Publishing. https://doi.org/10.1007/978-3-319-18305-3_1

Friedman, J. H. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, *38*(4), 367–378. www.elsevier.com/locate/csda

Hawkins, Douglas. (2004). The Problem of Overfitting. Journal of chemical information and computer sciences. 44. 1-12. 10.1021/ci0342472.

Healy, M. J. R. (1984). The Use of R^2 as a Measure of Goodness of Fit. *Journal of the Royal Statistical Society*, *147*(4), 608–609.

Hosseini, N., Sir, M. Y., Jankowski, C. J., & Pasupathy, K. S. (2015). Surgical Duration Estimation via Data Mining and Predictive Modeling: A Case Study. *AMIA Annu Symp Proc*, 640–648.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R Second Edition*. Springer New York, NY. https://doi.org/https://doi.org/10.1007/978-1-4614-7138-7

Kayış, E., Khaniyev, T. T., Suermondt, J., & Sylvester, K. (2015). A robust estimation model for surgery durations with temporal, operational, and surgery team effects. *Health Care Management Science*, *18*(3), 222–233. https://doi.org/10.1007/s10729-014-9309-8

Kotsiantis, S. B. (2007). Supervised Machine Learning: A Review of Classification Techniques. *Informatica*, *31*(3), 249–268.

Kurz, C.F., Maier, W. & Rink, C. A greedy stacking algorithm for model ensembling and domain weighting. BMC Res Notes 13, 70 (2020). https://doi.org/10.1186/s13104-020-4931-7

Larsson, A. (2013). The accuracy of surgery time estimations. *Production Planning and Control*, *24*(10–11), 891–902. https://doi.org/10.1080/09537287.2012.666897

Laskin, Daniel & Abubaker, A. & Strauss, Robert. (2013). Accuracy of Predicting the Duration of a Surgical Operation. Journal of oral and maxillofacial surgery : official journal of the American Association of Oral and Maxillofacial Surgeons. 71. 446-7.

Martinez, O., Martinez, C., Parra, C. A., Rugeles, S., & Suarez, D. R. (2021). Machine learning for surgical time prediction. *Computer Methods and Programs in Biomedicine*, *208*. https://doi.org/10.1016/j.cmpb.2021.106220

Ng, N. H., Gabriel, R. A., Mcauley, J., Elkan, C., & Lipton, Z. C. (2017). Predicting Surgery Duration with Neural Heteroscedastic Regression. *Proceedings of the 2nd Machine Learning for Healthcare Conference*, 100–111.

Olaronke, Iroju & Soriyan, Abimbola & Gambo, Ishaya & Olaleke, J.. (2013). Interoperability in Healthcare: Benefits, Challenges and Resolutions. International Journal of Innovation and Applied Studies. 3. 2028-9324.

Parsa, Amir Bahador & Movahedi, Ali & Taghipour, Homa & Derrible, Sybil & Mohammadian, Abolfazl. (2019). Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. Accident; analysis and prevention. 136. 105405. 10.1016/j.aap.2019.105405.

Pasha, G. R. (2002). SELECTION OF VARIABLES IN MULTIPLE REGRESSION USING STEPWISE REGRESSION. *Journal of Research (Science)*, *13*(2), 119–127.

Refaeilzadeh, P., Tang, L., & Liu, H. (2016). Cross-Validation. In *Encyclopedia of Database Systems* (pp. 1–7). Springer New York. https://doi.org/10.1007/978-1-4899-7993-3_565-2

Riekert, Martin & Premm, Marc & Klein, Achim & Kirilov, Lyubomir & Kenngott, Hannes & Apitz, Martin & Wagner, Martin & Ternes, Lena. (2017). Predicting the Duration of Surgeries to Improve Process Efficiency in Hospitals.

Ripon, Kazi Shah Nawaz & Nyman, Jacob. (2020). Hospital Surgery Scheduling Under Uncertainty Using Multiobjective Evolutionary Algorithms. 10.1007/978-3-030-31672-3_7.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Shahabikargar, Z., Khanna, S., Sattar, A., & Lind, J. (2017). Improved prediction of procedure duration for elective surgery. *Studies in Health Technology and Informatics*, *239*, 133–138. https://doi.org/10.3233/978-1-61499-783-2-133

Shinde, P. P., & Shah, S. (2018). A Review of Machine Learning and Deep Learning Applications. *Fourth International Conference on Computing Communication Control and Automation*. https://doi.org/10.1109/ICCUBEA.2018.8697857

Shi, Xiupeng & Wong, Yiik & Li, Michael & Palanisamy, Chandrasekar & Chai, Chen. (2019). A feature learning approach based on XGBoost for driving assessment and risk prediction. Accident Analysis & Prevention. 129. 170-179. 10.1016/j.aap.2019.05.005.

Strum, David P & May, Jerrold H & Vargas, Luis. (2000). Modeling the Uncertainty of Surgical Procedure Times: Comparison of Log-normal and Normal Models. Anesthesiology. 92. 1160-7. 10.1097/00000542-200004000-00035.

Sutherland, James & Peet, A. & Soulsby, R.. (2004). Evaluating the performance of morphological models. Coastal Engineering - COAST ENG. 51. 917-939. 10.1016/j.coastaleng.2004.07.015.

Tabachnick, B. G., & Fidell, L. S. (2007). *Experimental designs using ANOVA*. Thomson/Brooks/Cole.

Tan, K. W., Hoang, F. N., Nguyen, L., Gan, J., Shao, S., & Lam, W. (2019). Data-Driven Surgical Duration Prediction Model for Surgery Scheduling: A Case-Study for a Practice-Feasible Model in a Public Hospital. *IEEE 15th International Conference on Automation Science and Engineering (CASE)*, 275–280. https://doi.org/10.1109/COASE.2019.8843299

Thayer, J. D. (2002). *Stepwise Regression as an Exploratory Data Analysis Procedure*. Distributed by ERIC Clearinghouse. https://eric.ed.gov/?id=ED464932

Willmott, C. J. (1981). On the validation of models. *Physical Geography*, *2*(2), 184–194. https://doi.org/10.1080/02723646.1981.10642213

Wissmann, M., & Toutenburg, H. (2007). Role of Categorical Variables in Multicollinearity in the Linear Regression Model Role of Categorical Variables in Multicollinearity in Linear Regression Model. *Journal of Applied Statistical Science*.

