UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE FÍSICA

Improving Operating Room Schedule in a Portuguese Hospital: A Machine Learning approach to predict Operating Room Time

Alice Sousa Paulo

Mestrado em Engenharia Biomédica e Biofísica

Dissertação orientada por:
Dr. Rui Miguel Neves Cortes
Dr. Nuno Miguel de Pinto Lobo e Matela

2022

Acknowledgments

First, I would like to thank Doctor Rui Miguel Neves Cortes, from Lean Health Portugal, for accepting this challenge and guiding me in my dissertation with professionalism and attention. I also need to mention the support from the Lean Health Portugal team, Raquel Rodrigues, Millena Gama, Mariana Campos, and Carolina Ferraz.

I would also like to express my gratitude to my supervisor, Professor Nuno Matela, from the Faculty of Sciences of the University of Lisbon, for all the constructive comments and support during this year.

This project was only possible thanks to Joana Seringa and Sérgio Pedreiras from Centro Hospitalar Universitário Lisboa Central (CHULC), who made the materials available and gave their support, especially in the initial phase of this project. From CHULC, I would finally like to thank all the hospital administrators and the medical surgical directors from Urology, Luís Campos Pinheiro, General Surgery, Hugo Pinto Marques, and Orthopedics, Nuno Diogo, as well as the anesthesiologists, José Miguel Morais Silva Pinto and Cristina Ramos that had the availability to receive me and contribute to the project with their knowledge.

To all my close friends and family, I must express my appreciation and love for believing in me and giving me the strength and motivation I needed throughout these months. A special thanks to my parents and my sister Sónia, that gave me the opportunity and freedom to achieve my life goals.

Abstract

For most hospitals, the operating room (OR) is a significant source of expenses and income. A critical point of effective OR scheduling is the prediction of OR time for a patient procedure. An inefficient schedule results in two scenarios: underestimated or overestimated OR times. A solution reported in the literature is the implementation of machine learning (ML) models that include additional variables to improve the accuracy of these predictions.

This project goal is to improve the OR schedule efficiency in a hospital center by achieving precise OR time predictions. This goal was accomplished by developing two ML models (Multiple Linear Regression (MLR) and Random Forest (RF)), through two different approaches. Firstly, for all the specialties on the dataset (All Specialties Model). Second, a specialty-specific model for each (Urology, General Surgery, and Orthopedics Models). This leads to eight models where the predictive features were identified based on the literature along with consultations with the professionals.

The All Specialties Model presented a surgery median time of 115.0 minutes, with an R-squared surrounding 0.7. Urology had a median time of 70.0 minutes, with an R-squared of 0.822 and 0.831 and a MAE of 21.7 and 20.9 minutes for MLR and RF models, respectively. General Surgery had a median time of 110.0 minutes with an R-squared of 0.826 and 0.825 and a MAE of 26.2 and 26.1 minutes for MLR and RF, respectively. For Orthopedics, the RF was the only one able to model all the data with an R-squared of 0.683 and a MAE of 27.1 minutes.

When compared with the current methods, considering a 10% threshold, the models achieved reductions in underestimation surgeries (41%), and an increase of within predictions (19%). However, with a 22% increase in overestimation predictions. We conclude that using ML approaches improve the accuracy of OR time predictions.

Keywords: Operating Room Scheduling, Operating Room Efficiency, Operating Room Time, Machine Learning, Prediction

Resumo

O bloco operatório representa uma das unidades que gera maior despesas e receitas a nível hospitalar. Trata-se de um ambiente altamente complexo, onde é necessário alocar recursos materiais e humanos que são extremamente dispendiosos. Desta forma, o bloco operatório necessita de ser gerido de forma eficiente para garantir que o investimento inicialmente feito tem o seu retorno e é utilizado no seu máximo potencial. Paralelamente, os hospitais públicos, integrados no Serviço Nacional de Saúde, apresentam longas listas de espera às quais necessitam de dar resposta. Esta crescente demanda por serviços de saúde, que exige tratamento a nível de bloco operatório, é agravada pelo envelhecimento populacional, e leva a que todos os profissionais envolvidos neste ambiente coloquem os seus esforços no sentido de garantir que toda a população tem as suas necessidades asseguradas. Um ponto fulcral no problema descrito passa por, numa primeira instância, garantir um agendamento cirúrgico eficiente. Quando um paciente é eleito para uma cirurgia programável, cirurgia eletiva, é colocado em lista de espera e feito o seu agendamento, para mais tarde realizar o respetivo procedimento cirúrgico. No momento do agendamento é necessária a informação do tempo de sala de operação que o paciente irá requerer, para reservar o bloco de tempo de sala adequado ao seu procedimento cirúrgico. Um agendamento cirúrgico ineficiente pode gerar dois diferentes cenários que não são desejáveis. Por um lado, se existir uma subestimação do tempo de sala, situação em que o tempo previsto é inferior ao real, leva a que a cirurgia seja mais longa que o estimado e, consequentemente, atrase as operações seguintes. No pior dos cenários há operações que são canceladas. Por outro lado, se há uma sobrestimação, a cirurgia levou menos tempo que o estimado, não há um aproveitamento total dos recursos da sala de operação. Na maioria dos hospitais, esta previsão de tempo de sala é feita com base na experiência do cirurgião e a implementação de ferramentas de inteligência artificial para executar esta tarefa ainda é escassa. Este tipo de previsão leva a um elevado número de cirurgias subestimadas, pois o cirurgião, na sua maioria, não tem em consideração fatores do paciente e anestésicos que impactam o tempo de sala considerando, na maioria das vezes, somente o tempo necessário à cirurgia em si. Além disso, o cirurgião tende a alocar o maior número de cirurgias num curto bloco de tempo, o que leva a uma previsão irrealista.

Uma solução apontada na literatura é a implementação de algoritmos de aprendizagem automática para o desenvolvimento de modelos que implementem variáveis associadas ao paciente, operacionais, anestésicas e relacionadas com o staff. Este tipo de abordagens mostrou melhorar a precisão na previsão do tempo de sala.

O projeto apresentado foi baseado numa metodologia que, primeiramente, permitiu a compreensão dos métodos praticados no centro hospitalar abordado no projeto, o Centro Hospitalar Lisboa Central (CHULC), a validação da relevância do projeto e como objetivo principal, o aumento da eficiência do bloco operatório através da melhoria na precisão da predição do tempo de sala. Toda a metodologia foi desenvolvida tendo como fundamento a base de dados fornecida por esta instituição que contém todas as cirurgias relativas às especialidades de Urologia, Cirurgia Geral e Ortopedia realizadas nos últimos cinco anos (janeiro de 2017 a dezembro de 2021). Para alcançar o objetivo central de melhorar a predição do tempo de sala, foram propostos dois modelos de aprendizagem automática, cujo output é o tempo de sala, um modelo de regressão linear múltipla e de uma floresta aleatória (em inglês designado por Random Forest- RF) segundo duas abordagens. A primeira abordagem consistiu no desenvolvimento de um modelo único para todas as três especialidades apresentadas na base de dados e a segunda num modelo específico para cada especialidade individual. O que conduziu a um total de oito modelos, uma vez que em cada abordagem ambos os algoritmos de regressão linear múltipla e de RF foram implementados. As variáveis com potencial valor preditivo da base de dados do CHULC foram

identificadas com base na revisão de literatura assim como em reuniões marcadas com os diretores de serviço das especialidades abordadas, administradores hospitalares e anestesiologistas.

Uma vez abordada a metodologia atualmente implementada no CHULC para a previsão do tempo de sala, que é baseada na experiência do próprio cirurgião, foi avaliado o impacto do tempo controlado pelo cirurgião e relativo à anestesia no tempo de sala. O tempo controlado pelo cirurgião apresentou a maior correlação com o tempo de sala, com um coeficiente de Pearson de 0,966 seguido do tempo anestésico, com um coeficiente de 0,686. A elevada correlação do tempo controlado pelo cirurgião com o tempo de sala indica que, por um lado, a forma como a predição do tempo de sala é praticada atualmente não é totalmente errada, mas, por outro lado, não é tão realistas já que não considera todos os fatores que influenciam este tempo. Ao incluir as variáveis relativas ao paciente, hospital e anestesia nos oito modelos propostos, para uma mediana de tempo de sala de 115,0 minutos, o modelo de regressão linear relativo a todas as especialidades obteve um R-quadrado de 0,780 acompanhado por um erro médio absoluto de 26,9 minutos. Os modelos de Urologia apresentaram um R-quadrado de 0,822 e 0,831 e um erro médio de 21,7 e 20,9 minutos para o modelo de regressão linear e de RF, respetivamente, com uma mediana de cirurgia de 70,0 minutos. Para a Cirurgia Geral, a mediana de cirurgia é de 110,0 minutos com um R-quadrado de 0,826 e 0,825 e um erro médio de 26,2 e 26,1 minutos para os modelos de regressão linear e RF, respetivamente. No modelo de Ortopedia, o algoritmo de RF foi o único capaz de modelar todos os dados desta especialidade com um R-quadrado de 0,683 e um erro médio de 27,1 minutos, para uma mediana de cirurgia de 130,0 minutos. Nesta especialidade, a regressão linear conseguiu moldar todas as cirurgias com exceção das cirurgias relativas ao joelho e anca, com um R-quadrado de 0,685 e erro médio de 28,9 minutos. As possíveis causas foram levantadas e descritas em maior detalhe, a elevada variabilidade entre procedimentos e o perfil de doentes (polidiagnosticados e polimedicados) foram os pontos fulcrais apontados pelo diretor de cirurgia ortopédica do CHULC.

Quando comparado com os métodos atuais do CHULC, todos os modelos alcançaram uma diminuição significativa no erro de predição do tempo de sala. Considerando uma margem de 10%, todos os modelos apresentaram uma redução na percentagem de cirurgias subestimadas, cerca de 41%, e um aumento nas percentagens das cirurgias estimadas corretamente, rondando os 19%. No entanto, os modelos registaram um aumento de 22% nas cirurgias sobrestimadas. Futuros estudos no sentido de traduzir o impacto de cirurgias subestimadas e sobrestimadas serão necessários para complementar estes resultados.

A variável que apresentou um maior impacto em todos os modelos de RF foi a média do cirurgião com base no tipo de procedimento cirúrgico realizado. Dado o elevado grau de linearidade desta variável com o output do modelo, o tempo de sala, expresso por um coeficiente de Pearson de 0,865, levou a que o modelo de regressão linear conseguisse traduzir de forma precisa a relação entre estas variáveis, e, consequentemente, atingisse resultados semelhantes ao modelo de RF nas especialidades de Urologia e Cirurgia Geral.

Conclui-se que a implementação de abordagens de aprendizagem automática melhora a precisão na predição do tempo de sala e podem servir como uma ferramenta de apoio à decisão clínica para o auxílio do agendamento cirúrgico. Para operacionalizar estes resultados a nível hospitalar é necessário trabalho futuro.

Palavras-Chave: Agendamento Cirúrgico, Eficiência de Bloco Operatório, Tempo de Sala de Operação, Aprendizagem Automática, Predição

List of Figures

List of Tables

List of Acronyms

ACT	Anesthesia Controlled Time
AI	Artificial Intelligence
ANOVA	Analysis of Variance
ASA	American Society of Anesthesiologists
BMI	Body Mass Index
CHULC	Centro Hospitalar Universitário Lisboa Central
CV	Cross Validation
DT	Decision Tree
EHR	Electronic Health Record
ESS	Explained Sum of Squares
GBM	Gradient Tree Boosting Model
GEMs	General Equivalence Mappings
ICD	International Code of Disease
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
ML	Machine Learning
MLR	Multiple Linear Regression
MSE	Mean Squared Error
NCEPOD	National Confidential Enquiry into Patient Outcome and Death
NHS	National Health Service
OHE	One Hot Encoding
OLS	Ordinary Least Squares
OR	Operating Room
PACU	Post Anesthesia Care Unit

RAS	Robotic Assisted Surgery
RF	Random Forest
RMSE	Root Mean Squared Error
RSS	Residual Sum of Squares
SCT	Surgeon Controlled Time
SIGIC	Sistema Integrado de Gestão de Inscritos para Cirurgia
TSS	Total Sum of Squares
VIF	Variance Inflation Factor
XGB	Extreme Gradient Boosting

Contents

1. Introduction

1.1 Problem Context and Motivation

Over the past few years, the increase in digital health, storage capacity, and information processing have been improving healthcare services. The implementation of Electronic Health Records (EHR) in several hospitals illustrates this tendency. EHR is a digitized version of patient data that incorporates personal, medical, and procedure information. It provides efficient access to an extensive amount of information, reduces research expenses, accelerates new medical research, and can also be helpful in preventing medical mistakes. Adopting high-quality EHR has a significant influence on enhancing hospital healthcare quality and management [1].

Digital technologies and Artificial Intelligence (AI) are remodeling medicine, medical investigation, and public health. The implementation of AI technologies in the health sector has already provided essential contributions to some areas such as prediction-based diagnosis, health systems management and planning, and public health surveillance. With a pronounced implementation of these technologies, particularly Machine Learning (ML) algorithms, we witness a tendency to the usage of AI to aid healthcare providers and clinicians to avoid errors and allow these professionals to focus on more critical tasks and complex cases. The potential advantages and the economic benefits of AI for healthcare presage an expanded implementation of AI worldwide [2].

With an increased aging population, governments, the largest suppliers, and healthcare supporters, particularly outside the United States, became overloaded. Several numbers of countries put their efforts into reducing healthcare costs. Healthcare expenditures are elevated, and the Operating Rooms (ORs) represent the highest revenue costs for the hospitals [3]. This can be explained by the fact that numerous and expensive resources are needed, such as specialized and professional staff, cutting-edge technology, advanced equipment, and other medical supplies [4], [5]. Additionally, these resources need to be strategically distributed among hospital units and departments [6].

A significant number of patients admitted to the hospital are treated in the OR [4]. In ORs, problems related to the long surgery waiting lists and their waiting periods have become a growing concern for governments, which strive to create better conditions for the population. As a matter of fact, in 2012, Portugal, accompanied by the United Kingdom, was ranked at the bottom of the list in terms of healthcare accessibility amongst 34 other European countries. This result is heavily influenced by the extensive waiting lists for surgical procedures [7]. Despite its low ranking among European peers, the Portuguese healthcare system is a notable illustration of an improvement effort. Considering that all the population of Portugal has the right to access the National Health Service (NHS), there is an overflow of demand, resulting in higher wait times and waiting lists. Waiting lists and waiting periods have been a significant health policy subject in Portugal for many years, leading to the development of the Integrated Management System for the Surgery Waiting List, in Portuguese “Sistema Integrado de Gestão de Inscritos para Cirurgia” (SIGIC), in 2004 [8]. The primary objectives of SIGIC were to minimize

wait times, assure equity of access, increase overall system efficiency, and offer information quality and transparency [9]. The waiting times decreased in subsequent years after the SIGIC creation, although this trend has recently reversed. In recent years, the average waiting times have increased. According to the NHS, in January 2021, from 256 000 citizens registered for surgery, about 38.6% exceeded the maximum response time grenade [10]. It is important not to underestimate the negative impact of these delays on the economy, which justifies the investments to reduce the waiting lists, and surgery response. Although the government has made an effort to hire more health professionals in the last decade, more than 12% in 2019, it is not enough to respond to the growing need for healthcare services [11]. The Covid-19 pandemic has aggravated this problem since numerous health professionals leave the NHS for private institutions or other areas outside health [12].

Hence, part of the solution is not only to allocate more human resources but also the development of more technologies, tools, and models that will help the professionals and optimize the existing processes. AI provides a fundamental capacity for a more efficient data collection and process that complements health professional work, by reducing the data evaluation error. Additionally, it supports a massive knowledge evolution in health, and as long as the amount of collected and processed data increases as well as the data quality, the more accurate the predictions will be [13]. With this in mind, improving OR scheduling by implementing ML models can provide considerable advantages for health units like contributing to Value Health and, at the same time, maximizing the level of patient and staff satisfaction.

1.2 Problem Definition

The ORs are among the highest expenditure departments in hospitals. This is a complex and frequently unexpected environment, with various variables contributing to its inefficiency. Accordingly, surgeons and hospital administrators both present the responsibility to ensure that ORs are used to their maximum potential and that operating time in the theater is used wisely so that the return from the investment in the OR is maximized. Thus, in OR scheduling, patient contentment, and resource efficiency must be prioritized [3]. A solid planning and scheduling system in OR will allow more surgical activity, including emergencies, to be completed in a fair amount of time, improve the patient and caregiver experience, and boost personnel satisfaction and morale. For these reasons, many hospitals are increasingly scheduling OR's using scientific methodologies [14].

Several types of surgery and procedures performed in ORs can be categorized based on surgical urgency. According to the National Confidential Enquiry into Patient Outcome and Death (NCEPOD) [15], the types of intervention can be categorized into emergent, urgent, expedited, and elective. Elective surgeries encompass all surgeries scheduled in advance of regular admission to the hospital since they do not represent a medical emergency. On the other hand, expedited surgeries apply to patients who require an early treatment, where the condition is not an immediate threat to life, limb, or organ survival, generally performed within days of the decision to operate. If the intervention needs to be completed for acute onset of potentially life-threatening conditions within hours of decision to operate is classified as urgent. Lastly, emergent interventions are relative to immediate life, limb, or organ saving intervention within minutes of decision to operate [16].

The standard process for elective surgery consists of four main steps: the patient diagnosis by the physician and surgical decision, adding the patient to the waiting list and schedule, performing the surgery, and the postoperative recovery, as schematized in **Figure 1.1**. The OR time prevision will be recorded in the patient's EHR once the physician has performed the surgical checkup. The main surgical schedule is then planned on the system by the planning assistant [17].

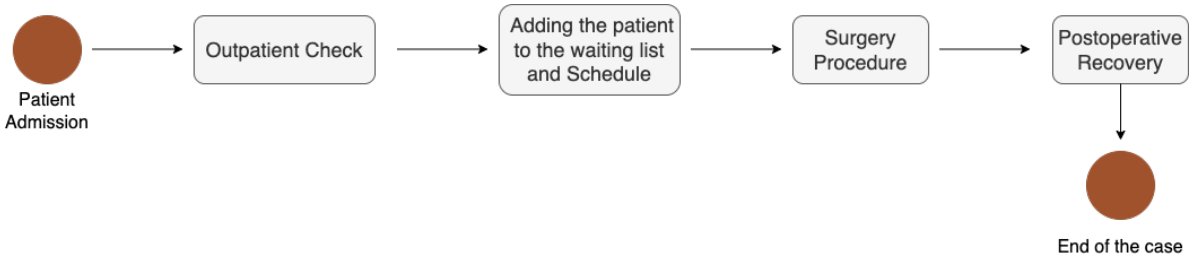

Figure 1.1- Standard flow of patient scheduling when admitted to an elective surgery.

When a patient arrives at the OR on the scheduled day for elective surgery, the ward nurse or the outpatient surgical department will first verify the patient's appointment time and basic information. Then, the anesthetist will interact with the patient in the holding area, assess the patient's operability and desire to undergo anesthesia, and select the anesthetic strategy. If the patient has already been hospitalized, the ward nurse and anesthesia assistant examine the patient before the operation. Following that, numerous activities might be completed simultaneously. The circulating nurse will evaluate the patient's vital signs, allergies, surgical risks, and special equipment before transferring the patient from the holding area to the operating theater. The anesthesia assistant will prepare the necessary supplies based on the anesthetic strategy, and the surgical nurse will prepare the surgical materials, position instruments, and adjust equipment, among other duties. The circulating nurse will move the patient to the OR. Once the OR has been cleaned and is ready from the last surgery, the surgical nurse will complete the final preparation.