Yu, B., Qiu, W., Chen, C., Ma, A., Jiang, J., Zhou, H., & Ma, Q. (2019). SubMito-XGBoost: predicting protein submitochondrial localization by fusing multiple feature information and eXtreme gradient boosting. *Bioinformatics*, *36*(4), 1074–1081. https://doi.org/10.1093/bioinformatics/btz734/5585744

Yuniartha, D. R., Masruroh, N. A., & Herliansyah, M. K. (2021). An evaluation of a simple model for predicting surgery duration using a set of surgical procedure parameters. *Informatics in Medicine Unlocked*, *25*. https://doi.org/10.1016/j.imu.2021.100633

Zhang, Y., & Haghani, A. (2015). A gradient boosting method to improve travel time prediction. *Transportation Research Part C: Emerging Technologies*, *58*, 308–324. https://doi.org/10.1016/j.trc.2015.02.019

Zhao, Jane & Forsythe, Raquel & Langerman, Alexander & Melton, Genevieve & Schneider, David & Jackson, Gretchen. (2020). The Value of the Surgeon Informatician. Journal of Surgical Research. 252. 10.1016/j.jss.2020.04.003.

Zhu, S., Fan, W., Yang, S., Pei, J., & Pardalos, P. M. (2019). Operating room planning and surgical case scheduling: a review of literature. *Journal of Combinatorial Optimization*, *37*(3), 757–805. https://doi.org/10.1007/s10878-018-0322-6

ANEXOS

Anexo 1 – Validação de Pressupostos 1

Figura A 1.1 – Histogramas da distribuição da variável dependente, no Data Set RLM por Especialidade.

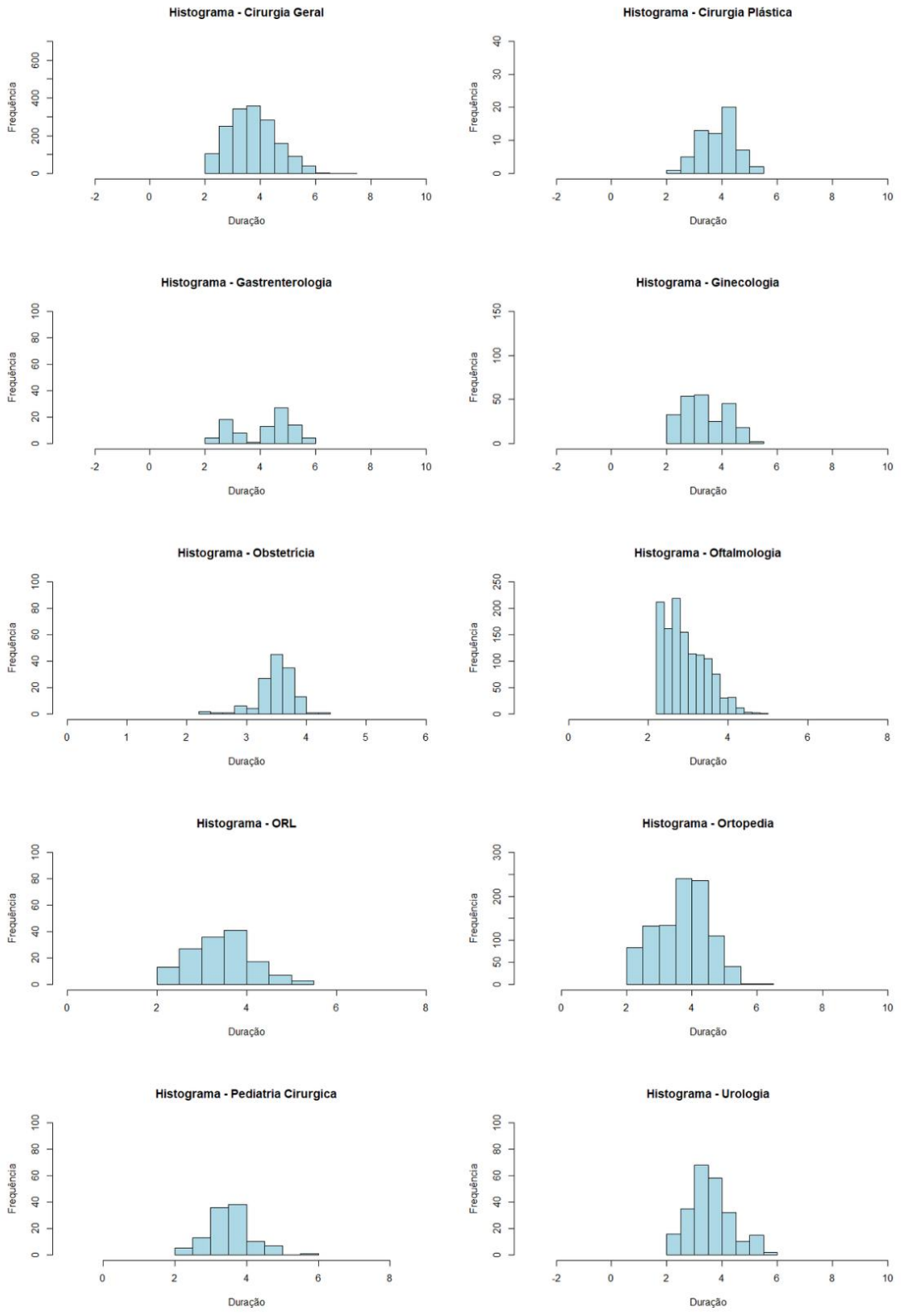


Tabela A 1.2 – Valores VIF para variáveis numéricas e binárias, no Data Set RLM por Especialidade.

	INT_ANTERIOR	OBESIDADE_C	NUTEIS_C	T_ESPERA	C_CIRURGIA	MEAN_MEDICO	MEAN_PRIORIDADE	MEAN_ICDPROC
Cirurgia Geral	1,0790		1,0790	1,6623	1,5376	1,5003	1,7085	1,4045
Cirurgia Plástica	1,1223		1,0937	1,3318	1,1770	1,5562	1,4770	1,3426
Gastrenterologia	1,4990	**390,1536**	1,0946	7,4756	1,3295	**343,2938**	5,5996	**28,9100**
Ginecologia	1,2473		1,0966	1,2847	2,9217	1,3843	1,2970	2,9788
Obstetrícia	**42,1003**		1,0519	5,3847	2,7189	1,3222	1,1345	**40,7780**
Oftalmologia	1,0540		1,0727	1,2524	1,3768	1,3167	1,7525	1,3749
ORL	1,0366		1,1758	2,1321	1,3182	1,0929	2,3472	1,4234
Ortopedia	1,0252		1,1606	1,6575	2,4581	1,5275	1,2681	1,8503
Pediatria Cirúrgica	1,0916		1,0888	1,1253	1,0822	1,7616	1,3208	1,5582
Urologia	1,0768		1,0984	1,7713	3,7869	3,8171	2,4152	1,2991

Anexo 2 – Seleção de Variáveis e Modelo Final

Tabela A 2.1 – P-Value resultante do teste da ANOVA.

	INT_ANTERIOR	ICD_PROC	ICD_DIAG	C_PRIORIDADE	SEMANA_C	MES_C	OBESIDADE_C	C_CIRURGIA	GDH	COD_EXTRACAO	COD_GNOSOLOGICO	ESPECIALIDADE_C	NUTEIS_C
Data Set RLM Agregado	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000	0,0000
Data Set RLM por Especialidade													
Cirurgia Geral	0,9010	0,0000	0,0000	0,0000	0,0000	0,0000		0,0000	0,0000	0,0000	0,0000		0,0000
Cirurgia Plástica	0,4380	0,0008	0,0002	0,8040	0,0430	0,2970		0,0746	0,0045	0,0001	0,0001		0,0481
Gastrenterologia	0,3660	0,0000	0,0000	0,0000	0,0109	0,5340		0,0119	0,0000	0,0000	0,0000		0,2600
Ginecologia	0,0000	0,0000	0,0000	0,0545	0,0000	0,0557		0,0000	0,0000	0,1210	0,0741		0,1430
Obstetrícia		0,0007	0,1170	0,1160	0,2930	0,1760		0,0002	0,0059	0,0002			0,5300
Oftalmologia	0,0097	0,0000	0,0000	0,0000	0,0000	0,0003		0,0000	0,0000	0,0000	0,2210	0,0000	
ORL	0,9420	0,0000	0,0000	0,0071	0,5800	0,3960		0,0000	0,0084	0,4070	0,5630	0,8880	
Ortopedia	0,0053	0,0000	0,0000	0,0000	0,0000	0,0026		0,0000	0,0000	0,0000	0,0000	0,0000	
Pediatria Cirúrgica	0,6000	0,0000	0,0000	0,0439	0,0249	0,0106		0,3190	0,0000	0,0000	0,0000	0,4110	
Urologia	0,0035	0,0000	0,0000	0,0000	0,0000	0,0005		0,0000	0,0000	0,0000	0,0000	0,0512	

Tabela A 2.2 – Algoritmo Backward Stepwise Regression com p-value de exclusão, no Data Set RLM Agregado.

	Iteração 1	Iteração 2	Iteração 3	Iteração 4	Iteração 5
INT_ANTERIOR	0,0011	0,0012	0,0012	0,0011	0,0011
OBESIDADE_C	0,0883	0,0882	0,0882	0,0882	0,0003
NUTEIS_C	0,2312	0,2316	0,2316		
MEAN_MEDICO	0,0000	0,0000	0,0000	0,0000	0,0000
T_ESPERA	0,7682				
C_CIRURGIA	0,0000	0,0000	0,0000	0,0000	0,0000
MEAN_PRIORIDADE	0,6342	0,6226			
MEAN_ICDPROC	0,1000	0,0998	0,0998	0,0998	
R^2	0,7751	0,7751	0,7751	0,7750	0,7750

Tabela A 2.3 – Iteração 1 da Backward Septwise Regression com p-value de exclusão, no Data Set RLM por Especialidade.

Iteração 1

	INT_ANTERIOR	T_ESPERA	NUTEIS_C	C_CIRURGIA	MEAN_MEDICO	MEAN_PRIORIDADE
Cirurgia Geral		**0,5964**		0,0001	0,0000	
Cirurgia Plástica		0,2543			0,4102	**0,7370**
Gastrenterologia		0,0220		**0,7763**		
Ginecologia	**0,3633**	0,0155		0,3397	0,0626	0,1401
Obstetricia		**0,1815**			0,0000	0,0121
Oftalmologia	0,0379	**0,2560**		0,1786	0,0000	
ORL		0,6558		0,0232	**0,7769**	
Ortopedia	0,0499	**0,0729**	0,0667	0,0000	0,0000	
Pediatria Cirúrgica		0,1736			**0,4302**	
Urologia	0,2612	0,7157		0,4709	**0,7262**	

Tabela A 2.4 – Iteração 2 da Backward Septwise Regression com p-value de exclusão, no Data Set RLM por Especialidade.