Following the period of surgical operation, two sub-processes might be carried out simultaneously. On the one hand, the anesthetist returned to the OR with the anesthesia assistant to end the anesthesia, observe the patient's vital signs, wait for the patient to wake up, and transfer the patient to the recovery area with the circulating nurse. On the other hand, the surgical nurse sorted out the knives and other supplies, closed the equipment, and cleaned the OR. Once the medical waste is at the specified location outside the OR, the cleaning team will arrive to remove it, dump it off, or clean and disinfect it. When the surgery is completed, most patients are sent to a Post Anesthesia Care Unit (PACU) to recuperate from anesthesia. Generally, patients must be observed in the postoperative surgery recovery area for two hours. If the patient's condition is stable, the patient will be moved to the general ward. At this moment, the surgical process is completed.

This is a highly complex process, as schematized in **Figure 1.2** with a high degree of variability and coordination that increases the uncertainty and unpredictability of surgery and OR time. Therefore, it is crucial in scheduling surgeries, human resource planning, and other logistical and planning procedures. The duration of the surgery and the total OR time are some aspects that must be considered while scheduling an operation in the OR [18]. However, it is well recognized that surgical durations have a significant degree of intrinsic variability, making it difficult to get the point estimates with minimal standard errors [19]. This can be explained by the fact that the duration of surgery is determined by

various factors, including different medical specialties, the surgical team's expertise, and the patient's health status. Patient factors might induce unexpected adjustments, such as violations of preoperative fasting periods or medical clearances that require last-minute cancellations, or intraoperative findings that affect the course of an operation. There are also staffing factors, such as start delays due to overlapping processes or poor communication, that induce variability in operating time. Finally, there are also system factors, including equipment supply chains and sterilization processes, patient transportation, and the efficient allocation and utilization of OR block time for each patient, operation, and surgeon [18].

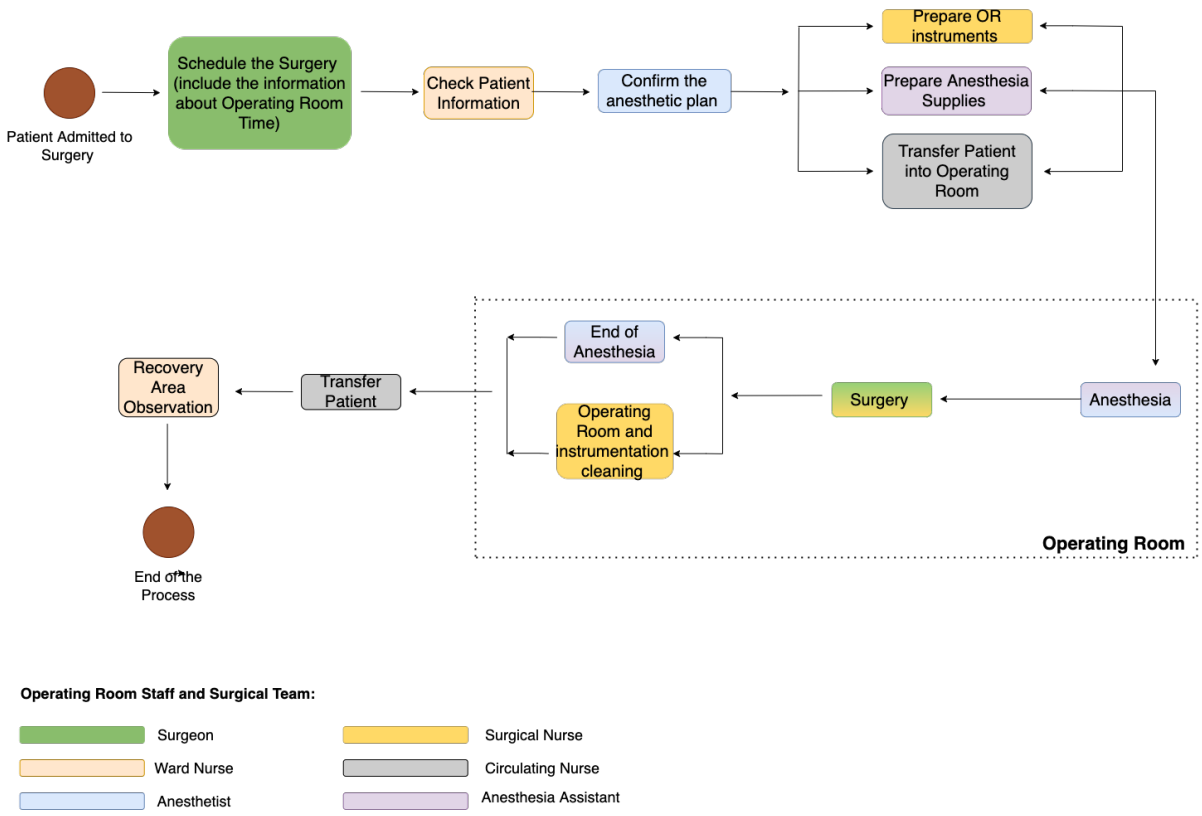

Figure 1.2- Patient's path during the surgery process mapping. The rectangles represent the different sequence tasks, and the rectangle color the Operating Room staff and surgical team personnel responsible for the respective task. The dots surround all the tasks that take place in the Operating Room.

The OR time is defined as the time elapsed between the patient's arrival into the OR and the patient's exit from the OR. This period includes the room setup, patient positioning, and the last recapitulation and confirmation of the patient identification, surgical location, and planned procedure, the preparation time. The preparation time is followed by two central moments: the Surgeon Controlled Time (SCT) and the Anesthesia Controlled Time (ACT). The ACT is divided into two moments: anesthesia induction time and anesthesia emergence time. The anesthesia induction time starts with the injection of the anesthetic drug until the induction conclusion, after all the catheters have been inserted and the patient is prepared to be positioned by the surgeon. The anesthesia emergence time is relative to the progressive return of consciousness following the cessation of anesthetic agents' delivery at the end of the surgical operation. The final time corresponds to the patient's transference to the OR holding and other small processes needed to perform by the staff before the patient leaves OR, depending on the

surgery type, patient condition, and other factors. The time elapsed after the patient left the room and before the next case is defined as the turnover time. During this period, staff performs various critical activities to clean and prepare the area for the following procedure. The schema of the time's definition is presented in **Figure 1.3** [20].

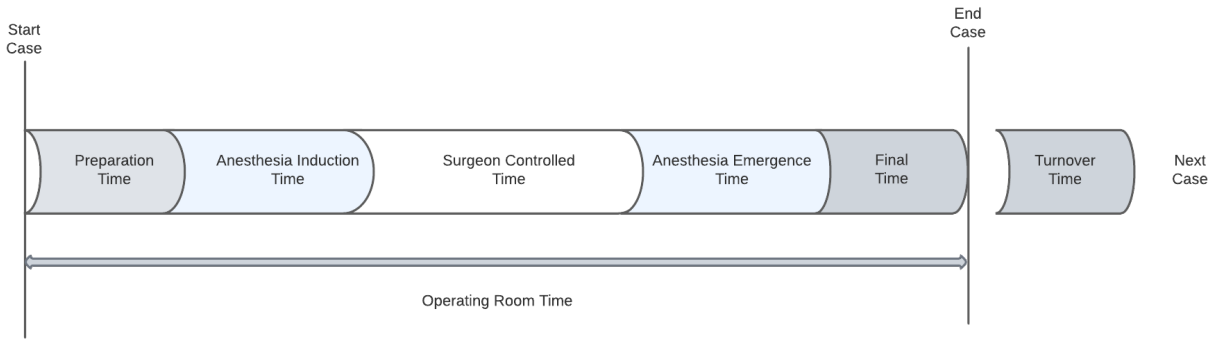

Figure 1.3- Operating Room Time division: each block represents the specific time that composed the total Operating Room Time.

When an operation takes longer than scheduled (underestimation) or less than scheduled (overestimation), it leads to inefficient utilization of OR resources and affects staff satisfaction. Underestimation of operation duration causes subsequent cases to be delayed or canceled, incurring additional unexpected costs of overtime work. Overestimation of surgical duration reduces OR utilization and throughput, and the hospital's resources are not utilized to their maximum potential. Both effects are undesirable, but they have different consequences. On the one hand, time is wasted, resulting in the underutilization of the operating theater, overestimation. On the other hand, procedures may be postponed, schedules may be interrupted, and cancellations can occur, underestimation [21].

The key point is to make more accurate predictions regarding the OR time for individual patients. Thereby, planning would be improved, and potential benefits noted, such as a more accurate prediction for a particular patient when compared to the average prediction for the group of patients undergoing the same operation and the variation around the prediction being smaller than the variation for the group as a whole.

Most hospitals have established estimates for surgery time using basic techniques. The preponderance of the hospital's estimations is based on the surgeon's expertise and, most of the time, does not consider specific factors such as patient conditions and anesthesia concerns. Additionally, case duration is frequently underestimated since surgeon estimations are typically formed by optimizing block scheduling to accommodate the maximum number of surgeries, potential cancellations, and cost savings. Furthermore, surgeries with increased uncertainty and unexpected outcomes throughout the surgery, anesthesia, and system variables that the surgeon may not have addressed add limitations to OR time prediction [19].

For OR time prediction and evaluation of the importance of input features, researchers have used linear statistical models such as regression or simulation to increase predictability. However, a common drawback of this research is that they employed fewer input variables or features in their models than other approaches, due to statistical techniques' limitations in handling with many input variables. ML has recently been proven to overcome this issue and demonstrated to be strong and useful in assisting healthcare management. Concretely, regression models have been used to forecast surgical duration and evaluate the impact of a large number of input variables [6].

The hospital of this case study is Centro Hospitalar Universitário Lisboa Central (CHULC). CHULC comprises 6 hospitals (Hospital de S. José, Hospital de S. António dos Capuchos, Hospital de Sta.Marta, Hospital D. Estefânia, Hospital Curry Cabral, and Maternidade Alfredo da Costa), with 1322 beds (Hospital de São José - 354, Hospital de Santo António dos Capuchos - 212, Hospital de Santa Marta - 200, Hospital Dona Estefânia - 128, Hospital Curry Cabral - 316 e Maternidade Dr. Alfredo da Costa - 112), and a total of 8412 professionals, including interns.

The CHULC works on health care practice training and research project creation with universities and institutes of higher education in health, such as medicine and nursing courses. Hence it is common the intern's presence during surgical procedures. This hospital was the first hospital unit in the Portuguese NHS with a surgical robot to perform assisted robotic surgeries, the Da Vinci Xi robotic system. Surgical robots are implemented in several specialties to replace the surgeon's hand, allowing higher precision and less invasive procedures. In CHULC, assisted robotic surgeries are not yet practiced in all specialties, with General Surgery, Urology, Gynecology and Cardiothoracic being an example of surgical fields where Da Vinci Xi is utilized. In 2021 the CHULC surgical elective activity recorded an increase of 14.7%, corresponding to more 2312 surgeries when compared to the same period of 2020. Moreover, ambulatory surgeries, i.e., surgeries where the patient does not need to stay overnight in the hospital, recorded a decrease of 0.4% [22].

In Portuguese public hospitals, demanding commitment regarding performance improvement and rigor in the management of NHS hospitals is required. Therefore, it is necessary to adopt measures to rationalize expenses, reduce waste, promote quality, improve efficiency in the organization of providers and the resources used in the provision of healthcare, and demand quality control. A determining factor for good management is the knowledge of the hourly cost of the OR and the cost per standard surgery. In Portuguese NHS hospitals, these values can fluctuate between 7€ to 11€/minute, the equivalent of 420€ to 660€/hour. In CHULC, the price per hour of an OR is 591€, where about 58% (345€) represents fixed costs and 42% (246€) variable costs, and the price for a standard surgery surrounded 1296€ [22]. These numbers highlight the fact that OR costs are elevated, and surgical activity is an essential element in the financing of hospital organizations, which is very dependent on the dynamics of the OR. Hence, any minimal improvement regarding ORs utilization represents cost savings for the hospital administration. The directors of surgical specialties approached in this project from CHULC also validated the exposed surgery scheduling problem, especially the OR time prediction task. They considered that the services could benefit from implementing ML models to improve the current scheduling methods.

1.3 Objectives

This thesis aims to develop two predictive models, Multiple Linear Regression (MLR) and Random Forest (RF) to predict the total OR time for a single patient in Urology, General Surgery and Orthopedics specialties. These specialties were selected since they are core specialties in CHULC, and their surgical directors declared to be open to embracing this challenge and demonstrating their interest. Additionally, these specialties represent high volume data, which is relevant for the problem. By achieving more accurate and realistic predictions of the OR time, it is possible to improve the surgery schedule for the services.

To reach this objective, this thesis focuses on the following lines of reasoning: 1) Understand the current methods and challenges of surgery schedule and OR time prediction on CHULC. 2) An examination of the literature regarding the application of ML models in surgery scheduling to understand the current approaches and how to improve them. 3) Evaluate the impact of both SCT and ACT times in OR time. 4) An exhaustive analysis of the dataset features, data distribution, and context. 5) Data Cleaning and Feature engineering. 6) Implement the predictive algorithms for all specialties and in a single specialty approach. 7) Elaborate specific models for each specialty and analyze the feature importance and significance of each one. 8) Evaluate the models and compare with the current methods.

1.4 Outline

This thesis is organized as follows: Chapter 2 presents the theoretical background with the concepts of ML to solve prediction tasks, model conceptualization, and evaluation metrics. Chapter 3 is relative to the literature review. First, the OR efficiency was approached, followed by the implementation of ML models to predict OR times, with a summary table for these studies that contain the relevant highlights, such as the input features, type of ML models, and evaluation metrics, amongst other criteria. Chapter 4, the Materials and Methods section, includes the CHULC dataset description and the framework to build the ML models, including the data preparation phase, the model conceptualization, evaluation, and, finally, the comparison with the current methods. The results are provided in chapter 5 and discussed in chapter 6. In chapter 7, the conclusions are pointed out, and the directions for future work are reported.

2. Theoretical Background

This chapter presents the methods and theories on which this dissertation is based. Firstly, section 2.1 introduces the workflow of ML problem resolution. In section 2.2, the concept of linear and non-linear ML approaches is introduced. Finally, the last sections report the model evaluation, hyperparameter tuning, and feature importance.

The employment of big data and current data science tools, such as ML, has received increased attention for their capacity to predict perioperative occurrences and support clinical decisions [23]. In particular, estimating the OR time is appropriate for a ML approach since the datasets are extensive and can possibly capture the multiple elements that might impact the OR time.

ML is a branch of AI and can be defined as the process of developing models that can learn and improve on their own by being adequately designed. These systems must seek patterns in collected data and employ them to make future predictions [24]. Data mining is also a recognized and popular AI area, and has a lot in common with ML. The prime difference between data mining and ML is that data mining focuses on extracting the rules from a massive amount of data, whereas ML is the process of teaching the computer how to learn and understand the specified parameters to seek these rules.

An important consideration in ML is the distinction between algorithm and model. The algorithm runs on the dataset and learns the patterns to build the model, a structure with the coefficients that are used to make the predictions on the data. The ML ground has been subdivided into multiple subfields that handle different types of learning tasks. The two major subfields are supervised and unsupervised learning. The primary distinction between the two approaches is that supervised learning is performed using the ground truth, i.e., the prior knowledge of the correct output values for the samples, labels. Hence, supervised learning aims to build a function that, given a sample of data and the desired outputs, best approximates the data-observable connection between the input and the output, the mapping function. Contrarily, unsupervised learning lacks labeled outputs. Therefore, its purpose is to infer the inherent structure existing in data points. Supervised learning can be divided into two types of problems: classification if the output is a discrete class variable and regression if the output is a continuous variable [24]. This dissertation will exclusively focus on supervised learning, specifically in regression problems, since this is the type of problem addressed in this project.

Supervised learning is used to identify data patterns, which can then be applied to an analytic process. The algorithm can predict the corresponding output variable when new input data is added. The fundamental goal of the training phase is to learn from labeled examples, included in the training set, to then identify unlabeled cases with high potential accuracy during the test phase. The workflow of this process is schematized in **Figure 2.1** and will be detailed and explained in the following sections [25].

ML can address many problems across different industry sectors by working with the correct datasets. Hence the first step for the ML problem resolution is data collection. The model's accuracy is determined by the quality of the data provided to the machine. If the data is erroneous or does not contain the information necessary to respond to the research problem, we will get incorrect results or irrelevant forecasts. Therefore, it is crucial to use data from a credible source since this will directly impact the output of the model [26]. The following steps of the ML problem resolution can be divided into: data preparation, model implementation and evaluation, parameter tuning, and finally, the deployment phase.

The deployment phase can be defined as taking a trained ML model and making its predictions in new data [27]. Data preparation, model development, evaluation, and parameter tuning phases will be described in greater detail in the following subsections.

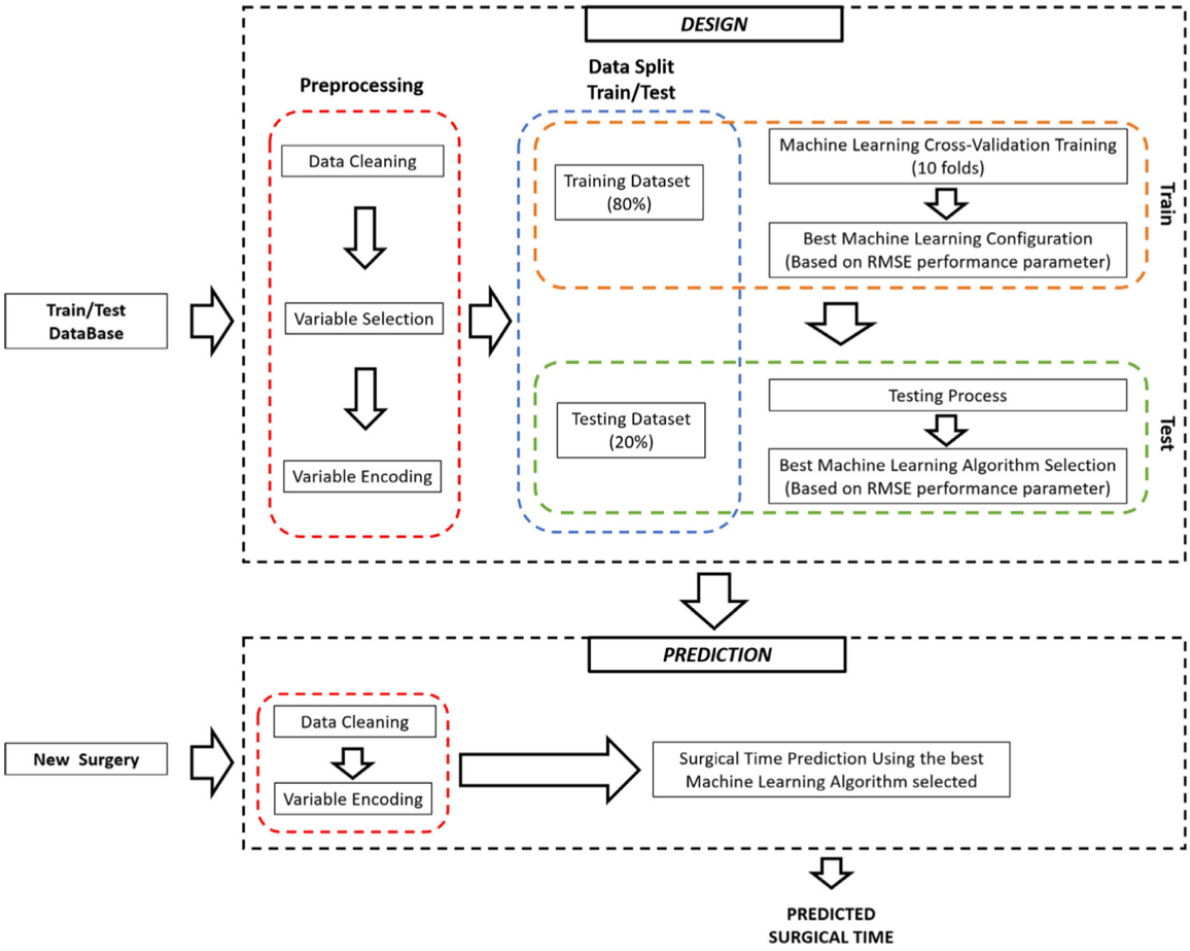

Figure 2.1- Design of a supervised Machine Learning model for a regression task, predict the Operating Room time for a surgery [25].