Iteração 2

	INT_ANTERIOR	T_ESPERA	NUTEIS_C	C_CIRURGIA	MEAN_MEDICO	MEAN_PRIORIDADE
Cirurgia Geral				0,0001	0,0000	
Cirurgia Plástica		0,2376			**0,3652**	
Gastrenterologia		0,0209				
Ginecologia		0,0157		**0,3397**	0,0372	0,1137
Obstetricia					0,0000	0,0276
Oftalmologia	0,0454			**0,1856**	0,0000	
ORL		**0,6830**		0,0179		
Ortopedia	0,0611		**0,0714**	0,0000	0,0000	
Pediatria Cirúrgica		**0,2319**				
Urologia	0,2659	**0,6395**		0,5219		

Tabela A 2.5 – Iteração 3 da Backward Septwise Regression com p-value de exclusão, no Data Set RLM por Especialidade.

Iteração 3

	INT_ANTERIOR	T_ESPERA	NUTEIS_C	C_CIRURGIA	MEAN_MEDICO	MEAN_PRIORIDADE
Cirurgia Geral						
Cirurgia Plástica		**0,3195**				
Gastrenterologia						
Ginecologia		0,0538			0,0372	**0,1137**
Obstetricia					0,0000	0,0276
Oftalmologia	**0,0560**				0,0000	
ORL						
Ortopedia	**0,0541**			0,0000	0,0000	
Pediatria Cirúrgica						
Urologia	0,2502			**0,5960**		

Tabela A 2.6 – Iteração 4 da Backward Septwise Regression com p-value de exclusão, no Data Set RLM por Especialidade.

Iteração 4

	INT_ANTERIOR	T_ESPERA	NUTEIS_C	C_CIRURGIA	MEAN_MEDICO	MEAN_PRIORIDADE
Cirurgia Geral						
Cirurgia Plástica						
Gastrenterologia						
Ginecologia		**0,0538**			0,0372	
Obstetricia						
Oftalmologia					0,0000	
ORL						
Ortopedia				0,0000	0,0000	
Pediatria Cirúrgica						
Urologia	**0,2403**					

Tabela A 2.7 – Iteração 5 da Backward Septwise Regression com p-value de exclusão, no Data Set RLM por Especialidade.

Iteração 5

	INT_ANTERIOR	T_ESPERA	NUTEIS_C	C_CIRURGIA	MEAN_MEDICO	MEAN_PRIORIDADE
Cirurgia Geral						
Cirurgia Plástica						
Gastrenterologia						
Ginecologia					0,0440	
Obstetricia						
Oftalmologia						
ORL						
Ortopedia						
Pediatria Cirúrgica						
Urologia						

Tabela A 2.8 – Valores β da constante e das variáveis numéricas e binárias obtidos nos modelos finais.

	Constante	INT_ANTERIOR	T_ESPERA	NUTEIS_C	C_CIRURGIA	OBESIDADE_C	MEAN_MEDICO	MEAN_PRIORIDADE	MEAN_ICDPROC
Data Set RLM Agregado	2,0832	0,0611			0,2374	4,1262	0,0055		
Data Set RLM por Especialidade									
Cirurgia Geral	2,8941				0,0497		0,0008		
Cirurgia Plástica	5,3181								
Gastrenterologia	5,3152		0,0002						
Ginecologia	2,7073						0,0020		
Obstetrícia	1,3886						0,0096	0,0054	
Oftalmologia	1,8429						0,003		
ORL	2,4312				0,1828				
Ortopedia	3,2008				0,0882		0,0016		
Pediatria Cirúrgica	2,5637								
Urologia	2,8548								

Anexo 3 – Validação de Pressupostos 2 (Data Set RLM por Especialidade)

Figura A 3.1 – Q-Q Plot dos Resíduos (da esquerda para a direita): Cirurgia Geral, Cirurgia Plástica, Gastrenterologia, Ginecologia, Obstetrícia, Oftalmologia, ORL, Ortopedia, Pediatria Cirúrgica e Urologia.