2.1 Data Preparation

Data preparation is the process of converting raw data, used for ML algorithms, to find insights or make predictions, in **Figure 2.1** is labeled as “Preprocessing”, inside the red dots. The process addresses missing data, improperly formatted/structured data, inconsistent values, and non-standardized categorical variables [28].

During the data preparation phase of a ML project, standard tasks may be employed. These tasks include [28]:

-Data cleaning: Process of identifying and rectifying inaccuracies or errors in data;

-Feature Engineering: Process of selecting, manipulating, and converting raw data into features that can be employed in supervised learning. This includes data transformation (the process of converting the scale or distribution of variables), creating new variables from existing data and feature selection (the process of identifying the input variables that are most relevant for the problem).

2.1.1 Data Cleaning

Data cleaning involves the transformation of messy and complex data into clean data. Messy data includes statistical noise, missing values, and redundant examples. Complex data refers to raw data that may include complex non-linear connections that must be uncovered [29].

This process is conducted by identifying outliers, missing values, duplicates, and redundant samples. Missing values are common in real-world data due to various reasons, such as unrecorded observations or information leakage. The missing data diminishes the statistical power of the analysis and distorts the conclusions' validity. Many ML algorithms do not accept data with missing values, therefore, handling missing data is critical.

After identifying all missing values, commonly represented by NaN, the most basic technique for dealing with this type of data is deleting entries with a missing value. However, this strategy is not recommended since it can result in the elimination of some essential and valuable information data from the original dataset. For that reason, imputing the missing values is a plausible strategy for handling this problem [30]. There are several methods for replacing missing values. Among them, it is possible to highlight [30]:

-Replacing with an arbitrary value: The process of replacing all instances of missing values within a variable with an arbitrary value, such as "0", "Missing", or "Not defined". However, this process can distort the variables' original distribution, and the arbitrary values can create outliers;

-Replacing with mean/mode/median: This is the most often used approach for filling in missing values in numeric columns. If there are outliers, the mean, the measure of central tendency, will not be adequate. The median, the middle value when data points are organized in order, will be more suitable for imputation since the mean is statistically sensitive to outliers. The mode is the most frequent value of the data points. However, the mode might lead to ambiguity when dealing with continuous data. If none of the values are repeated, there may be more than one mode or, rarer, none at all. The mode can also be utilized to fill in missing values for categorical data columns;

-Replacing with the previous value (Forward fill) or with the next value (Backward fill): In some circumstances, especially for time series data, imputing the values with the previous value rather than the mean, mode, or median is preferable. This is referred to as forward fill. When the next value is used to impute the missing value, it is called backward fill;

-Interpolation: Interpolation is a technique for estimating unknown data points between two known data points, more common in time series and image processing;

-Most probable value: This can be calculated using regression, k-Nearest Neighbor imputation techniques, or decision tree induction.

A dataset may contain extreme values outside the range and differ from the rest of the data, the outliers. Understanding and even deleting these outlier values may enhance ML modeling and model quality. There are different methods to detect outliers, and they can be based on the distance and density of the other data points in the dataset or on a threshold previously defined by the user. The threshold is selected based on the user's domain knowledge regarding the dataset and the defined problem [31].

Duplicates refer to rows with similar data that may be ineffective in modeling if they are not critically identified during the model review. ML algorithms will improve the performance by detecting and eliminating duplicate data entries. Duplicate rows will lead to a misleading performance in an algorithm evaluation. For example, in a train/test split or k-fold cross-validation, duplicate rows may exist in both the train and test sets, and the model evaluation on these rows will be accurate. As a result, a biased performance estimation on new data will be provided [31].

2.1.2 Feature Engineering

i. Data Transformation

Data transformation refers to the process of transforming raw data into a format or structure that is more appropriate for model building. This process is determined by the data's characteristics, such as variable types, and the algorithms that will be used to model it, which may impose specific requirements on the data. This process typically leads to a change in the type or distribution of data variables [32].

Most real-world datasets contain both categorical and continuous variables. While continuous variables easily fit into all ML models, categorical variables are implemented distinctly in models and programming languages. There are some approaches to handle this problem such as discretization and encoding techniques. Discretization is the process that transforms a numeric variable into an ordinal variable. Alternatively, it is also possible to encode a category variable as an integer or a boolean variable, which is necessary for most regression and classification problems [32]. The following points describe some of these common transformations [32]:

-Discretization Transformation: Encodes a numeric variable as an ordinal variable;

-Ordinal Transformation: Encodes a categorical variable as an integer variable;

-One Hot Encoding (OHE): Encodes a categorical variable as a binary variable. OHE transforms each category value into a new categorical column and attributes it a binary value of 1 or 0. A binary vector is used to represent each integer value. The values are zero, and the index is denoted by a 1. This allows categorical coding variables with independence between them, which is beneficial for avoiding correlation biases that the algorithm may learn. However, it raises the dimensionality of the model in the same

proportion as the number of categories to be encoded. This approach may introduce dimensional difficulties when the number of categories encoded is large.

Numeric variables are processed by computers. There is a significantly higher resolution in the range, from 0 to 1 than in the data type's larger range. As a result, it may be advantageous to normalize variables to this range. If the data has a Gaussian probability distribution, shifting it to a standard Gaussian with a mean of zero and a standard deviation of one may be more beneficial [32];

-Normalization Transformation: Scales a variable into the range 0-1;

-Standardization Transformation: Scales a variable to a standard Gaussian distribution.

ii. Feature Creation

Feature creation entails determining which variables will be most relevant in the model prediction. This is a subjective process that needs human interaction and prior knowledge about the data as well as the relationship between the features. The existing features are combined using math operations to generate new derived features with higher predictive values or based on domain knowledge about the data [33]. As a note, the terms feature and independent variable meaning the same in this dissertation. In ML the term feature is more common to use whereas independent variable in statistics.

iii. Feature Selection

When creating a predictive model, feature selection is the process of minimizing the number of input variables. It is preferable to limit the number of input variables to reduce modeling computational costs and, in certain situations, increase the model performance. Some techniques, such as filter, wrapper, and embedding algorithms, are used for feature selection [33]. For linear regression models, it is a common practice to examine the relationship between the independent and dependent variables, based on the p-value [34].

Another essential point in feature selection is the collinearity analysis between features, this is particularly important for linear models. In statistics, collinearity or multicollinearity occurs when there is a correlation between two or more predictor variables. Correlation describes how two or more variables are linked and the strength of this relation. In a MLR model, one predictor variable that presents a strong correlation with other variables can be linearly predicted from the others with a high degree of accuracy. Since the critical point of a regression model is to separate the relationship between each independent variable and the dependent variable, if the correlation between variables is high, it might bring some issues when fitting the model and in the result's interpretation. It weakens the capability of the statistical measure to rely on p-values to detect significant independent variables, because it diminishes the power of coefficients. As a result, it is challenging to analyze independent factors and individual effects on the dependent variable. Multicollinearity can also lead to overfitting when the model performs well on train data but unsatisfactorily on test data [35]. The two most common approaches to detect multicollinearity are the correlation matrix and the Variance Inflation Factor (VIF).

A correlation matrix is a table that displays the correlation coefficient between variables. A single variable (X_i) in the table is associated with the table's other values (X_j). This allows us to examine which pairs have the highest correlation. The most common correlation coefficient is the Pearson's

correlation coefficient, but there are others. The Pearson's coefficient is a measure of linear correlation that describes the strength and the direction of the relationship between two variables. This metric ranges from -1 to +1, where +1 indicates a positive and strong relationship, -1 indicates a negative relationship, and 0 means there is no relationship. An absolute Pearson's coefficient value lower than 0.40 indicates a low correlation, between 0.40 and 0.59 a moderate correlation, between 0.60 and 0.79 a high correlation, and above 0.80 to 1 a very high correlation.

The VIF is a metric for determining the multicollinearity of a group of multivariate regression variables. In a regression model, the VIF of a variable is equal to the ratio of the total model variance to the variance of a model that contains the specific independent variable under analysis, according to **equation 2.1**, where R_i^2 represents the coefficient of determination of the i^{th} variable on the remaining ones.

$$VIF_i = \frac{1}{1-R_i^2} \tag{2.1}$$

The variance of other independent variables cannot be predicted from the i^{th} independent variable if R_i^2 is equal to 0. Consequently, the VIF is equal to 1, and the i^{th} independent variable is unrelated to the others, implying that multicollinearity is absent in this regression model. A high VIF ratio suggests that the independent variable linked with it is significantly collinear with the other variables in the model.

When collinearity/multicollinearity exists, a possible solution to handle it is to eliminate one of the highly correlated independent variables or combine and add them together [35].

2.2 Model Selection

Model selection is the process of selecting the model that best generalizes data. Based on the literature review in chapter 3, Linear Regression and RF were common regression models used for solving similar problems to the one approached in this project. The process of model selection in this dissertation is detailed in chapter 4. The following sections focus on the theoretical principles behind these two common ML models, Linear Regression and RF.

In the process of selecting a ML model, an important point is the model's ability to generalize new data when learning the target function from training data and find a balanced point between over- and under-fitting. Overfitting occurs when a model predicts the training data too well and learns the information as well as the noise in the training data. This phenomenon severely impairs the model's performance on new data and implies that the model picks up the noise or random oscillations in the training data and learns them as ideas or rules. The issue is that this information does not apply to new data and has a negative impact on the models' capacity to generalize. Underfitting is when a model is unable to learn the training data or generalize it to new data. An underfit ML model is unsuitable and will be noticeable due to poor performance on training data [36].

2.2.1 Linear Regression Models

Linear Regression is a supervised ML model that captures the linear relation between the dependent and independent variables. Simple Linear Regression is the case where there is only a single independent variable, described by **equation 2.2**, where β_0 and β_1 are the coefficients, y the target variable, x the predictor variable and ε the error term [37].

$$y = \beta_0 + \beta_1 x + \varepsilon \tag{2.2}$$

Multiple Linear Regression (MLR) is used when there are multiple explanatory variables. The model for MLR is described by **equation 2.3** and the corresponding prediction by **equation 2.4**.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + \varepsilon \tag{2.3}$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_m x_m + \varepsilon \tag{2.4}$$

In a linear regression model, the Ordinary Least Squares (OLS) is used to estimate the parameters of a linear function. The principle of least squares bases this process. The objective of OLS is to minimize the sum of the squares of the differences between the observed dependent variable, y, and the values predicted by the linear regression model of the independent variable $(x_1, \ldots, x_m)$. This can be traduced in the **equation 2.5** [34]:

$$\hat{\beta} = (X^T X)^{-1} X^T y, \quad \text{where } \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_m \end{bmatrix}, \; X = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1m} \\ 1 & x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n1} & \cdots & x_{nm} \end{bmatrix}, Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \tag{2.5}$$

Although it is unlikely to have an exact solution for this system, the objective is to identify the coefficients that better fit the equations by solving a quadratic minimization problem. Linear Regression constructs and fits a linear model with coefficients (β) to minimize the Residual Sum of Squares (RSS) between the observed and predicted values in the dataset. A residual is a deviation from the estimated value to the regression line that best fits the data. In statistics, the RSS is used to analyze if a statistical model is a good fit for the data. The RSS measures the variance that can not be explained by the model, the variability of the error term. The smaller the RSS, the better the model fits the data. Mathematically, RSS is defined by **equation 2.6**, where y_i represents the i^{th} value of the variable to be predicted, $\widehat{y_i}$ the respective predicted value and n the upper limit of summation.

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{2.6}$$

The Total Sum of Squares (TSS), **equation 2.7**, translates the total variation present in the dependent variable. TSS is decomposed in the variability that can be explained by the regression model, Explained Sum of Squares (ESS), and by the error term, the RSS. TSS can also be defined as the sum of the squared difference between the observation of the i^{th} variable, y_i, and the overall mean, $\bar{y}$.

$$TSS = ESS + RSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{2.7}$$

The greater the model fits, the closer the predicted value is to the actual value. As a result, the coefficient of fit from linear regression, R-squared , R^2, is closer to 1. This coefficient can be defined in terms of RSS and TSS, by **equation 2.8**.

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2} \tag{2.8}$$

R-squared is a statistical metric that measures the proportion of variation in the dependent variable that the independent variables can explain. This metric provides the regression model's fit quality, and its value typically ranges from 0 to 1. The optimal R-squared value is 1. The closer the R-squared value is to 1, the better the model fits. Contrarily, a value closer to 0 means that the model fits poorly. R-squared quantifies the fitness and accuracy in different regression models, not just linear models. For instance, an R-squared of 70% indicates that the regression model explains 70% of the variability presented in the output variable. A higher R-squared means that the model can explain more variability [26].

The significant advantage of Linear Regression ML models is that they are easy to implement, and the output coefficients are easier to comprehend. If the input/output relationship is linear, this type of algorithm is even better to implement since it is less complex when compared to other algorithms. Nevertheless, outliers in the linear regression approach can significantly impact the regression. Additionally, the model's boundaries are linear, and the dependent and independent variables are assumed to have a linear relationship. However, when this relation is not strictly linear, the model does not capture the learning patterns well [37].

2.2.2 Random Forest Model

A Decision Tree (DT) algorithm is a ML system that classifies data using a tree structure and hierarchical logic. The concept behind a DT is that the tree is built by splitting the data into smaller subsamples based on predefined criteria that determine the tree's structure. The decision tree begins with the root node, representing the best potential factor for splitting the data. Typically, the splitting criteria are the Mean Absolute Error (MAE), or the Mean Squared Error (MSE). The tree becomes deeper and deeper until no improvements to the specified splitting criterion are made, or there are no samples or factors to divide the data. The leaf node is the node at the end of the DT without any possible split [38].

The RF algorithm is an ensemble learning algorithm (a combination of different learning algorithms to achieve higher prediction performance) for classification or regression problems based on a DT. In the RF algorithm, bootstrap samples are taken from the original data. It constructs untrimmed classification or regression trees for each bootstrap sample, which is a data sample obtained from a training set with replacement. This is also known as bagging or bootstrap aggregating. However, instead of picking the best split among all predictor variables at each node, it randomly chooses the predictor variables and selects the best split among them, ensuring that DTs have a minimal correlation. This is a significant distinction between DT and RF. RF chooses only a subset of the potential feature splits, whereas DT analyzes all possible feature splits [38].

RF algorithms have some key hyperparameters, a parameter that is defined prior the training phase and will be used to control the learning process. For instance, these hyperparameters include the number of trees, the number of features sampled, and the maximum path between the root and leaf node, amongst another. A single tree in the ensemble is made up of a bootstrap sample, and a collection of DT constructs the algorithm. One-third of the training sample is reserved as test data, the out-of-bag sample. The prediction of the model will differ depending on the type of task. Individual DTs will be unweighted averaged for the regression task. The most common categorical variable of all the trees will give the final predicted class for the classification task.

Every node in the DT is a condition on a particular feature, meant to divide the dataset such that identical response values are in the same set. Impurity refers to the metric used to select the ideal condition. Hence, an essential point in RF models is to find the most significant feature to split using a suitable criterion. For classification tasks, Gini index and entropy are commonly used. For regression tasks, the variance reduction using MSE or the MAE are commonly used [39].

The low correlation between the trees is a significant benefit of this approach. Another advantage is the reduced risk of overfitting. If there are many DTs in RF, then the classifier will not overfit the model since averaging uncorrelated trees reduces the overall variance and prediction error [38].

Linear regression models are the most straightforward and intuitive ML techniques and do not provide the black box effect that RF does. Linear Regression models are extensively utilized, and when used in the appropriate data set, they are a potent prediction tool. However, they are substantially less agile than RF since they only employ linear data. RF produces best forecasts when the data is non-linear [38].

2.3 Model Evaluation

After the training phase, the model is assessed in a test dataset. There are different techniques to perform this task, the two most common are [42]:

- Hold-out validation: Technique that divides the data into separate groups, one for training and the other for testing. The training set is used to train the model, whereas the test set is used to evaluate how well the model works on data that has not yet been observed by the model. When employing the hold-out approach, a typical split is to use 80% of the data for training and the remaining 20% for testing. However, there are others possible splits, such as 70% for training and 30% for testing.

-Cross Validation (CV): Technique that divides the original dataset into two parts: a training set that trains the model and an independent set for validation, a test set. The most common method is the k-fold CV. The process requires a unique parameter, k, that specifies the number of samples into which a given data sample will be divided. The sample is subdivided into k smaller subsets, and k-1 folds are used as training data to train a model, and the resulting model is validated using the remaining data. The k-fold CV performance metric is the average of the values obtained in the loop. This method is computationally intensive, yet it does not waste much data [42].

Evaluating a model is a critical component of developing an effective ML model, and different evaluation metrics can be employed based on the type of problem, classification, or regression.

For regression problems, besides the R-squared, **equation 2.8**, the two most common model evaluation error metrics are the root mean squared error (RMSE) and the MAE, **equation 2.9** and **2.10**.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\widehat{(y_i} - y_i)^2} \tag{2.9}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \tag{2.10}$$

The RMSE, presented by **equation 2.9**, measures the discrepancy between the predicted values from the model ($\hat{y}_i$) and the actual value of the estimated variables (y_i), and n is relative to the number of fitted points. The deviations between the predicted and actual values are squared separately and averaged through the sample, then, the mean square root is computed. Because errors are squared before averaging, the RMSE attributes more weight to larger errors, which is particularly beneficial in cases where large errors are especially undesirable [40].

MAE of the model is the average of the absolute values of each prediction error over all occurrences in the test set, **equation 2.10**. Statistically, it evaluates the accuracy of two continuous variables. MAE is a linear score; therefore, all individual errors are equally weighted [40].

The MAE and RMSE can be used in conjunction to identify error variance in a collection of model predictions. These metrics allow to analyzing how far the predictions differed from the actual result. However, they do not provide any indication regarding the direction of the error (under- or over-predicting the data). The RMSE value is always equal to or greater than the MAE, and the bigger the difference between these two metrics, the bigger the difference between the single errors in the sample. If they are equal, all errors have the same magnitude.

Another standard metric used in statistics is the Mean Absolute Percentage error (MAPE). MAPE is a not scale-dependent metric that measures the forecast system's accuracy. It is determined by **equation 2.11**, where A_i represents the actual value and F_i the forecast value. Since this is a not scale-dependent metric it is common to multiply by 100% the **equation 2.11** to get the MAPE as a percentage [40].

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{A_i - F_i}{A_i}\right| \tag{2.11}$$

MAPE penalizes negative errors more severely than positive ones. As a result, when MAPE is used to evaluate the accuracy of prediction models, it is biased since it will continually select a method with too low predictions [40].

2.4 Hyperparameter Optimization or Parameter Tuning

The parameter tuning process can be defined as selecting an ideal combination of hyperparameters for the learning algorithm. The model hyperparameters define the structure of the prediction model and the quality of the predictions. For more complex models, such as RF, it is necessary to set these hyperparameter values to build the model, such as the number of trees, the depth of the decision tree, the maximum number of features, etc. [41].