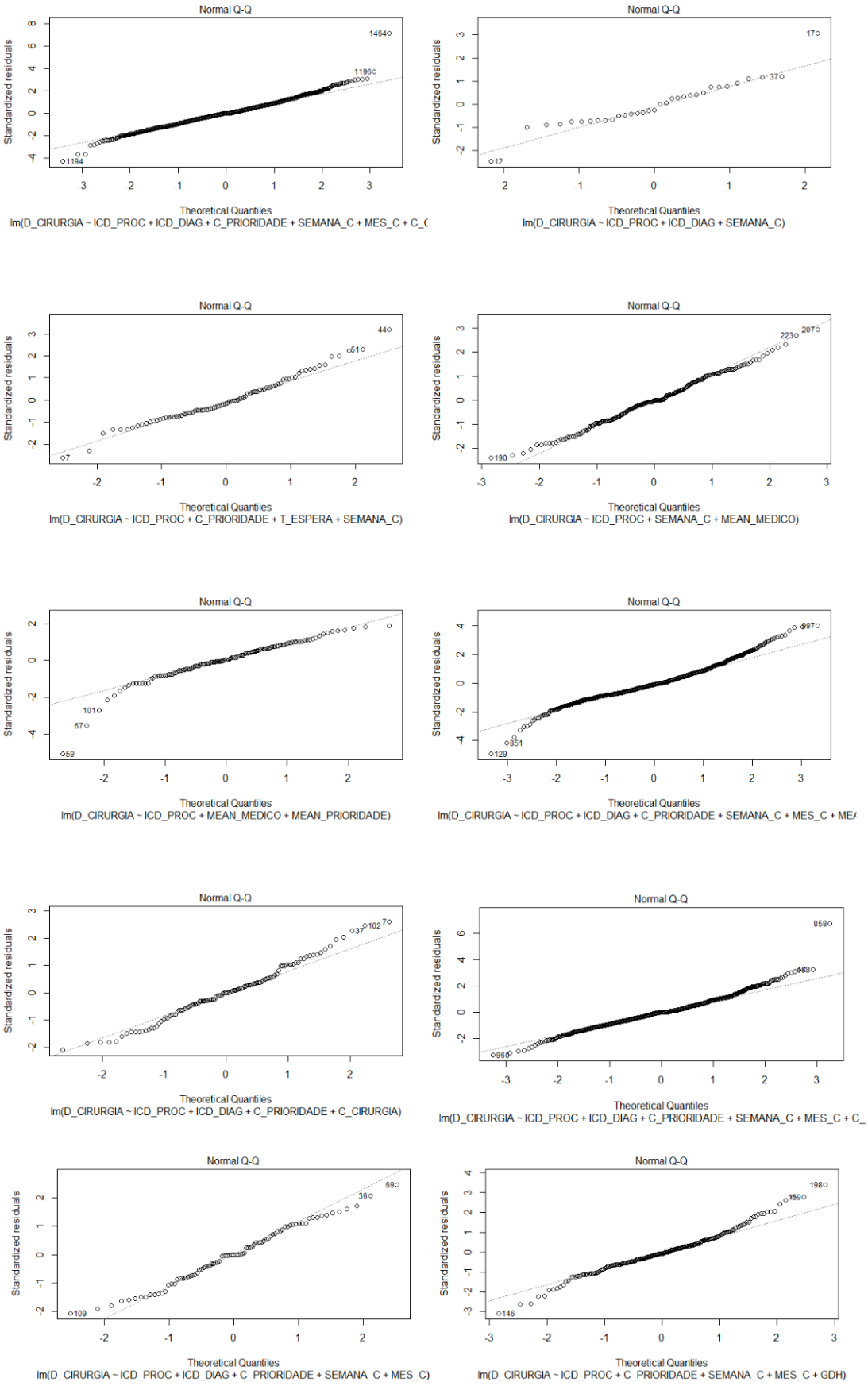


Tabela A 3.2 – Resultados do teste de Durbin-Watson, no Data Set RLM por Especialidade.

	DW	***P-Value***
Cirurgia Geral	**2,0178**	0,7605
Cirurgia Plástica	**1,6922**	0,1500
Gastrenterologia	**1,7052**	0,0531
Ginecologia	**1,7834**	0,0800
Obstetrícia	**1,9488**	0,3608
Oftalmologia	**2,0059**	0,9007
ORL	**1,9278**	0,7645
Ortopedia	**2,0239**	0,7628
Pediatria Cirurgica	**1,9034**	0,7702
Urologia	**2,1339**	0,7622

Figura A 3.3 – Gráficos de dispersão (da esquerda para a direita): Cirurgia Geral, Cirurgia Plástica, Gastrenterologia, Ginecologia, Obstetrícia, Oftalmologia, ORL, Ortopedia, Pediatria Cirúrgica e Urologia.