Hence, when building ML models, the critical point is to set up and combine the values of the hyperparameters to get the best model performance. The first step in an optimization technique is to define a search space in which each dimension represents a hyperparameter, and each point indicates a different model configuration. The optimization procedure aims to identify the vector that results in the highest model performance after learning, such as maximum accuracy or least error. Several optimization techniques can be utilized, but the most frequent are random search and grid search methods [41]:

-Random Search: defines a bounded search space for hyperparameters values and randomly selects points in this space;

-Grid Search: defines the search space as a grid composed of different hyperparameter values and evaluates every single position of the grid.

Although it generally takes longer computation time, random search is more helpful in finding hyperparameter combinations where there is any a priori intuitive knowledge about these values. The

grid search method is excellent for testing combinations that work well in the specific model. Both random and grid search evaluate models for a specific hyperparameter vector are tested using CV. Multiple iterations of the complete k-fold CV process are run to tune the hyperparameters with a different model setup, **Figure 2.2**. After comparing all models, the best one is chosen, trained on the entire training set, and assessed on the testing set [42].

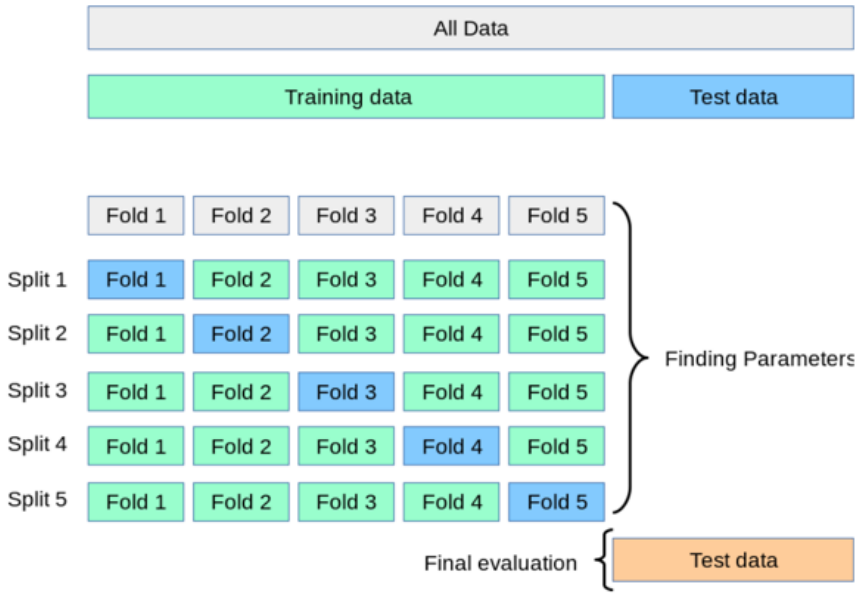

Figure 2.2- Cross Validation Schema [42].

2.5 Feature Importance and Significance

Feature importance is the process that calculates the scores for each input feature of the model. The scores indicate the feature impact on the model. The bigger the score, the greater the impact on the model used to predict a given variable. These scores are beneficial for data understanding, model improvement, and interpretability. They allow to explain the connection between features and the output variable and determine which features are more significant to the model. Analyzing the features scores is an essential tool to reduce the dimensionality of the model since the higher scores present a higher predictive power of the model and, therefore, should be kept. The lower scores may be eliminated without negatively influencing the model's performance. Another relevant point is the model interpretability, especially in healthcare projects, since the feature importance provides a most straightforward model interpretation, especially for black-box models such as RF, facilitating the model presentation to other stakeholders [38].

Commonly, the p-value is used in linear models to understand if the variable is significant for the output prediction or not. A p-value is a number determined on a statistical test. This value describes the probability of finding a result when the null hypothesis is true and is used to accept or exclude the null hypothesis. If the p-value is less or equal to the pre-defined alpha level, the null hypothesis is rejected. When determining the feature importance, the p-value is determined on the t-statistic, statistical propriety that provides the distance between the mean to zero in terms of standard deviation and gives the probability if there is any relation between the feature and the model's output.

On the t-test test, the slope of the independent variable is assumed to be equal to zero, indicating that it has no significant influence on the dependent variable's prediction. Regression coefficients (β) are therefore deemed significant when their p-values are less than the defined alpha level, usually defined in the literature as 0.05. If the p-value is less than the alpha value it is possible to affirm that there is a relation between the output and the feature or that there is a 5% probability that there is no relation between the output and the feature, therefore the feature must be maintained in the model [34].

At the end of this chapter, all the theory behind the ML approach to solving regression tasks, such as the problem under analysis, is described and was the basis for this dissertation framework.

3. State of the Art

Predicting the duration of surgeries accurately is a challenging subject that has received substantial attention in the literature. Searching for keywords such as: OR efficiency and scheduling, operation case duration, model prediction, and ML yielded a set of relevant research published in open databases. Most of them on ML prediction are from the previous five years, showing that the study direction of this theme is cutting-edge. This section aims to understand, firstly, which are the challenges and then how ML algorithms can improve the OR time prediction. The first section describes the studies regarding OR efficiency, and the second describes the ML approach to estimate OR case-time duration. In section 3.2 a resume table, **Table 3.1**, resumes the key characteristics of the ML studies found in the literature.

3.1 Studies related to Operating Room Efficiency

It is critical to analyze the current efficiency of the OR, i.e., to determine the criteria for measuring OR efficiency, in order to increase its efficiency. The criteria used to determine if an OR is efficient varies depending on area, hospital, and OR size. However, researchers have developed several common assessment markers based on case studies, fact-finding, quantitative data, and other methodologies.

Olivares et al. [43] estimated that the costs of OR underutilization are 60% bigger compared with OR overutilization. OR underutilization is caused by a variety of factors. To investigate which factors contribute the most to OR underutilization, Tankard et al. [44] conducted a study where the means of different features that contribute to OR underutilization were compared. These features were: patient in the room, turnover time, scheduling gaps, OR holds and closed rooms. The authors concluded that mid/end-of-day gaps (i.e., when ORs were unoccupied and had at least one case completed, but setup/cleanup times could not be shifted out of this period) and closed rooms (i.e., when a staffed OR remained unoccupied for a whole day, without cases, for a maximum of 8 hours per room per day) presented a bigger impact in OR underutilization. While turnover time (i.e., time longer than 45 minutes between "patient out" of room to next "patient in" room in the same OR) and ‘‘patient in the room’’ (i.e., the time that elapses following a 4:59-minute grace period is added to the scheduled start time of the first case of the day in each room) presented a lower impact.

There is no singular metric or set of measurements that has been established to evaluate surgical care efficiency best. However, Dexter et al. [45] measured the OR efficiency in hours of OR time that are underused or overutilized. The authors defined OR inefficiency as the sum of two products: the hours of underused OR time and the cost per hour of underutilized OR time plus the hours of overutilized OR time and the cost per hour of overutilized OR time. Hence, the OR time is maximized by minimizing the sum of these products. The equation is simplified on the day of operation since the cost per hour of underused OR time is a sunk cost, i.e., invested money that cannot be recovered. As a result, by the day of surgery, increasing OR efficiency is achieved by reducing the overutilization of OR time. The most crucial step in increasing OR efficiency is properly allocating OR time, which can be achieved by precise OR time predictions. Then, on rare occasions, services should find themselves in a scenario where their allotted OR time is complete, but they still have another case to schedule.

3.2 Studies related to Machine Learning Models to predict Operating Times

Considering the focus of this dissertation, in enhancing case duration prediction to maximize OR efficiency, this section presents the recent studies that apply ML to predict OR times and case-time duration. The utilization of the EHR based on past data for a specific operation and/or surgeon is a standard method to calculate OR times and case-time duration. Most investigations compared the performance of prediction models to this old technique approach. This section compiles some of the relevant studies found in the literature followed by **Table 3.1** which summarizes the most relevant topics of the ML surgery time prediction studies found in the open databases. This table includes the most relevant topics under analysis for this project: the title of the study, the input model variables, the model output, the ML algorithm implemented, the metrics for model evaluation, the study period followed by the specialties where the model was applied, and finally some observations and results relevant for the analysis.

It is possible to notice in **Table 3.1** that most authors implemented linear regression-based algorithms in their studies, followed by regression DT-based algorithms such as RF, Extreme Gradient Boosting (XGB), Gradient Tree Boosting model (GBM), etc... Fewer studies, such as Devi et al. [46], apply other ML algorithms such as Artificial Neural Network or Adaptive Neuro-Fuzzy Inference Systems.

Firstly, before implementing a ML model, a relevant point is understanding the data distribution. Strum et al. [47] established a comparison of log-normal and normal models in 40 076 surgical cases to determine which distribution fits better the surgical data. In this study, the Shapiro-Wilk test was used to determine the goodness-of-fit for both models. The test shows that the log-normal model outperforms the normal model regarding a diversified group of procedures. Using the normal distribution might skew results from commonly used statistical techniques, while the log-normal distribution is better suited. Whereas nonparametric techniques, such as Kruskal-Wallis instead of Analysis of Variance (ANOVA), can be used to sidestep the challenge of model selection, these analyses are often less powerful than similar parametric ones when the data is normally distributed.

Edelman et al. [48] employed a Linear Regression model to estimate the surgery time for six different university hospitals. When only a few factors such as patient's age, type of anesthesia, and pre-surgical length predictions were employed, their results indicated reduced prediction time errors (an error of 39.5 min with a MSE of 3859.6 min in a sample of surgical procedures with a total procedure mean time of 150 min). Except for some levels of the categorical variables for type of anesthesia and type of surgery, all variables in the linear regression models were highly significant predictors ($p < 0.01$). Since the overall effect of these variables was significant, these variables were maintained in the model. Similarly, Eijkemans et al. [49] also employed a Linear Regression model to predict the OR time, but with a higher number of factors. Besides the patient characteristics, the authors also include team characteristics, such as the number and age of surgeons and anesthetists and session characteristics. Among these features, the operation and team features are shown to have the best predictive performance, and the patient characteristics have a smaller significance. Additionally, the authors reported that operating time was shorter for patients over 60 years old and higher for patients with a greater Body Mass Index (BMI). Implementing the prediction model rather than the surgeon's prediction based on historical

averages would reduce under- and over-predicted OR times by 2.8 and 6.6 minutes, which corresponds to a reduction of 12% and 25%, respectively.

Kayis et al. [50], in their study, implemented a Linear Regression model with Elastic-net regularization in two years of surgical data across 21 different specialties. The authors proved that their model, with an R-squared of 0.64 and a mean absolute deviation of 42.65± 0.59 minutes improves estimation accuracy by 1.98 ± 0.28 minutes, particularly by decreasing large errors and not only operational and temporal parameters but also medical staff and team experience-related characteristics (number of nurses and the frequency with which the medical team collaborates) might be used to improve the current estimations.

By implementing variables from a data mining processing of the hospital's clinical histories, Hosseini et al. [51] used a classical least square linear regression and a stepwise regression to estimate the surgical time across 15 specialties. The authors included six input variables: surgery type, procedure type, physical status, patient age, surgery scope, and specialty. The results revealed a satisfactory approximation of surgical time, equivalent to a manual prediction approach based on the average duration of the surgeon's most recent procedures and an adjustment made by the scheduler based on experience.

Many authors included linear models, DT-based models, and different types of ML algorithms to analyze which responds with a better performance to the proposed problem. For example, Shahabikargar et al. [52] used ensemble algorithms (e.g., RF, Bagging, and LSBoost) to estimate surgical duration at the specialty level, such as cardiothoracic, urology, plastic surgery, and a few others. The prediction model's performance changed a lot according to the specializations. Ophthalmology showed the most promise, and the RF technique reduced the overall prediction error by 44 % (MAPE reduction from 0.68 to 0.38) when the authors applied filtered data.

Besides the linear regression model, Huang et al. [53] also applied RF and XGB across 25 different specialties. The authors did not report significant differences between the RF and XGB. Even though the RF evaluation metrics were equal to the XGB model ones, the XGB model was still able to lower the overall prediction error (in minutes) compared to the average, Reg, and log reg models. The XGB model presented an R-squared of 0.82 and a MAE of 30.2 minutes. Bartek et al. [54] also concluded that the XGB algorithm outperforms the linear regression and the current case-time duration estimation, as well as estimates provided by surgeons themselves. For their model, the authors came to the conclusion that the preponderance of the information used in the models was based on the procedure and staff information. Shahabi Kargar et al. [52] implemented a Linear Regression, a Multivariate Adaptative Regression Splines, and RF algorithms to predict procedure time estimation in twelve specialties. The authors concluded that the linear regression model was poor compared to the hospital estimate of procedure time. The authors speculate that this could be due to the fact that surgeons estimate the time based on their experience and implicitly consider the interactions between variables. In contrast, these interactions were not considered by their linear regression model. The study results reveal that the RF model outperforms other methods, by reducing 28% the MAPE when compared to the current hospital estimations. Martinez et al. [25] examined a ML-based surgical time predictive model. The authors analyzed four different ML models (Linear Regression, Support Vector Machines, Regression Trees, and Bagged Trees) to estimate the surgical surgery length. The models were compared in terms of RMSE. The four algorithms presented an error between 26 to 37 minutes. The Bagged Trees presented the best performance (RMSE= 26 min and 3.16 min of training time) when the database with nine specialties was used, representing 80% of the total surgeries. The authors also noted that with a reduced RMSE, the Bagged Trees model outperformed the experience-based technique.

Some authors only focused on data originated from single departments or specialties. Devi et al. [46] developed a methodology to predict the surgery time in an ophthalmology department using surgical predictors such as the experience of the surgeon in years, the experience of the anesthetist in years, type of anesthesia, etc. in three ML models (Adaptive Neuro-Fuzzy Inference Systems, Artificial Neural Networks, and Multiple Linear Regression Analysis). However, the authors only included three types of surgeries (corneal transplant, cataract, and Oculoplastic surgery) which leads to some questions and limitations regarding the generalization capacity of the model when applied to other ophthalmic surgery procedures.

More recently, due to a more extensive implementation and the tendency to perform robot-assisted surgery, some authors use ML models to predict the Robotic Assisted Surgery (RAS) case duration. For instance, Zhao et al. [55] tried to predict RAS by implementing six different models (Multivariable linear regression, Ridge regression, Lasso regression, RF, Boosted Regression Tree, and Neural Network) with twelve input variables, that included the robot model, in twelve different procedures. The authors discovered that all ML models reduced the average RMSE compared to the baseline model. The Boosted Regression Tree presented the lowest RMSE, of 80.2 minutes, by using this model the authors were able to increase the number of schedule cases from 148 to 219 cases. Although there are not many ML model studies to predict the RAS time duration, the tendency to increase the performance of RAS by the development of cut-edging technology accompanied by the higher prediction capacity of ML algorithms will allow the research on these types of operations across different specialties.

As reported in this section, many recent studies address the problem of predicting OR time across different specialties, departments, and different types of procedures. Predominantly, the ML algorithms identified were Linear Regression models and DT-based algorithms, such as RF and XGB. The number of model features was highly variable, depending on the authors and the used data. Nevertheless, it is possible to distinguish a few categories of variables: patient-related variables (i.e., sex, BMI, medical history, age…), procedure characteristics, such as the primary procedure, and surgery team characteristics (i.e., surgeon/anesthetist unique identifier, team size…). Additionally, it is possible to test two approaches: develop a model for different type of procedures and specialties or for a single specialty.

Table 3.1- Resume table of the studies found in the literature relative to Machine Learning models to predict surgical time.

Title	Model Input (Features)	Model Output	ML Algorithm	Model Evaluation	Study Period	Specialties	Observations/Conclusions
A Machine Learning Study to Improve Surgical Case Duration Prediction [53]	Primary Surgeon's Prior Events: -Total number of previous surgeries -Total minutes spent on previous surgeries within the same day as well as within the last 7 days -Number of urgent and emergent operations prior to the case that was being performed by the same surgeon Patient: -Age -Gender -ICD code -In- /outpatient -ASA status -Hypertension, Anemia and Diabetes Surgical team: -Primary surgeon's ID -Surgeon team size Specialty -Primary surgeon's gender -Primary surgeon's age Operation: -Procedure type -Sub-procedure type -Anesthesia type -Facility and Room No. -Day of the week and time of the day	Case Duration	-Linear Regression -Random Forest -Extreme Gradient Boosting (XGB)	-R-squared -MAE -Percentage overage (actual duration longer than prediction), underage (shorter than prediction) and within (within prediction), with a 10% threshold	January 2017 to December 2019 (170,748 records)	25 different specialties (422 types of procedures): Ophthalmology, Cardiovascular; Otorhinolaryngology, Head and neck, Trauma and acute care, Obstetrics, and gynecology, Colorectal, Urology, Body science and metabolic disorders, Breast surgical oncology, Plastic and reconstruction, General, Orthopedics, Thoracic, Neurosurgery Oral and maxillofacial, Anesthesiology, Gastroenterology and hepatology, Dermatology, Pediatric, Bariatric and metabolic	Even though the RF evaluation metrics were equal to the XGB model, the XGB model was still able to lower the overall prediction error (in minutes). When compared to the average, Reg, and logReg models, the XGB model performed better in terms of prediction. The XGB model was more computationally efficient in that it finished the training process in less time when compared to the other algorithms.
Machine learning for surgical time prediction [25]	-Surgeon -Procedure -Surgeon Experience Specialty -Patient Destination -Patient Life Stage	Surgical Time	-Linear Regression -Regression Trees -Support Vector Regression -Bagging Regression Trees	RMSE	December 2004 to April 2019 (206,587 records)	9 specialties: Orthopedics and traumatology, general surgery, gynecology and obstetrics, neurosurgery, urology, plastic surgery, otorhinolaryngology,	There were three scenarios elaborated to select which is the best to apply the ML algorithms: All data; Nine specialties; one specialty. In terms of RMSE and computation time, Bagged Trees performed the best. The best scenario for Bagged Trees'

						ophthalmology, and head and neck surgery	performance was the nine specialties scenario.
Improving Operating Room Efficiency: Machine Learning Approach to Predict Case-Time Duration [54]	Patient factor: -Age -Sex -BMI -Patient admit class -Preoperative diagnosis -Medical history condition Procedure factor: -Primary procedure -Primary procedure category -First/second/third/fourth/fifth sub procedure -Operative modifier Personnel factor: -Surgeon unique identifier -Historic primary procedure duration -Historic sub procedure duration	Case-time duration	-Linear regression -Extreme Gradient Boosting (XGBoost)	-R-squared -Accuracy -Predictive capability of being within a 10% tolerance threshold: 1.Overage (case duration > predicted + 10%). 2.Underage (case duration < predicted - 10%) 3. Predictive capability	January 2014 to December 2017 (46,986 records)	12 surgical services: General, Cardiac, Thoracic, Vascular, Transplantation, Neurosurgery, Plastic, Orthopedics, Gynecology, Urology, Otolaryngology, and Oral-Maxillofacial Surgery	The XGBoost ML surgeon-specific models outperformed linear regression and current case-time duration estimation, including a historical average per surgery type and surgeon, as well as estimates provided by the surgeons.
Improving the efficiency of the operating room environment with an optimization and machine learning model [56]	-Procedure -Weight -Age and Sex -Scheduled postop destination Service -Scheduled procedure length -Patient class -Location -Radiology	Required PACU time for each type of surgical procedure	Gradient tree boosting model	Accuracy	May 2014 to June 2016	NA	Despite the presence of uncertainty in procedure and recuperation periods, the authors demonstrated that they might obtain a considerable improvement over the previous scheduling system. The author's methodology has decreased overall PACU holds by 76% without reducing operating room use in the second half of 2016 data.
Predicting Procedure Duration to Improve Scheduling of Elective Surgery [52]	Patient characteristics: -Age -Gender -Type of admission Classification -CCI -Referral center Operation characteristics: -Procedure indicator Unit -Specialty -Theatre -Order -Ward	Procedure time estimation	-Linear Regression (LR) -MARS (Multivariate Adaptive	-RMSE -MAPE -R-squared	June 2018 to June 2012	12 specialties Cardio-thoracic, Enterology , General, Gynecology, Neurosurgery, Ophthalmology, Orthopedics, Plastic and Reconstructive, Urology, Vascular, Other Surgeries	The performance of the linear regression model was poor when compared to the baseline, i.e., the hospital estimates of procedure time. The authors speculate that the reason for this could be the fact that surgeons estimate the time based on their experience and implicitly consider the interactions between variables, whereas these interactions were