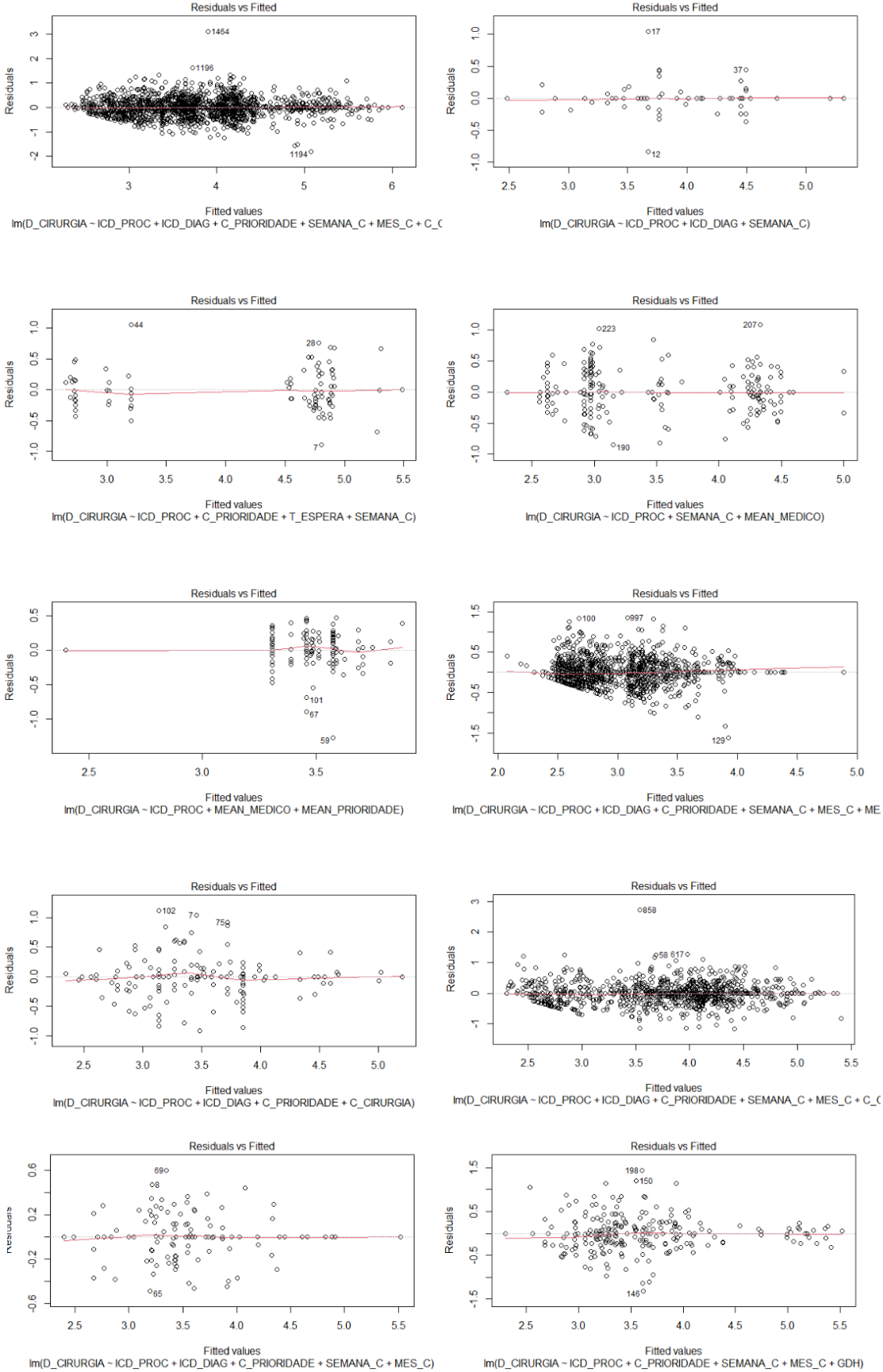
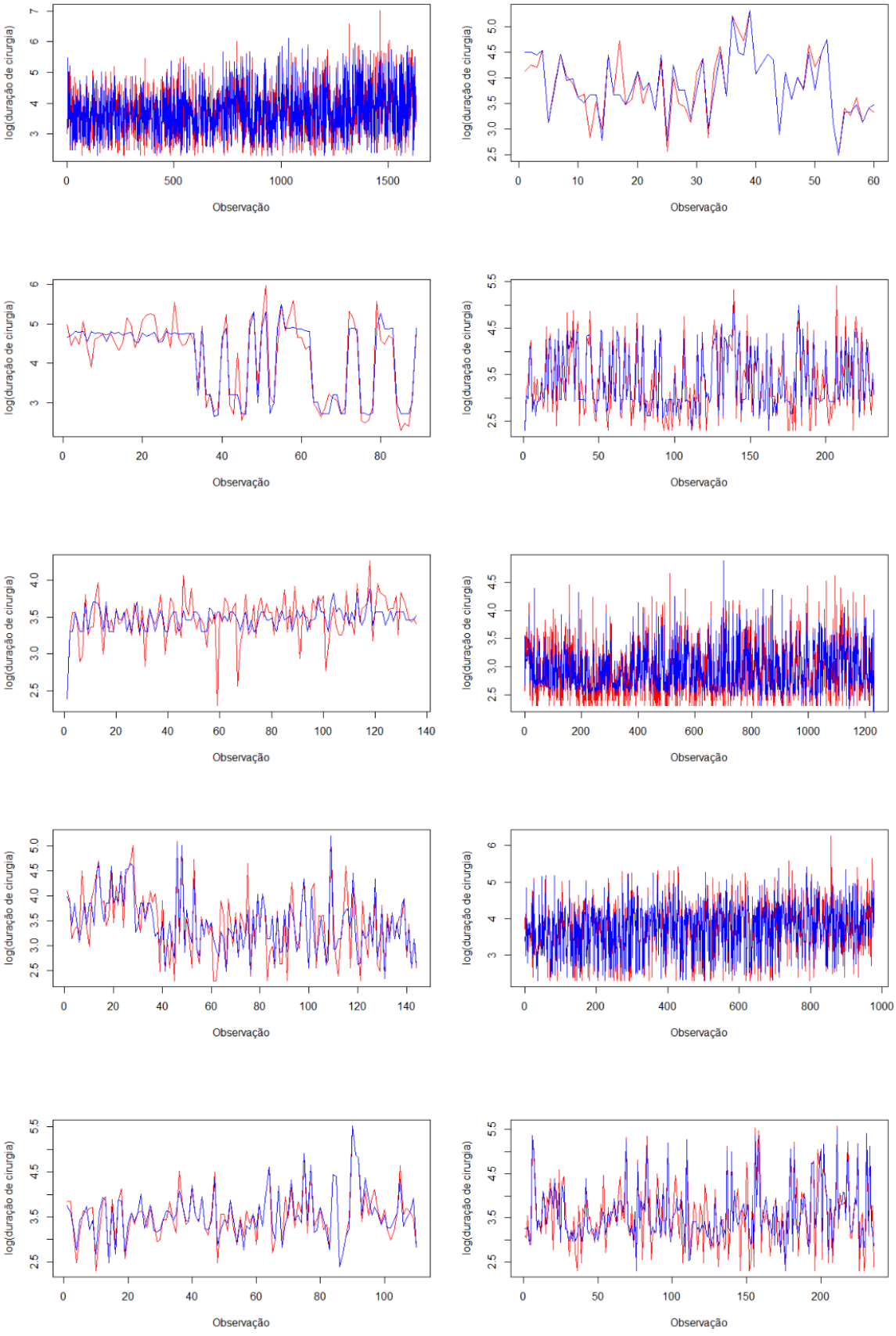


Figura A 3.4 – Valores reais vs. Valores Previstos (da esquerda para a direita): Cirurgia Geral, Cirurgia Plástica, Gastrenterologia, Ginecologia, Obstetrícia, Oftalmologia, ORL, Ortopedia, Pediatria Cirúrgica e Urologia.



Anexo 4 – XGBoost

Figura A 4.1 – Resultados RMSE por iteração (da esquerda para a direita): Cirurgia Geral, Cirurgia Plástica, Gastrenterologia, Ginecologia, Obstetrícia, Oftalmologia, ORL, Ortopedia, Pediatria Cirúrgica e Urologia.

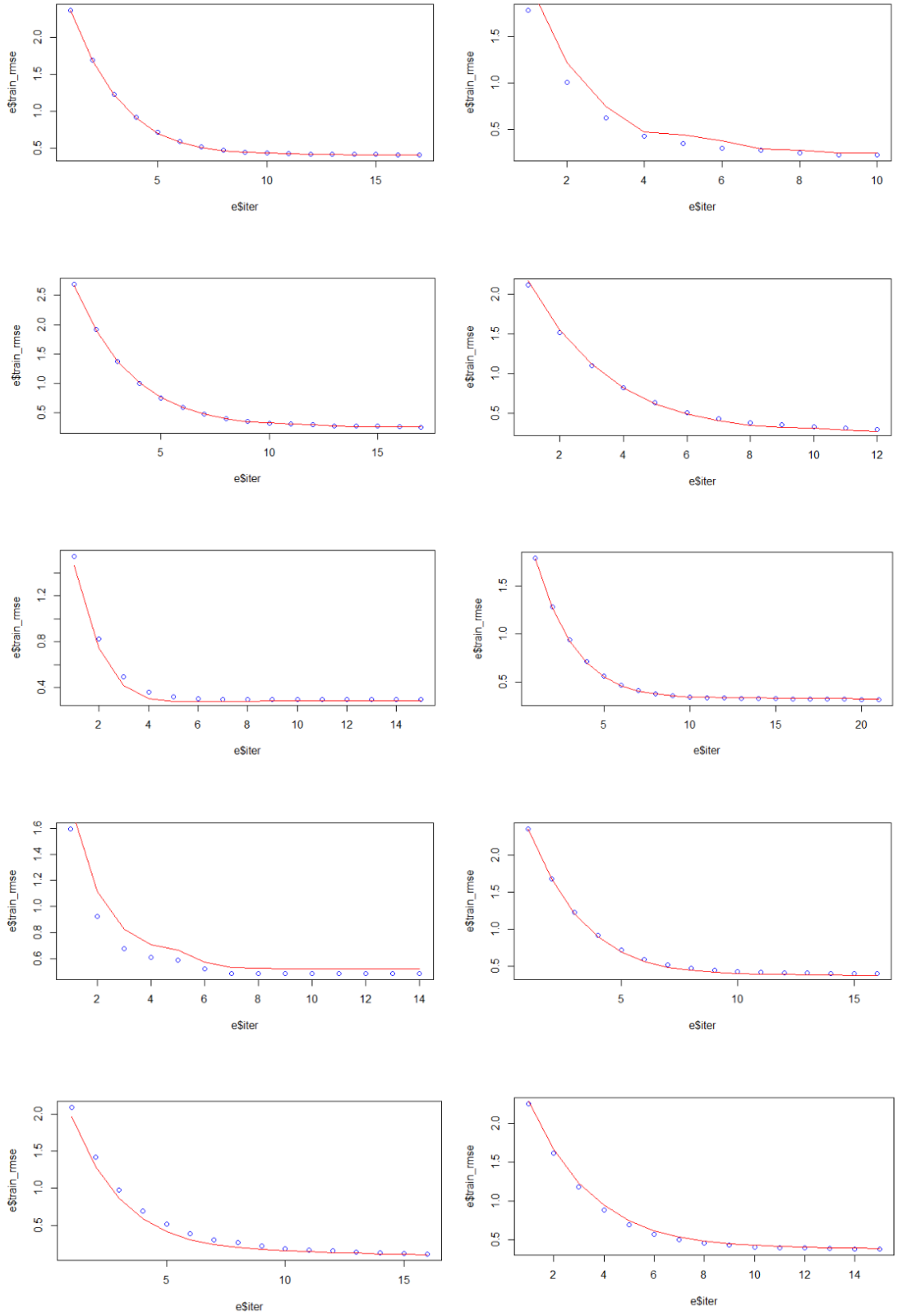


Figura A 4.2 – Valores SHAP (da esquerda para a direita): Cirurgia Geral, Cirurgia Plástica, Gastrenterologia, Ginecologia, Oftalmologia, ORL, Ortopedia, Pediatria Cirúrgica e Urologia.