	-Sub specialty <u>Procedure:</u> -Primary Procedure class -Session type <u>Surgery team characteristics:</u> -Consultant -Con. category -Surgeon -Surgeon category -Surgeon-Consultant Surgeons -Anesthetists -Team size		Regression Splines) -RF				not considered by their linear regression model. Cross-validation results reveal that the RF model outperforms other methods.
Case study of the prediction of elective surgery durations in a New Zealand teaching hospital [57]	-Type of Surgery -Estimate of duration (possibly transformed) for a surgeon chosen to be the baseline -Surgeon	Procedure time estimation	Hierarchical linear regression (HLR)	-RMSE -MAE	1 February 2015 to 14 May 2018	Ear, nose and throat (ENT) elective surgeries	HLR approach performs marginally better than the Arithmetic Mean approach, and either approach generally gives better predictions than the surgeons' estimates
Machine Learning Can Improve Estimation of Surgical Case Duration: A Pilot Study [23]	<u>Patient:</u> -Age, sex, height, weight -Allergies -Medical conditions -ASA physical status classification <u>Providers:</u> -Surgeon(s) -Anesthesiologist(s) -Scrub nurse(s) and technicians(s) -Circulator nurse(s) and technicians(s) -Whether the assigned group has worked on this type of case before -Facility / Room -Hospital bed census Equipment -Day of the week -Time of day -Procedure -Procedure Type	Case-time duration	Combination of supervised learning algorithms (Leap Rail)	-Absolute Median Difference -Absolute differences by subspecialty	January to March 2018 (3 months)	<u>16 specialties</u> Cardiovascular, Otorhinolaryngology, Gastroenterology, General Surgery, Gynecology, Interventional Radiology, Neurosurgery, Oral/Maxillofacial, Orthopedics, Plastic Surgery, Podiatry, Interventional Pulmonology, Spine Surgery, Thoracic Surgery, Urologic Surgery, Vascular Surgery	Leap Rail was more accurate for 14 of the 16 subspecialties; however, only the findings for Gastroenterology, General Surgery, Orthopedics, and Urology were statistically significant among those 14 subspecialties. Regardless of the prediction model, both groups showed many outliers, demonstrating that intra-operative variability was challenging to account for in prediction.

	-Surgeon comments -Procedure modifiers -Anesthesia type -Implants / tissues used -Prior Events -Last food/drink intake -Timing of prior perioperative milestones -Case delays -Cancellations -Room turnover time						When the results were split down by specialty, cardiology showed the most significant substantial improvement (albeit this impact was not statistically significant), followed by orthopedics and urology
Improving the Prediction of Total surgical Procedure Time Using linear regression Modeling [48]	-Estimated surgeon-controlled time (eSCT) -Patient age -Type of operation -American Society of Anesthesiologists (ASA) physical status classification -Type of anesthesia used -Main specialism	Total procedure time (TPT)	Linear Regression Models	-MAE -MSE -R-squared of the model	2012-2016 (79,983 records)	<u>20 specialties:</u> Ophthalmology, Ear, nose, and throat, Cardiothoracic surgery, Orthopedic surgery, Neurosurgery, Plastic surgery ,Oral and maxillofacial surgery, Obstetrics and gynecology, Abdominal surgery, Urology, Surgical oncology, Traumatology, Obstetric and gynecological oncology, Miscellaneous, Pediatric surgery, Vascular surgery, Hepatobiliary surgery, Transplant surgery, Anesthesiology, Pediatric gastroenterology	Linear regression model using the estimated surgeon-controlled time, type of operation, ASA classification, and type of anesthesia conduct to a better TPT prediction. When the patient's age was included in the model, it did not significantly improve the model. The Linear Regression model developed by the authors outperforms the current methods of using a standard duration for the ACT or a fixed ratio between SCT and TPT.
Predicting the Unpredictable [49]	<u>Session characteristics:</u> -Number of separate procedures -Laparascopic procedure -Year of surgery <u>Team characteristics:</u> -Number of surgeons -Summed ages of the surgical team -Age of the youngest surgeon -Age of the oldest surgeon. -Number of anesthesiologists	OR time	Linear mixed modeling (with the logarithm of the total OR time as the dependent variable)	-Adjusted R-squared -Adjusted R-squared Gain, Relative to the Base Model (%)	Elective operations performed by the department until June 2005 (18,838 records)	General Surgery	The operation and team features had the best predictive performance. However, patient factors had a small but significant influence on OR time.

	-Summed ages of the anesthesiologists -Age of the youngest anesthesiologist -Age of the oldest anesthesiologist Patient characteristics: -Age and Sex -Number of previous admissions -Length of the current admission -First operation -BMI -Presence of cardiovascular risk factors						The surgeon's assessment made a substantial and independent impact on the prediction
Improved Prediction of Procedure Duration for Elective Surgery [58]	Patient characteristics: -Patient age and gender -Urgency category -Type of admission -Patient payment class -Referral center -Charlson Comorbidity Index Hospital and operation characteristics: -Hospital unit and specialty -Ward and theatre -Session -Surgery team category -Number of surgeons: -Anesthetists -Their professional category and specialty	Surgery duration prediction	-Generalized Linear Model (GLM) -Multivariate Adaptive Regression Splines (MARS) -Random Forests algorithms -LS Boost. -Bagging algorithms	MAPE	July 2008 to June 2012 (60362 records)	104 different types of procedures across 11 surgical specialties: Cardio-Thoracic, General Surgery, Neurosurgery, Orthopedics, Urology, Gynecology, Ophthalmology, Plastic Surgery, Vascular Surgery	Using filtered data results in a 44 % reduction in overall prediction error (MAPE reduced from 0.68 to 0.38). By employing the Random Forests algorithm while using the newly developed ensemble approach delivered, the authors achieve a MAPE of 0.31, representing a 55% reduction relative to the original error and a reduction of 18% compared to the RF implemented on filtered data.
Data-Driven Surgical Duration Prediction Model for Surgery Scheduling: A Case-Study for a Practice-Feasible Model in a Public Hospital [59]	Surgery Factors: -Department Code -Op. Theatre Code (OT) -OT Location Code -Proc. Code (PC) -Proc. Desc. -Proc. Surgical Table Code -Operation Risk -Type Of Anesthesia -Method of Operation Patient Factors: -Type Of Operation	Surgical duration prediction	-Two-step data-mining model (The model first uses domain knowledge to estimate the first surgeon rank, and then uses this predicted attribute, as well as other predictors (surgical team, patient,	RMSE	2016 and 2017 (41,000 surgical records)	NA	The experimental findings reveal that the suggested the author's two-step model is more parsimonious and beats the hospital's existing moving averages strategy. Compared to the baseline model, the Gradient Boosting Model achieves a much lower root mean square error.

	-Gender and race -Weekday -ASA Status <u>Surgical Team Factors:</u> -Surgical Team Size -1st, 2nd and 3rd Surgeon ID -1st, 2nd, 3rd Surgeon Title -P.Anaes. ID -P. and Asst Anaes. Title -Surgical and Anaes. Student -Consultant ID and Title <u>Temporal Factors:</u> Moving Avg Dept, TC, OT, Diag, PC		temporal, and operational factors) in a tree-based model to predict surgical durations) -Ensemble approach using GBM				
Prediction of Surgery Times and Scheduling of Operation Theaters in Ophtholmology Department [46]	-Experience of accompanying theater staff -Type of anesthesia -Experience of anesthetist -Patient preconditions (like existence of Redness of eye, diabetics, hypertension, watery eyes or any other sources of infection) -Patient age	Estimation of surgery times	-Adaptive Neuro Fuzzy Inference Systems (ANFIS) - Artificial Neural Networks (ANN) -Multiple Linear Regression Analysis (MLRA)	-Average root mean square error -RMSE	NA	Three ophthalmologic surgeries: Cataract surgery, Corneal transplant surgery, Oculoplastic surgery.	The ANFIS model outperforms the other two models. In simulations, encouraging findings from the authors for optimal Operation Theater scheduling were found, which suggest that the same must be validated in a real-world setting in a hospital.
A robust estimation model for surgery durations with temporal, operational, and surgery team effects [50]	-OR Suite Number -Specialty, Procedure, Attending Surgeon, and Encounter ID -Inpatient/Outpatient Estimated -Number of Surgeons, Anesthesiologists and Nurses Estimated -Number of Cases on the Day and in the Same OR on the Day -Sequence Number of the Case in OR -Number of Same Specialty Cases on the Day and in same OR -Number of Cases Attending Surgeon has on the Day -Time of Day -Weekday, Month and Year -Estimation Type -Attending Surgeon and Anesthesiologist Joint Experience	Surgery duration estimation	Linear regression with Elastic-net regularization	-RMSE -R-squared - Mean Absolute Deviation	Two years (10292 records)	633 different procedure types under 21 different specialties	The authors discovered that not only operational and temporal parameters but also medical staff and team experience-related characteristics (such as the number of nurses and the frequency with which the medical team collaborates) might be used to improve the already used estimations.

	-Attending Surgeon-Nurse Joint Experience -Attending Anesthesiologist-Nurse Joint Experience -Attending Surgeon-Anesthesiologist Joint Experience Frequency -Attending Surgeon-Nurse Joint Experience Frequency -Attending Anesthesiologist-Nurse Joint Experience Frequency						
A Machine Learning Approach to Predicting Case Duration for Robot-Assisted Surgery [55]	-Scheduled duration Procedure group Elderly (age > 65) -Obese (BMI >30) -Gender -Combined case -Robot model -Malignancy -Tumor chest, abdomen, head and neck, pelvis and retroperitoneum -Hypertension Smoking history -Atrial fibrillation	Robotic Assisted Surgery (RAS) case duration	-Multivariable linear regression -Ridge regression. -Lasso regression. -Random Forest -Boosted Regression Tree. -Neural Network	RMSE	January 2014 to June 2017	<u>12 Procedures:</u> Bowel resection, Cystectomy, Low anterior resection, Myotomy, Partial nephrectomy, Radical nephrectomy, Radical prostatectomy, Salpingnoophorectomy (benign), Simple hysterectomy (benign), TAH-BSO (malignant), Trans-oral robotic surgery, Ureteral reimplantation	The authors discovered that compared to the baseline model, all machine learning models reduced the average RMSE. The boosted regression tree had the lowest average RMSE, considerably lower than the baseline model.
Surgical Duration Estimation via Data Mining and Predictive Modeling: A Case Study [51]	-Specialty -Priority Class -ASA Class -Age -Encounter Class -Procedure Code	OR time	-Classical Least Square Linear Regression (LIN) -Stepwise Regression (STEP)	-R-squared -RMSE -MAE	3.25 years	<u>15 Specialties:</u> Orthopedics ,General Surgery ,Otolaryngology ,Urology, Ophthalmology, Neurosurgery, Surgical Oncology, Plastic Surgery ,Thoracic ,Vascular ,Obstetrics ,Psychosurgery, Acute Care Surgery , GYN Oncology , Gynecology	Both LIN and STEP yield better forecasts than the existing state-of-the-art and relative to the current practice in the service. The hybrid technique using STEP regression predicts better for orthopedics, general surgery, and surgical oncology specialties, which account for more than half of all procedures.

4. Methods

This chapter describes the materials and methods used for the development of the ML models to predict OR time in the two proposed approaches: the first approach in all specialties and a single specialty model for the second. Firstly, the data-collection process, description and preparation are detailed. The models' conceptualization is then described and evaluated, along with feature importance, and comparison with current methods. Python 3.9.7 was used for this project, with the open-source library SciPy for the statistical analysis and the Scikit learn library for the ML algorithms construction.

4.1 Materials: Data Collection and Description

The dataset used in this project consists of the collection of 48737 surgeries performed on CHULC for five years, from January 2017 until December 2021, including three specialties of Urology, General Surgery, and Orthopedics. The study protocol and notification regarding the general data protection regulation were submitted and approved by the hospital's management committee and ethics commission. All data was anonymized before analysis. The surgeons and anesthetists are portrayed anonymously throughout the project to safeguard their privacy. Among the three specialties analyzed in this dissertation, General Surgery and Urology are the only ones where RAS is currently implemented.

The initial dataset presented 52 variables containing information regarding the case ID, specific details about the patient, surgical team characteristics, and the hospital's information respecting the place and type of surgical procedure. These variables can be grouped into patient, surgical procedure, and surgery team characteristics. Patient characteristics comprised the patient's age, gender, urgency, priority, waiting days, type of admission and scheduling, diagnosis code according to the International Code of Disease (ICD), ICD-9 or ICD-10, and anesthetic risk, according to the American Society of Anesthesiologists (ASA). Procedure characteristics include all factors linked to the hospital and operation, such as hospital unit, specialty, ward and theater, as well as the type of operation and anesthesia. Finally, the surgical team's characteristics include all factors related to personnel engaged in the surgery, such as the principal surgeon and anesthetist's ID. The specific time of each surgery includes the initial and final time of patient arrival and leaving the OR, as well as the time of starting and finishing the anesthesia and surgery. This enabled an examination of operation length from the surgeon's, anesthesiologist's, and OR's points of view.

Table 4.1 contains all the initial database variables and the newly generated variables, which will be described in detail in the following sections. The table presents the variable name followed by a detailed description, the variable type (numerical or categorical), and the predictor type, i.e. if the variable contains information about the patient, procedure, or surgical team, these characteristics were included for all variables. For categorical variables, the number of categories/possible values was also added. The color code gives the information if the variable was implemented in the model, represented by the green and blue color, or if it was not, represented by orange and red color. The difference between the green and blue colors is relative if the variable was initial on the database, green color, or if that variable was generated through mathematical operations or based on the domain knowledge of the reality of data, blue color. The red color refers to redundant variables that were eliminated, as described in section 4.2.3, and the orange color gathers the variables that were not selected for the models. The

feature selection criteria for variable selection are explained in section 4.4. All variable names were maintained in the original name as they were extracted from the CHULC database since this is how they are currently described in the hospital's database and for future processes, when new data from this database is added, it will facilitate the comparison among variables. Additionally, when discussing with the stakeholders the original variable name facilitates communication because their designations are standard in their clinical quotidian.

Table 4.1- Centro Hospitalar Universitário Lisboa Central dataset variables and new generated variables description.

Variable Name	Variable Description	Type	Nb of levels	Predictor Type
ID	Identification of the surgical patient	Categorical	48737	Patient characteristic
SEXO	Patient's gender	Categorical	2	Patient characteristic
IDADE	Patient's age	Numerical	———	Patient characteristic
DIAS_EM_LIC	Days in waiting list before surgery	Numerical	———	Patient characteristic
PRIORIDADE	Surgical priority level (Normal, urgent…)	Categorical	3	Patient characteristic
DTA_INTERV	Date of the Surgery (Day/Month/Year)	Date/Time	———	Procedure characteristic
ANO_INTERV	Year (2017 to 2021)	Categorical	5	Procedure characteristic
MES_INTERV	Month of the Year (January to December)	Categorical	12	Procedure characteristic
NDIA_SEMANA	Day of the Week (Monday to Sunday)	Categorical	7	Procedure characteristic
AMBULATORIA	If it is an ambulatory surgery, a surgery that doesn't require hospital admission (Y/N)	Categorical	2	Procedure characteristic
COD_TIPO_INTERV	"TIPO_INTERVENCAO" hospital' code	Categorical	4	Patient characteristic
TIPO_INTERVENCAO	Surgical schedule type (SIGIC, basic or additional schedule)	Categorical	4	Patient characteristic
TIP_INTERV_A	Contains additional information regarding to the level of urgency of the surgery	Categorical	3	Patient characteristic
DIAS_PRE_OP	Number of days before the surgery	Numerical	———	Patient characteristic
COD_ESP	"BLO_ESP" hospital' code	Categorical	20	Procedure characteristic
BLO_ESP	The hospital and the surgical service performing the procedure	Categorical	20	Procedure characteristic
ESPECIALIDADE (*)	Surgery Specialty	Categorical	3	Procedure characteristic
COD_BLOCO	"DES_BLOCO" hospital' code	Categorical	10	Procedure characteristic
DES_BLOCO	Description of the Operating Room	Categorical	10	Procedure characteristic
COD_SALA	"DES_SALA" hospital' code	Categorical	37	Procedure characteristic
DES_SALA	Description of the specific operating theater inside the block where the procedure will be performed	Categorical	37	Procedure characteristic
COD_ESPEC_PROV	"DES_PROVENIENCIA" hospital' code	Categorical	8	Procedure characteristic
DES_PROVENIENCIA	Patient's provenance for surgery (from an external hospital, urgence service, ...)	Categorical	8	Procedure characteristic
COD_ESPEC_DEST	"DES_DESTINO" hospital' code	Categorical	12	Procedure characteristic
DES_DESTINO	Patient's destination after the surgery	Categorical	12	Procedure characteristic
TIPO_ADMINT	Type of Patient Admission (Priority or Urgent)	Categorial	2	Patient characteristic
COD_VERDIAG	International Code of Disease (ICD-9 or ICD-10)	Categorical	2	Patient characteristic
COD_DIAGNOSTICO (**)	Diagnosis code for the specific disease based on ICD-9 or ICD-10	Categorical	2996	Patient characteristic
DES_DIAGNOSTICO	Detailed description of the diagnosis code, "COD_DIAGNOSTICO"	Categorical	2996	Patient characteristic
GRUPO_DIAGNOSTICO	Group of diagnosis code based on ICD-9 or ICD-10	Categorical	27	Patient characteristic
COD_ACTO	Surgical procedure code	Categorical	2761	Procedure characteristic
DES_ACTO	Detailed description of the procedure code, "COD_ACTO"	Categorical	2762	Procedure characteristic