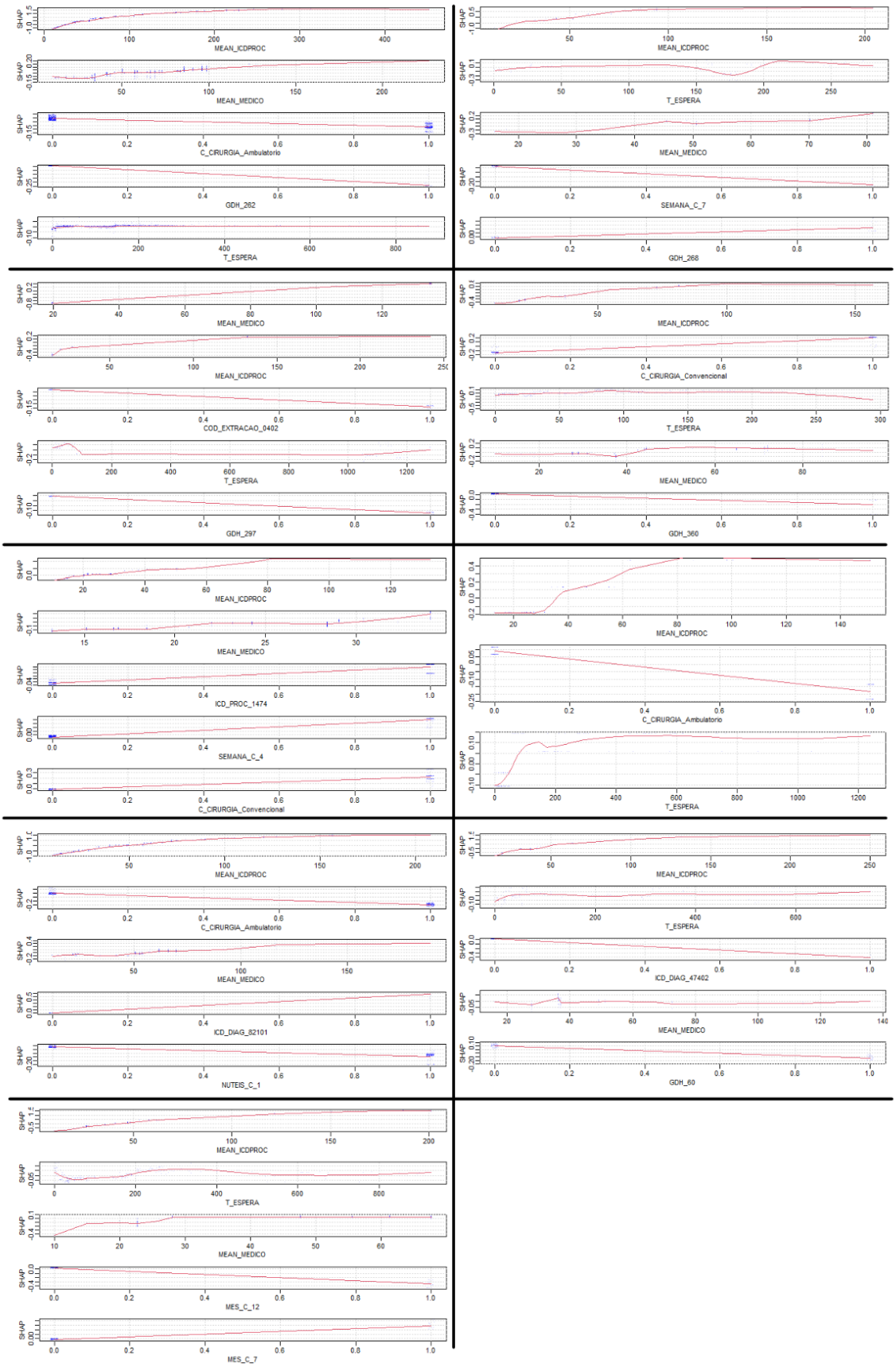


Figura A 4.3 – Importância das Variáveis (da esquerda para a direita): Cirurgia Geral, Cirurgia Plástica, Gastrenterologia, Ginecologia, Oftalmologia, ORL, Ortopedia, Pediatria Cirúrgica e Urologia.

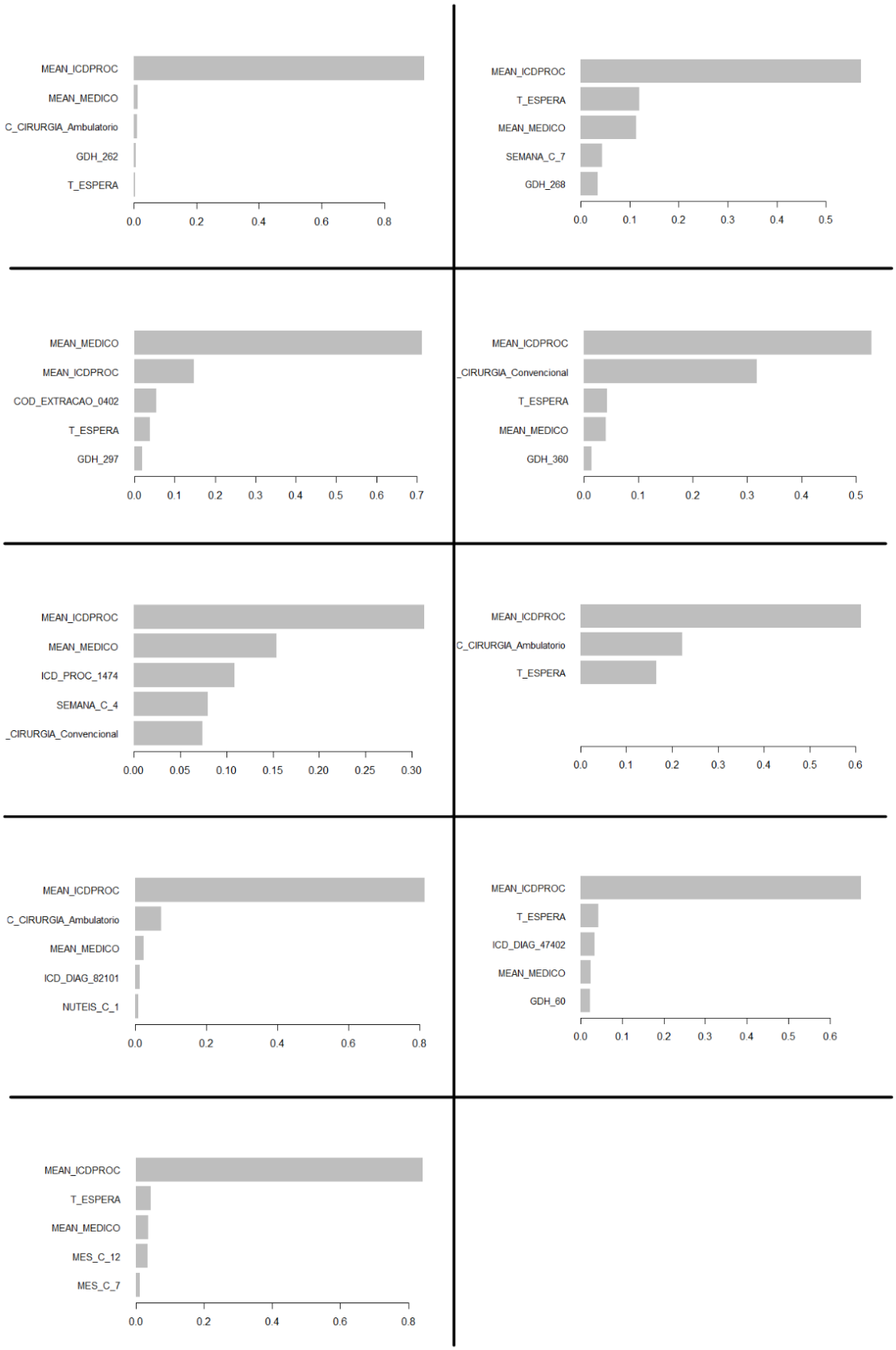


Figura A 4.4 – Valores reais vs. Valores previstos (da esquerda para a direita): Cirurgia Geral, Cirurgia Plástica, Gastrenterologia, Ginecologia, Obstetrícia, Oftalmologia, ORL, Ortopedia, Pediatria Cirúrgica e Urologia.