ACTO_ROBOT (****)	The type of robotic procedure (if applicable)	Categorical	9	Procedure characteristic
CIRURG_PRINC	Surgeon unique identifier	Categorical	348	Surgical Team characteristic
ANEST_PRINC	Anesthetist unique identifier	Categorical	164	Surgical Team characteristic
TIPO_ANESTESIA	The type of anesthesia (general anesthesia, regional anesthesia...)	Categorical	10	Procedure characteristic
ANESTESIA	The type of anesthesia (same information as "TIPO_ANESTESIA" variable)	Categorical	10	Procedure characteristic
COD_R_ANEST	The risk of anesthesia based on ASA	Categorical	6	Patient characteristic
RISCO_ANEST	"COD_R_ANEST" hospital' code	Categorical	6	Patient characteristic
ANEST_ADICIONAL	If it is necessary additional anesthesia	Categorical	2	Patient characteristic
Ini_Sala	Time that the patient enters to the Operating Room	Numerical	———	Procedure characteristic
Ini_Aneste	Anesthesia induction starting time	Numerical	———	Procedure characteristic
Ini_Cirurg	Surgeon controlled starting time	Numerical	———	Procedure characteristic
Fim_Cirurg	Surgeon controlled finishing time	Numerical	———	Procedure characteristic
Fim_Anest	Anesthesia emergence finishing time	Numerical	———	Procedure characteristic
Fim_Sala	Time that the patient leaves the Operating Room	Numerical	———	Procedure characteristic
Prep_Sala_Ini	Initial time of cleaning Operating Room	Numerical	———	Procedure characteristic
Prep_Sala_Fim	End time of cleaning Operating Room	Numerical	———	Procedure characteristic
Pernoita	If the patient stays overnight	Categorical	2	Procedure characteristic
Alta	If the patient is stable and able to go home without medical supervision	Categorical	2	Procedure characteristic
Duracao_Prevista_no_Agendamento	Predicted Operating Room Time by the Surgeon (Used as a comparative term for analysis)	Numerical	———	Comparative term
Cod_media (***)	Arithmetic mean of the OR time for the surgical code procedure	Numerical	———	Surgical Team characteristic
Cirur_media (***)	Arithmetic mean of the OR time surgeries performed by the specific surgeon	Numerical	———	Surgical Team characteristic
Anest_media (***)	Arithmetic mean of anesthesia time performed by the specific anesthetist	Numerical	———	Surgical Team characteristic
Cirur_Ato	Arithmetic mean of surgical time performed by the specific surgeon based on the surgical procedure code	Numerical	———	Surgical Team characteristic
Anestesista_tipo	Arithmetic mean of anesthesia time performed by the specific anesthetist based on type of anesthesia	Numerical	———	Surgical Team characteristic
Tempo_Sala	Operating Room Time (Model output)	Numerical	———	Procedure characteristic
Tempo_Prep (*****)	Preparation Time	Numerical	———	Procedure characteristic
Tempo_Cirurgia (*****)	Surgeon-controlled time (SCT)	Numerical	———	Procedure characteristic
Tempo_Anestesia (*****)	Anesthesia-controlled time (ACT)	Numerical	———	Procedure characteristic
Tempo_Final (*****)	Final Time	Numerical	———	Procedure characteristic

(*) Only included in All Specialties Model
(**) Only included in Urology Model
(***) Removed after the collinearity analysis
(****) Not include in the Orthopedics model
(*****) Only used for the impact of SCT and ACT in OR time analysis

Variable included in the Model
New variable generated
Eliminated variable due to redundancy
Variable not included in the model

4.1.1 Anesthesia and Diseases Codes

Some dataset variables are written according to international standardized classifications, such as the anesthesia patient risk, the patient disease code, and the surgical procedure code. It is fundamental to comprehend the form and how the classification is designed and attributed to a better understanding when analyzing the variable. The standard classification used in this dataset regarding the risk of anesthesia, presented by the "COD_R_ANEST" variable, is the ASA physical status classification system. ASA was created with a six-level categorization to assist physicians regarding the patient's physiological status to assist in surgical risk prediction. An ASA I corresponds to a healthy patient without acute or chronic disease and an average BMI. An ASA VI is referring to a vegetative state where a patient has been declared brain dead and whose organs are being extracted for donation. The in-between classifications correspond to progressive states of systemic diseases [60]. **Table 4.2** contains the detailed ASA six-level classification with the respective definition.

Table 4.2- American Society of Anesthesiologists classification of patient's physical status for surgical risk.

ASA Class	Definition
I	Normal healthy patient
II	Patient with a mild to moderate systemic disease
III	Patient with a severe systemic disease which is not incapacitating
IV	Patient with an incapacitating systemic disease that is a constant threat to life
V	Moribund patient who is not expected to survive 24h with or without operation
VI	Declared brain dead whose organs are being removed for donation

The World Health Organization (WHO) published the ICD, a widely used diagnostic tool that established codes for diagnosis and procedures relative to hospital services for epidemiology, health management, and clinical applications. The ICD was used to write the "COD_DIAGNOSTICO" and "COD_ACTO" variables in this dataset. This tool provides the collection, review, and comparison of mortality and morbidity data obtained in different periods and areas. It guarantees semantic data interoperability that promotes international comparability. There are different versions of ICD, based on the year they were developed. Since the first ICD, ICD-6 published in 1946, the WHO ensured the update and revision of this classification.

In the CHULC dataset, the versions of ICD were the ICD-9 from January 2017 until October 2020 (three years and ten months) and ICD-10 from October 2020 until December 2021 (one year and two months). For diagnosis codes, the ICD-9 clinical modification system has 13 000 codes, each composed of an alphabetic or numeric first digit followed only by numbers, with a minimum of three digits and a maximum of five. The first digits (1st-3rd) are relative to the disease category, and the remaining digits (4th -5th) indicate the etiology and the anatomic site of manifestations. The ICD-10 system is an expansion of ICD-9 (68 000 codes). It consists of three to seven alphanumeric digits, where the first three digits represent the category, the fourth to sixth digits the etiology, manifestation, and severity, and the seventh the extension [61].

Since ICD-10 has a novel structure that does not directly translate to ICD-9 codes one-to-one, different mapping methodologies based on the same ICD-9 codes can result in different sets of ICD-10 codes. The US Centers for Medicare & Medicaid Services established forward and backward General Equivalence Mappings (GEMs) to make mapping ICD-9 codes to ICD-10 codes easier. Although the

efforts of the Centers for Medicare & Medicaid Services in creating GEMs, their application is not simple and strict, and different approaches can result in different outcomes. As the name implies, the forward and backward GEMs are not mere mirror copies of one another. They are separated maps with vastly different scopes and coverage. The forward map does not include the bulk of ICD-10 codes, while the backward map does not include many ICD-9 codes, making it difficult to translate and connect them [62].

4.1.2 Current Methods

When defining the objectives of this dissertation, one of the first points was to understand the current methods and challenges of surgery schedule and OR time prediction on CHULC. To reach this goal, meetings were scheduled with the directors of specialties addressed in this project and anesthesiologists from CHULC. Besides these reunions that were a preponderant point in the conceptualization of the project, a visit in OR was made to understand the dynamics, environment, and the mapping process of a surgical procedure since the moment the patient arrive in the OR until he leaves the OR.

In general, in CHULC, quantitative tools to assist surgery schedules are uncommon, and the experience-based technique of programming surgery services prevails, leading to an inefficient OR workflow. Presently, this estimation is made by the primary surgeon based on their own experience, and the knowledge about the OR time regarding the past similar surgeries.

Regarding this task, the director of the Urology specialty points out three main factors that influence and complicate the OR time prediction: 1) In current methods, the surgeons only consider the surgeon's time and exclude the anesthesia time. 2) There is variability between professionals depending on the experience and the school they have learned. For the same type of procedure, two surgeons can present different surgical times. 3) There is also the variability of the own surgeon, which is harder to consider or predict. For the points 2) and 3), relative to the in-between and personal variability, requesting a surgeon to change their technique in order to homogenize this variability is expensive and represents quality risks for the hospital, which is not desirable. Hence, this variability must be taken into account when making OR time predictions.

The director of General Surgery also points out that the actual OR time prediction is mainly based on the surgeon's personal experience. Furthermore, he highlights the difficulty in standardizing some types of lengthier procedures regarding more complex patients. For the Orthopedics specialty, the director points out the high number of emergencies necessary for this specialty. Therefore, this will difficult the schedule of elective surgeries. Additionally, the typical profile of the patients admitted to the orthopedics surgeries are polymedicated and polydiagnosed elderly patients, which induces a high variability regarding both SCT and ACT, and, therefore, in OR time.

Like the surgical directors, the anesthesiologists related that the OR time prediction is only based on the surgeon's perspective. Moreover, they considered that the variability among anesthetists is minor compared to the surgeons and point out factors such as the patient's physiognomy, hypertension, and cardiac insufficiency as factors that can influence the ACT.

A common factor highlighted in all the professional's reunions was that the CHULC is a hospital school, and the intern's presence in surgeries leads to higher OR times. Apart from this, another salient fact in all specialties is the demand for individualization and specification in groups/units of procedures instead of treating the specialty as a whole. This will allow higher precision when discussing surgical factors and problems. When confronted about the idea of having a ML model that works as a decision-support tool to aid in OR time prediction, all agreed that this would be beneficial for their service.

However, they demonstrated their preference for having a specific-service model than a global one that includes all hospital specialties, since the specific-service model is more reliable and avoids some possible skepticism, in their opinion. This information was taken into account and was the basis for developing the specialty-level models instead of having a unique model for all specialties.

The described reunions were a preponderate point in this project, especially on the conceptualization and building phases. This allows the clinical surrounding from the starting point on this project, which was reflected by a better model understanding and conduct to less skepticism, and a better acceptance.

4.2 Data Preparation

This chapter addresses all tasks regarding the data preparation phase. It describes the sequence of all processes implemented in this project to transform raw data into legible data for the ML algorithm implementation and model development.

4.2.1 Time Variables Definition

The main goal of this dissertation is to predict the total OR time necessary for an individual patient. The CHULC database only contains the starting and finishing times of each task that compose the total OR time, as schematized in **Figure 1.3**. Hence, before any variable transformation, it is necessary to define the times that correspond to the preparation time, ACT, SCT, final time, and, most importantly, the OR time, the output of all models. Based on the variable's name from **Table 4.1**, each time was defined as follows:

$$\text{Preparation Time} = \text{Ini\_Aneste} - \text{Ini\_Sala} \tag{4.1}$$

$$\text{Anesthesia Time} = (\text{Ini\_Cirurg} - \text{Ini\_Aneste}) + (\text{Fim\_Anest} - \text{Fim\_Cirurg}) \tag{4.2}$$

$$\text{Surgeon Controlled Time} = \text{Fim\_Cirurg} - \text{Ini\_Cirurg} \tag{4.3}$$

$$\text{Final Time} = \text{Fim\_Sala} - \text{Fim\_Anest} \tag{4.4}$$

$$\text{Operating Room Time} = \text{Fim\_Sala} - \text{Ini\_Sala} \tag{4.5}$$

At the end of this task, five new variables were generated. The “Tempo_Prep”, based on **equation 4.1**, the “Tempo_Anestesia”, on **equation 4.2**, “Tempo_Cirurgia”, on **equation 4.3**, “Tempo_Final” on **equation 4.4**, and “Tempo_Sala”, the model’s output on **equation 4.5**.

4.2.2 Distribution of Operating Room Time

It is helpful to understand how the model's output, the OR time, is distributed when modeling. Strum et al. [56] proposed that the log-normal distribution fits better to modeling surgical procedures times when compared with a normal distribution, as reported in chapter 3. The logarithm of a log-normal distribution is normally distributed. For a log-distribution time in a surgical context, a small number of procedures may take substantially longer than average since it can handle values from zero to infinity

and has a big right tail. The initial idea to analyze if the CHULC data follows a log-normal distribution to then choose the suitable statistical tools to analyze data was first to apply a logarithm of data and then the Shapiro-Wilk test for the normality analysis.

The Shapiro-Wilk hypothesis test consists of a null hypothesis (H0) that the population is normally distributed and an alternative hypothesis (H1) that the population is not normally distributed. The null hypothesis is rejected if the p-value is smaller than the selected alpha level, indicating that the data examined are not normally distributed. To perform this test, the SciPy library was initially considered, however, as it is explained in the next paragraph, it was not adequate.

Statistical hypothesis tests have big power. The power of a statistical hypothesis is the probability of not accept the null hypothesis when it is false. It is given by **equation 4.6** in terms of type II errors. Type II errors occur when the null hypothesis is not rejected and is false, false negative.

$$\text{Power of a test} = 1 - \text{Probability of type II error} \tag{4.6}$$

Consequently, any minor difference between the data distribution and the null distribution (normal distribution) is significant and leads to the null hypothesis being rejected. The Shapiro-Wilk test's hypothesis from the SciPy library is sensitive when the number of samples (N) is bigger than 5000. If N>5000, the p-value may not be accurate, which is the case of the data under analysis.

For large sample sizes, the infringement of the normal assumption does not result in significant concerns, which implies that it is possible to use parametric procedures, even if the data is not normally distributed. For samples of hundreds of observations, the distribution of the data can be ignored. The central limit theorem states that regardless of the data distribution's nature, the distribution tends to be normal for big samples. Although pure normality is recognized as a myth, searching for it graphically is a common approach [63].

Taking into consideration that the data under analysis got thousands of samples, after the log transformation, the p-values of the Shapiro-Wilk hypothesis test for normality would not be accurate. Therefore, the data distribution was plotted, to analyze graphically, but the hypothesis test was not performed [63].

4.2.3 First Data Cleaning Phase

A substantial number of the initial 52 variables were redundant, i.e., two variables containing the same information since one variable includes the description of the hospital room or procedure amongst other information, and the other variable has the respective hospital code of that variable. For instance, the variable “DES_SALA” contains the description of the OR theater inside the block where the procedure will take place, and the variable “COD_SALA” contains the hospital code of that specific OR theater. Hence, these two variables contain the same information. Before continuing into the dataset analysis, when two variables were redundant, one of them was eliminated. This step removes non-essential variables that reduce accuracy and increase the model's complexity. The eliminated variable does not have specific elimination criteria because both are categorical variables, and the OHE method was selected to encode them. Thus, the variables are converted to binary numbers, therefore there is no difference between eliminate the variable that contains the hospital code or the one which contains the detailed description

Due to the purpose of this project, to predict the OR time for elective surgeries, emergency surgical cases were eliminated since they are not possible to schedule in advance. Additionally, those procedures performed in less than ten instances during the five years under study were excluded because they do not represent the bulk of the surgeries. Due to the lower number of these surgeries, above the defined threshold, in the training phase, the model might not learn the data patterns and not generalize well in the test phase. Consequently, these procedures can represent noisy data for the model, which is not desirable and, for that reason, were eliminated. For the General Surgery Model, this number was twenty rather than ten since this is an extensive specialty with many different types of procedures across different body areas compared to other specialties. The definition of this threshold to exclude the surgeries was discussed with the respective specialty directors. Finally, all cases with negative or less than ten minutes of OR time were excluded. The described framework allows the exclusion of some outlier points.

4.2.4 Missing Values

In the theoretical background chapter, the question of missing values was introduced as well as some common approaches to handle it. One of the most important considerations when deciding which imputation approach to applying is to obtain the best effective value for the missing variables. This section details the specific methods selected to handle missing values on CHULC dataset variables.

For the age variable, “IDADE”, two central tendency measures, the median, and the mean, were considered to impute the missing values in the first instance. The selection was based on data distribution. If the data distribution is symmetric, the mean value is selected to replace the missing values. Otherwise, if data is skewed distributed, the median value is selected. The mode was not considered for this step since age is a continuous variable. After the data distribution analysis, the median value was selected to impute missing values for the age variable since the data was not symmetric.

The number of missing values of “DIAS_EM_LIC” and “PRIORIDADE” variables were linked since the number of missing values of both variables corresponded to the same entry, i.e., the same person's operation. After a thorough examination, it was possible to notice that these operations correspond to situations where a reoperation took place in less than 24 hours. Hence, the missing values of the “DIAS_EM_LIC” variable, corresponding to days on the waiting list, were placed as 0, and the “PRIORIDADE” variable, corresponding to surgical priority level, was replaced with the maximum level, “4 Urgencia Diferida”.

The “DES_DESTINO” variable contains the information about the patient's destination after the surgery. It is not possible to infer this variable's missing values based on other variables. Since this is a categorical variable, the mode or searching for similar cases would be a possibility to fill the missing values. However, there are different possible levels of categories, and to not induce bias in the model by adding information that might not correspond to reality, a level for missing values was created designated as “Not defined”.

Similarly, for the health professional’s ID variables, “ANEST_PRINC”, and “CIRURG_PRINC”, for missing values, a new level for missing values was created as “Not defined”.

For the anesthesia variable, “COD_R_ANEST”, the missing values depend on patient conditions and health habits, such as BMI, systemic diseases, smoking habits... Consequently, predicting or filling these values based on other CHULC dataset variables or by looking into similar cases is not accurate. Since anesthetic variables are highly individual and valuable for the model’s prediction from

the anesthetic point of view, they cannot be eliminated or predicted based on another dataset information. One more time, a new level for missing values was created as “Not defined”.

In the case of the “ACTO_ROBOT” variable, the missing values do not indicate missing information. They refer to surgeries where the robotic procedure was not applied. Hence, the missing values were filled with “Not Applicable”.

When time-related variables, except for “Ini_Sala”, and “Fim_Sala”, which are the variables that compose the OR time, the output of the model, presented missing values was not considered a relevant concern since this information of the specific times that compose the total OR time is not included in any of the models, only the total OR time. However, these times are necessary to analyze the impact of SCT and ACT times in OR time. For this parallel analysis, the surgeries with missing values regarding “Ini_Aneste”, “Ini_Cirurg”, “Fim_Cirurg”, “Fim_Anest” were eliminated since, without these values, the SCT and ACT, as well as preparation and final time cannot be defined, according to **equations 4.1** to **4.4**. The “Prep_Sala_Ini” and “Prep_Sala_Fim” time-related variables are included in the turnover time, which is not included in OR time, and, therefore, unnecessary for the described analysis, so they were eliminated.

Lastly, variables with a percentage of missing values upper than 70%, except for the “ACTO_ROBOT” variable, where missing values correspond to non-applied RAS, were eliminated. The remaining variables that were not addressed in this subchapter do not present missing values, therefore the approach to deal with them was not required.

4.2.5 Categorical Variables

Based on **Table 4.1**, it is possible to notice that most of the variables are categorical variables. As mentioned in section 2.1 in regression analysis, categorical variables require special care since they cannot be inserted into the regression equation, unlike continuous variables. Instead, they must be encoded into a set of variables before being included in the regression model. The OHE method was selected to encode these types of variables, since it ensures independence between variables. However, for features with high cardinality, i.e., attributes that presented too many unique values, OHE became a problem since there is a different column for each unique value (showing its presence or absence) in the categorical variable. Consequently, this causes two issues. The first is space consumption, and the second is the curse of dimensionality. As the number of levels of a categorical feature grows, the proportion of data that the model needs to distinguish and learn and generalize it from increases exponentially [71]. Hence, high dimensional data showed to be prejudicial when analyzing the data to detect patterns, and during the ML training phase.

The variables in the dataset that presented a high level of cardinality were: “COD_DIAGNOSTICO”, “COD_ACTO”, “CIRURG_PRINC”, and “ANEST_PRINC”, as possible to notice by the number of levels of these variables in **Table 4.1**. The following paragraphs focus on the implementation approaches to reduce the cardinality of these variables.

The “CIRURG_PRINC”, and “ANEST_PRINC” variables contain the information about the primary surgeon and the anesthetist ID, respectively. The leading information of these two variables for the model is whether the specific professional tends to do the surgical procedure more quickly or slowly.

There is variability between professionals inside the same specialty and even for similar procedures. This happens since there is a learning curve for all the professionals and depending on their years of experience and the medical school they frequented, there are techniques that are done differently, consequently, this will influence the OR time. Even for surgeons with extensive years of experience, some techniques have changed. The development of new technologies such as RAS will change all professionals' learning curves, independent of their years of experience. With this in mind, these two categorical variables were converted into four continuous variables. Rather than having the primary surgeon ID, the average time of both surgeons and anesthetists' specialists were calculated and presented by the variable names: "Cirur_media" and "Anest_media", respectively. Additionally, the surgeon's mean of the surgical time for the last surgeries of that specific procedure was implemented, the "Cirur_Ato" variable. This variable was added since the mean will not accurately capture the information for a surgeon who performs various types of surgeries with very different surgical times. In these cases, there is heterogeneity in the surgical times. For instance, for a surgeon who performs mainly three types of procedures with a considerable amplitude separating these surgical times, the mean will not be a precise metric to measure and capture these patterns. Similarly, for "ANEST_PRINC", instead of having the professional ID, the mean of the anesthesia time based on the type of anesthesia induced in the patient was considered, the "Anestesista_tipo" variable. These new metrics were validated by the medical directors and anesthesiologists of the approached specialties.

The "COD_DIAGNOSTICO" is relative to the diagnosis code for the specific disease based on ICD-9 or ICD-10 codes. This variable presented the highest number of levels, hence the highest cardinality. However, the number of levels was not homogeneously distributed among the three specialties. The General Surgery specialty presented 1168 unique diagnosis codes, Orthopedics 1657, and Urology 285.

Due to the high levels of "COD_DIAGNOSTICO", the group of diagnoses was added for All Specialties, General Surgery and Orthopedics specialties instead of the specific diagnosis code. The ICD diagnosis codes are divided and classified according to the type of disease in larger groups. For instance, the Neoplasms category encompasses all the ICD-9 diagnoses encoded with numbers between 140 to 239, and for ICD-10, the codes between C00-D49. **Table A1** in the Annex describes all the ICD groups of diagnoses descriptions with the corresponding ICD-9 and ICD-10 diagnosis codes. This step allows reducing the variable levels to 27 and, consequently, cardinality reduction was observed.

Since Urology was the specialty with lower variable cardinality, the "COD_DIAGNOSTICO" variable was maintained in this specialty-specific model. Although this was the specialty with fewer categorical levels, there were still a considerable number of levels. To reduce them, the similar diagnosis codes were grouped based on GEMs and on the structure of the code, i.e., the codes that were relative to the same type of condition in the same body anatomy. The medical opinion was also preponderating in this phase. To support this step, the Kruskal-Wallis statistical test was applied. Since the OR time distribution for the same diagnosis code does not present a normal distribution, a non-parametric test was selected. This is a non-parametric test, equivalent to the parametric one-way ANOVA, to determine if samples belong to the same population. Hence, for the similar diagnosis codes, based on the respective OR time distribution, the hypothesis for this test was:

H0: The samples belong to the same population (i.e., population medians are equal)

H1: At least one sample does not belong to the same population (i.e., population medians are different)

If the null hypothesis was accepted, with a significance level of 0.05, the diagnosis codes are similar, corresponding to a very approximated diagnosis, and result in an approximated OR time distribution, hence can be grouped. Otherwise, when the null hypothesis was rejected, the diagnosis were not grouped. The same line of thinking was not applied in the General Surgery and Orthopedics specialties since applying the described steps in these specialties will not significantly reduce cardinality, as tested.

The "COD_ACTO" variable contains the information about the type of procedure required for the patient. This is a key identifier for the surgery. Hence this metric must be maintained in its original form. For this reason, unlike the "COD_DIAGNOSTICO" variable, the codes relative to the procedure were not grouped. Nevertheless, implementing the filter for excluding all the procedures that were not executed at least ten times in the previous five years and twenty times for the General Surgery specialty allows reducing the cardinality of the variable by only including the most common surgical procedures.

The remaining categorical variables were maintained, and the OHE technique was applied for all categorical variables.

4.2.6 Feature Creation

As discussed in the theoretical introduction, creating new features aims to obtain variables with more predictive value. Over the methods section, the critical thinking behind the new feature creation is sequentially explained. In **Table 4.1**, these features are colored blue, and they are associated with the diagnosis codes group, "GRUPO_DIAGNOSTICO", to the surgeons' and anesthetists' metrics, "Cirur_media", "Anest_media", "Cirur_Ato", "Anestesista_tipo", and time definition, "Tempo_Sala", "Tempo_Prep", "Tempo_Cirurgia", "Tempo_Anestesia", and "Tempo_Final". Additionally, a new variable regarding the type of operation was added, as proposed by a CHULC specialist, the "Cod_media", which represents the mean of the OR time for the specific procedure.

4.3 Feature Selection

After a thorough literature review, presented in chapter 3, along with consultations with clinicians and hospital administrators of CHULC, potential predictors for the regression models were identified from the CHULC database, presented by the green and blue without * colors.

In addition to the initial variables presented on the CHULC dataset, in the feature selection process, the collinearity between the new continuous variables regarding the surgical team and procedure were analyzed. These variables were: "Cirur_media", "Anest_media", "Cirur_Ato", "Anestesista_tipo", and "Cod_media".

To analyze the collinearity between them the correlation matrix with the Pearson's coefficient was displayed. If two variables present a Pearson's Coefficient higher than 0.600, this indicates that collinearity most likely to exist, therefore one of them must be eliminated. The elimination criteria was based on two factors. First, the variable containing most of the information should be maintained. Second, the variable with the highest correlation with the model output, traduced by the highest Pearson's coefficient with the OR time, should be maintained.

At the end of this section, all data preparation and feature selection methods were exposed, and the dataset was prepared for the ML implementation.

4.4 Impact of SCT and ACT times in OR time

When defining this dissertation's objectives, one of them was to evaluate the impact of both SCT and ACT times in OR time. This analysis aims to support the necessity to include all OR time parcels and not only to focus on the SCT when predicting the OR time for the patient, as the currently practiced methods. By looking into all the OR time parcels, it is possible to define the OR time block necessary for an individual patient with high accuracy and improve the OR schedule.

For this analysis, besides the core of OR time, i.e., the SCT and the ACT, the preparation and final time were also considered. Although these times have a minor impact on OR time, it is important to analyze them since they provide insights about the current hospital's time in processes that must be performed, but are not part of the pivotal moment in the OR as the operation itself and the anesthesia. Quantifying these processes will provide an analysis tool that will allow the discussion to reduce these non-essential times and optimize OR time.

To analyze and measure the relationship' strength between these times and the total OR time, first, the time variables were plotted to analyze if there is a linear relation between them and the OR time. After this, the correlation was analyzed by the correlation matrix. Pearson's correlation coefficient was the selected metric for the correlation analysis. If this coefficient value is higher than 0.600, the correspondent time variable was considered as highly correlated with the OR time.

4.5 Model Development

Before the algorithm implementation, a fundamental step is the exploratory data analysis to understand the variable distribution and the relationship with the model output. For this task, statistical measures to summarize the principal data characteristics and plot the graphics provided the visual relationship between them and supported the knowledge about data distribution. The exploratory data analysis enables the characterization of the data and the information for developing predictive models. For exploratory data analysis, the data distribution, and the relation between variables, particularly with the output variable, the OR time, were plotted. The most relevant analysis are presented in the results section.

4.5.1 Model Selection

In this project, the model selection was based on the literature review studies that apply ML to predict OR times. In Chapter 3, particularly by examining **Table 3.1**, it is possible to conclude that linear

and DT-based models are the most selected models by the authors to solve this type of problem, as also concluded by the end of that section.

Moreover, it was also considered the interpretability of the model and the black-box model problem regarding clinical applications. A black-box model is a model that only captures the functional behavior between the inputs and outputs of the system. Therefore, it is not possible to comprehend how variables are integrated to then create predictions. While the black box systems may not be a concern in other industries, they create healthcare risks and consequences that require conscious attention. The opacity generated by these systems makes the coding harder to access and algorithms that internalize data in ways that are difficult for humans to audit or understand. This leads to a lack of transparency that both clinicians and patients require. Models with direct clinicians' applications should prioritize transparency to ensure more trust in the model and eliminate possible skepticism. By adopting more transparent models, it is also possible to analyze which variables have more impact on the system, which is a critical point for healthcare problems [64].

Considering these two points, two different types of models were selected. For the linear regression approach, a MLR model was selected because it is a common prediction approach that might provide a good comprehension of variable relationships. For the DT-based model, an RF model, due to its capacity to handle many predictors efficiently and its consistent performance across a range of ML tasks. Although most authors treated the RF as a black-box model, as approached in chapter 2, a forest comprises a vast number of deep trees, each of which is trained on bagged data with a random selection of characteristics, it is possible to comprehend the decision process by studying each tree thoroughly. Additionally, computing feature importance is one technique to gain insight into an RF.

4.5.2 Model Description

The theoretical fundamentals of both MLR and RF were presented in section 2.2. For both models, first, the global CHULC dataset was applied, i.e., the dataset containing all the three specialties surgeries, the first approach, All Specialties Models. Then, this dataset was divided into three subsets, one for each specialty, the second approach, Urology, General Surgery, and Orthopedics Models. This results in eight different models corresponding to the two approaches for the two ML algorithms (RF and MLR):

1. All Specialties Models (RF and MLR models)
2. Urology Model (RF and MLR models)
3. General Surgery Model (RF and MLR models)
4. Orthopedics Model (RF and MLR models)

In data modeling, the process of training a predetermined algorithm to predict the values from the selected variables when applied to new data, the dataset was split into train and test sets, 80% for train and 20% for test.

The initial features used for both models are colored in green and blue in **Table 4.1**. The scikit-learn, a free machine learning library for Python, was used for the data modeling process. For both RF and MLR algorithms, the default parameters were initially implemented. For RF, the default parameters correspond to a number of trees of 100, and the minimum number of samples required to split an internal node of 2.

Due to the shape of data distribution, some authors point out that the logarithmic transformation on the output variable, OR time, would perform better. Since this was not a transversal practice in all studies, it was tested, and if this step does not improve evaluation metrics (section 4.6), the logarithmic transformation will not be considered for analysis.

4.5.3 Model Tuning

The model tuning corresponds to the phase where the optimal hyperparameters are found to maximize the models' predictive accuracy and reduce the prediction errors. The model tuning was only applied in the individual specialty models since the specialty professionals from all the services preferred a robust and detailed model for their department instead of a general model that englobes many different procedures. The key focus from this point was to focus only on the specialty's models, reduce the error for these models and enhance their performance at a specialty level, as desired by the stakeholders.

This phase was applied to the RF model since it was initialized with predefined settings. The goal is to find the optimal values for the number of trees, the maximum depth of the tree, and the minimum number of samples, amongst other settings. First, the random search and then the grid search, both along with k-fold CV, were applied for this process in each trial.

Firstly, the random search was applied to understand the range of optimal values for the hyperparameters. This step gave a hint about the ideal possible values for the hyperparameters. Then, based on the result values of the random search, the grid values were defined, and the grid search method was applied. The k-fold CV was used with a k value of 5. Although other k values can be used, this was the suitable value based on training times and the literature. The grid search was performed until no more significant improvements to reduce the MAE were found.

This was an interactive process where the model with the best set of hyperparameters was selected based on the lower MAE. Although the MAE was selected as the prime selection parameter, the run time was also registered. It is essential to look at the run time, as when making predictions on new data, and considering the high volume of data, it is desirable for the algorithm to run in a short period of time. Ideally, in this project, the dataset would be updated with a given periodicity to improve the model's performance and update the dataset, which would mean a higher volume of data. Consequently, the run time will be an even more relevant aspect as the data volume increases.

4.6 Evaluation Metrics

For both models, the selected evaluation metrics were the R-squared, RMSE, and MAE. These are the three most common metrics for regression models evaluation. As reported in **Table 3.1**, it was also the common metrics used by the authors to evaluate the models in the literature. The R-squared was obtained in both training and test phases. Although the R-squared of the test phase is more important, it is relevant to compare the R-squared from the test phase with the R-squared of the training phase to ensure that the model does not over or under-fit. A high R-squared on the train compared with the R-squared on the test means that the model does not generalize well. If the R-squared test is much higher than the training phase, it might indicate that the train/test split was inadequate for the problem, and CV is needed.

The MAPE was not included as an evaluation metric in this project since this is an asymmetric measure that gives a higher weight to negative errors and, consequently, induces bias in the model evaluation. Once negative and positive errors are not desirable in equal weight, this metric was not included.

Besides the described evaluation metrics, the percentage of underestimated, overestimated, and within test samples were also used as evaluation metrics, since these metrics are also reported as evaluation metrics in the literature. It was determined with a 10% tolerance threshold and defined by **equation 4.7** to **4.9**. The equations and the 10% threshold were defined based on the literature.

Underestimated Cases: Real OR Time > Predicted OR Time + 10% tolerance threshold (4.7)

Overestimated Cases: Real OR Time < Predicted OR Time - 10% tolerance threshold (4.8)

Within Cases: Predicted OR Time - 10% tolerance threshold < Real OR Time < Predicted OR Time + 10% tolerance threshold (4.9)

4.7 Feature Importance and Significance

Lastly, the feature importance was analyzed. This provides a score to input features depending on their value in predicting the target variable. This is an important task to ensure the transparency of the models and to improve the efficiency and efficacy of a predictive model through dimensionality reduction and feature selection.

In the MLR model, the feature significance was first analyzed based on the p-value by the t-statistic test. The process consists in comparing the p-value for each term in the model to the significance threshold previously defined, 0.05, to evaluate whether the relationship between the output and every single feature in the model is statistically significant. The null hypothesis is that the term's coefficient (β) equals zero, indicating that the term and the response have no relationship. The hypothesis test is defined as follows:

H0: $\beta = 0$ There is no relation between the feature and the output

H1: $\beta \neq 0$ There is a relation between the feature and the output

With a significance threshold of 0.05, there is a 5% chance of a relationship existence while there is any. If the p-value is less than or equal to 0.05, the null hypothesis is not accepted, and it is possible to infer that the output variable and the feature have a statistically significant relationship.

For categorical variables, it is not possible to add them up in order to get the overall significance of the variable. One method to determine the significance of categorical variables in a linear model is when there is a significant t-test at one level of the categorical variable, assuming that the overall significance will also be impactful. After this analysis, the variables that were not statistically significant for the MLR model were eliminated for this model, making the model less complex since only the significant variables were included and achieving better run times.

For RF, the feature scores were displayed along with the feature weight, based on a function of the scikit-learn library, and the first five features were considered the most impactful ones for the RF model. However, it is crucial to have a critical evaluation when analyzing the importance of the RF features; due to the high levels of the categorical features.

4.8 Comparison with Current Methods

After the model development phase, the final goal was to compare the model predictions with the current methods in the test set. The CHULC database contains a variable called "Duracao_Prevista_no_Agendamento", the variable that contains the OR time predicted by the surgeon. This variable was used as a comparative term for the analysis. First, the plot containing the predicted time by the surgeon versus the actual OR time was displayed. Based on **equations 4.7, 4.8,** and **4.9**, the labels of underestimated, overestimated, and within percentages were attributed and compared with the model's percentages. To estimate the surgeon's prediction errors, based on **equations 2.9** and **2.10**, the MAE and RMSE were calculated for the surgeon's predictions, where $\hat{y}_i$, represents the OR time predicted by the surgeon, y_i, the actual value of the OR time and n the number of surgeries.

5. Results

This chapter aims to present the relevant results of this project. First, the data preparation phase and exploratory data analysis were presented as well as the impact of the ACT, SCT, preparation, and final times in the OR time. The bulk of this chapter is constituted by the MLR and RF model results. The results are divided into the two approaches initially proposed for all specialties and a single specialty, with a particular focus at a specialty level. Finally, the model's feature importance was analyzed and compared with the current CHULC methods.

5.1 Data Preparation

The initial data was composed of the operating data for five years, from January 2017 to December 2020. Based on exclusion criteria presented in section 4.2.3. from the initial dataset, containing 48737 surgeries, 9025 were eliminated since they represent emergency cases, 7018 with negative OR time, and 5643 that do not present the minimum number of surgeries, 10 in the case of Orthopedics and Urology and 20 for General Surgery. This resulted in a total of 27051 surgeries for the model development, with the division by specialty: 11885 surgeries for General Surgery, 5210 for Urology, and 9956 for Orthopedics.

From the initial 52 variables, 13 were eliminated based on the redundancy criteria, represented by the red color in **Table 4.1**. This allowed dimensionality reduction of the initial dataset without losing relevant information. Additionally, based on the criteria of presenting more than 70% of missing values, five variables were eliminated, the “Prep_Sala_Ini”, “Prep_Sala_Fim”, “Pernoita”, “Alta”, and “ANEST_ADICIONAL” variables.

For the Urology Model, the patient diagnosis categorical variable, “COD_DIAGNOSTICO” reduced from the initial 285 levels to 62, when the minimum number of surgeries filter was applied, and the steps explained in section 4.2.5 were implemented. This variable is denoted by (**) in **Table 4.1** since it was only maintained in the Urology Model.

Using the “GRUPO_DIAGNOSTICO” variable instead of “COD_DIAGNOSTICO” in the All Specialties, General Surgery, and Orthopedics Models resulted in a reduction into 27, 22, and 15 levels, respectively.

5.2 Variable Selection

The predictors for all the regression models were identified based on the literature review and consultations with the experts, as explained in chapter 4. The set of variables selected for the models is presented in **Table 4.1** according to the color code, by the blue and green colors.

The feature marked with (*) contains the information about the specialty in which the surgical procedure is included, this was only relevant for the All Specialty Models, since all surgeries from different specialties entries are mixed, therefore it was only included in this model. The feature marked with (**) is relative to the diagnosis code, which was only included in the Urology Model, as previously explained. The feature relative to RAS is marked with (****), and it was not included in the Orthopedics Model since this specialty does not perform RAS in the period under analysis. However, if more data is included where RAS for Orthopedic specialty is performed, this variable must be included in this model.

Finally, features marked with (***) were removed after the collinearity analysis, and the ones marked with (*****) were only included in the impact of SCT and ACT in OR time analysis.

After displaying the correlation matrix, with the Pearson's coefficient analysis for the new continuous variables generated, **Table 5.1**, three variables, the "Cod_media", "Cirur_media" and "Anest_media", from the set of variables selected based on the described process in section 4.3, were eliminated since the Pearson's coefficient was higher than the threshold defined in the Methods section of 0.600. The variable "Cod_media", the mean of the surgical procedure, and "Cirur_media", the surgeon mean time, and the "Cirur_Ato", the mean of the surgeon for that surgical procedure, presented a Pearson's coefficient of 0.902 and 0.683, respectively. Therefore, they presented high collinearity, and one of them must be eliminated. Considering that "Cirur_Ato" contains additional information on the mean of the surgical procedure time, once it presents the mean based on the professional and type of operation, this variable was maintained and the "Cod_media" and "Cirur_media" were eliminated. The "Anest_media", mean of the anesthetist time, and "Anestesista_tipo", mean of the anesthetist time depending on the type of anesthesia, presented a Pearson's Coefficient of 0.878, also indicating high collinearity. With the same line of reasoning, the "Anestesista_tipo" was maintained since it contained additional information regarding the type of anesthesia that influences the mean of the anesthetist, and "Anest_media" was eliminated. This allows us to reduce the dimensionality and maintain the variable with less information loss and with the highest correlation coefficient with the output, "Tempo_Sala", as it confirmed in **Table 5.1**. The maintained variable "Cirur_Ato" presented a Pearson's Coefficient with the output, "Tempo_Sala", of 0.865, whereas the variable "Cod_media" had a coefficient of 0.808, and "Cirur_media" of 0.602. Similarly, the variable "Anestesista_tipo" presented a coefficient of 0.535, whereas the eliminated variable, "Anest_media", had a coefficient of 0.475. The eliminated variables after this analysis are noted with (***) in **Table 4.1**, and the remaining variables, "Cirur_Ato" and "Anestesista_tipo", were maintained.

Table 5.1- Correlation Matrix for the new generated variables. The dark color indicates a high Pearson's Coefficient above the defined threshold of 0.600 indicating that one of the variables must be deleted due to the high collinearity.

	Cod_media	**Cirur_media**	**Anest_media**	**Cirur_Ato**	**Anestesista_tipo**	**Tempo_Sala**
Cod_media	1.00	0.638	0.450	0.902	0.572	0.808
Cirur_media	0.638	1.00	0.350	0.683	0.391	0.602
Anest_media	0.450	0.350	1.00	0.427	0.878	0.475
Cirur_Ato	0.902	0.683	0.427	1.00	0.498	0.865
Anestesista_tipo	0.572	0.391	0.878	0.498	1.00	0.535
Tempo_Sala	0.808	0.602	0.475	0.865	0.535	1.00

5.3 Impact of Anesthesia, Surgeon-controlled, Preparation and Final Times in Operating Room Time

The study of the impact of ACT, SCT, and preparation and final times in OR time was defined as a parallel study of the ML model development to confirm the importance of the subject of predicting the total OR time by looking at the total OR time and not only into the SCT.

Graphics that describe the relationship between the SCT and ACT with the output variable, OR time, were plotted since these were the times that most influenced the OR total time, **Figures 5.1** and **5.2**. **Table 5.2** is the correlation matrix, with the corresponding Pearson's coefficient describing the variable's relationship for all time variables under study. Pearson's coefficient above the defined threshold, 0.600, is colored green. It is notable in **Figure 5.1** the linear relation between the SCT and OR time that is then reflected in **Table 5.2** by the very high Pearson's coefficient of 0.996. The ACT presented a Pearson Coefficient of 0.686, also indicating a high correlation with the output.

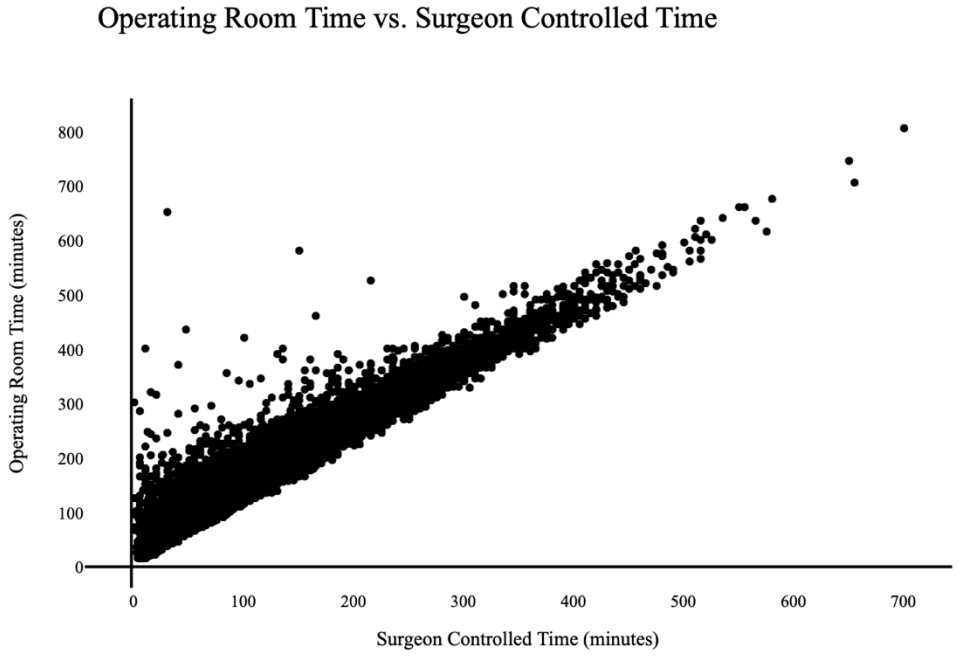

Figure 5.1- Relationship between Operating Room time and Surgeon Controlled time.

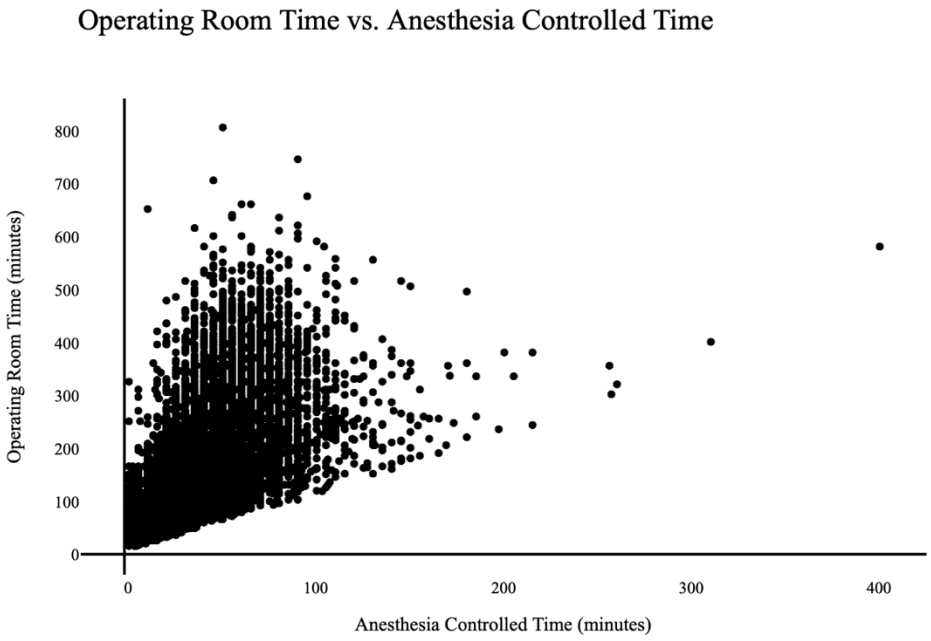

Figure 5.2 - Relationship between Operating Room time and Anesthesia Controlled time.

Table 5.2- Correlation Matrix with Pearson's Coefficient between the times that compose the Operating Room time and the Operating Room time. The green color indicates a Pearson's Coefficient between the variables above the defined threshold of 0.600.

	Preparation Time	**Surgeon Controlled Time (SCT)**	**Anesthesia Controlled Time (ACT)**	**Operating Room Time (OR time)**	**Final Time**
Preparation Time	1.00	0.132	0.107	0.234	0.0372
Surgeon Controlled Time (SCT)	0.132	1.00	0.526	0.966	0.0745
Anesthesia Controlled Time (ACT)	0.107	0.526	1.00	0.686	0.0566
Operating Room Time (OR time)	0.234	0.966	0.686	1.00	0.191
Final Time	0.037	0.0745	0.0566	0.191	1.00

5.4 Exploratory Data Analysis

The Exploratory Data Analysis subsection aims to present the data visualization, summarized statistics, and the interaction between the inputs and the output feature, the OR time.

Figures 5.3, **5.4**, **5.5** and **5.6** presents the response/output variable distribution, the OR time for all specialties, and at the individual specialty perspective, accompanied by the median and interquartile range (noted as IQR in the figures). There was reasonable variation among specialties, but in all distributions, there was a positively skewed tail. Although any statistical test was performed in order to analyze the data distribution, as explained in section 4.2.2, by the data distribution graphics in **Figures 5.3** to **5.6**, it is possible to identify some peaks. The median surgical time for all specialties is 115.0 minutes (≅ 1.9 hours), with 99.4 minutes for the Urology specialty, 140.3 minutes for General Surgery, and 137.0 minutes for Orthopedics.

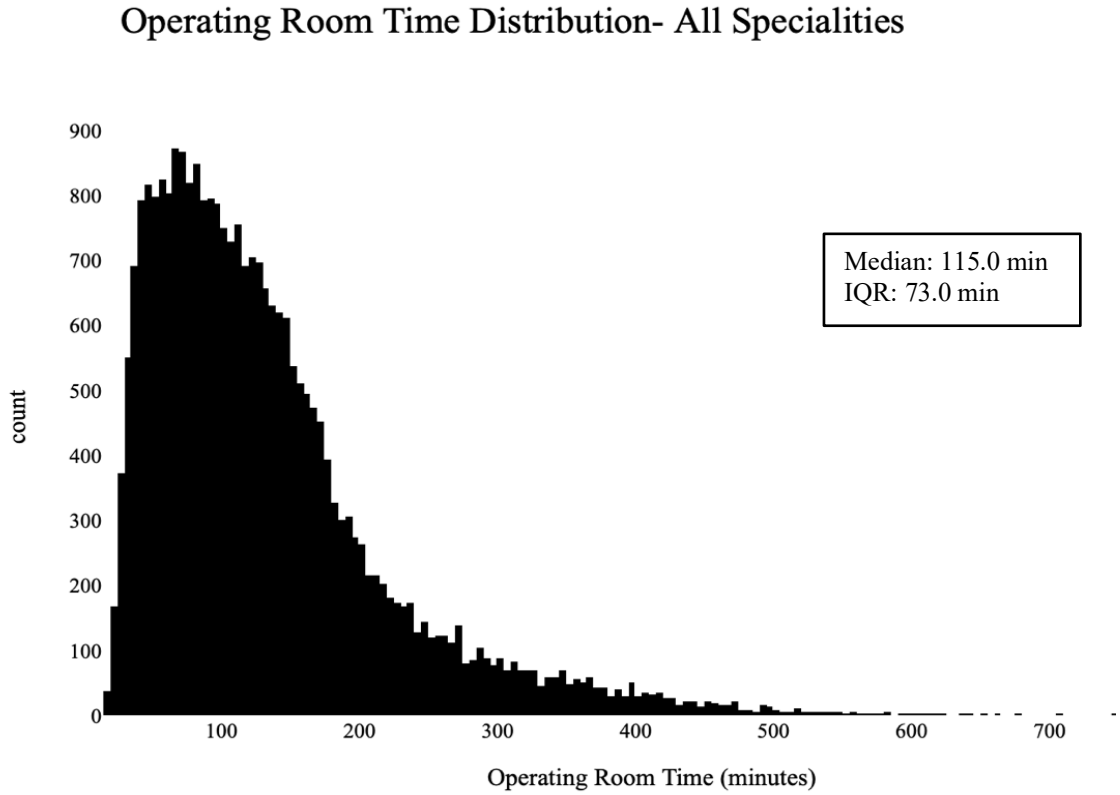

Figure 5.3-Distribution of Operating Room time for All Specialties.

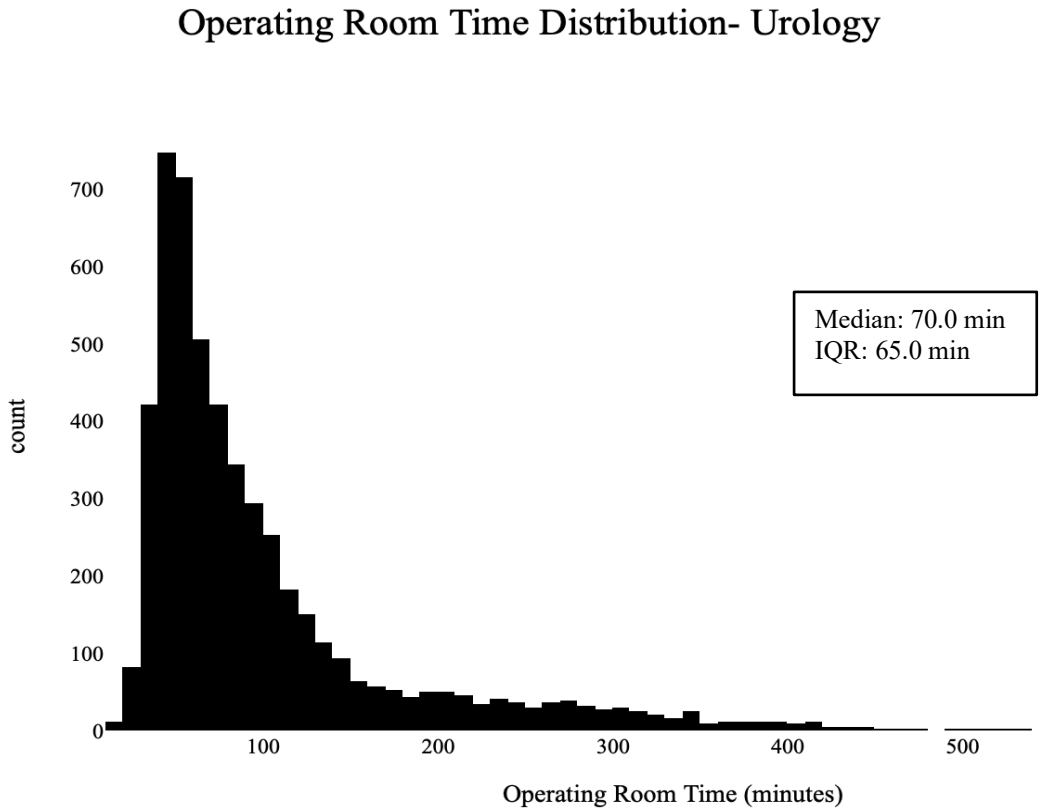

Figure 5.4- Distribution of Operating Room time for Urology specialty.

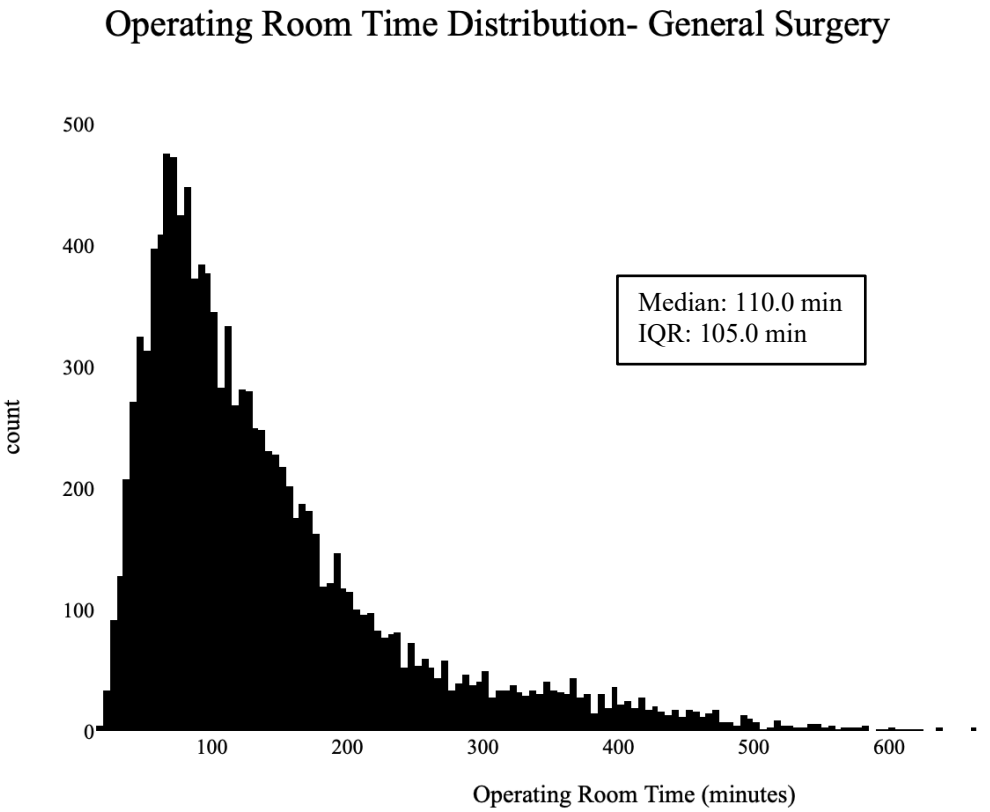

Figure 5.5- Distribution of Operating Room time for General Surgery specialty.

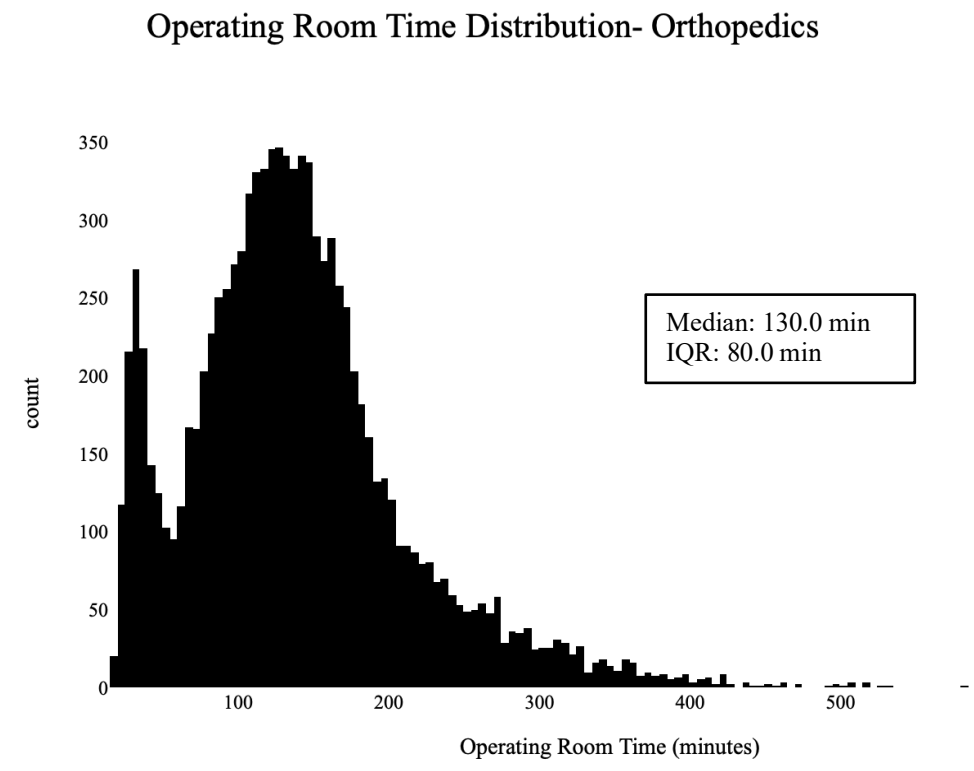

Figure 5.6 - Distribution of Operating Room time for Orthopedics specialty.

After obtaining the data distribution, some analysis regarding the relationship between the input and output feature was plotted and a few of them are presented in this section. Starting with continuous variables, **Figure 5.7** describes the relationship between the days on the waiting list, X-axis, and the OR time, Y-axis. In general, shorter days on the waiting list correspond, most of the time, to more urgent patient status, leading to higher OR times when compared to higher days on the waiting list.

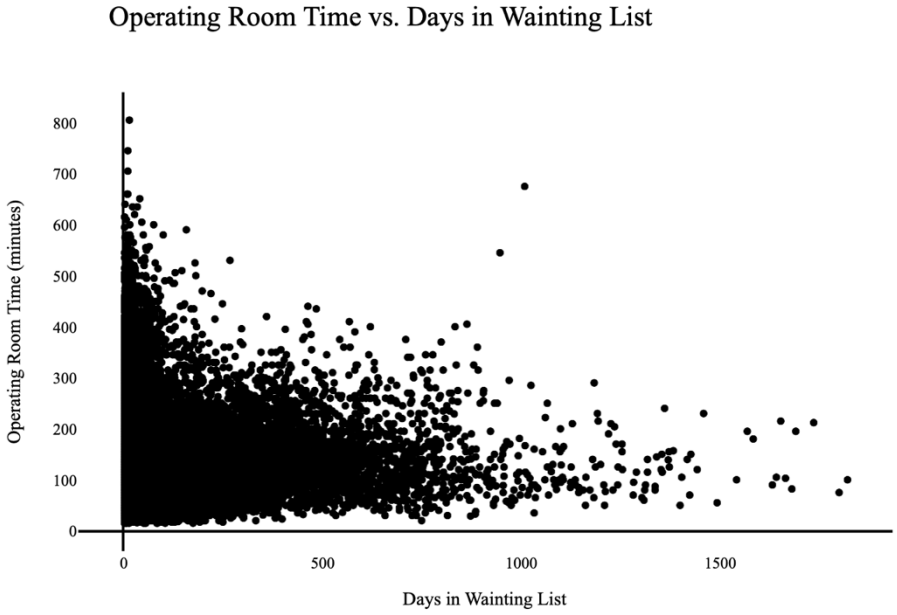

Figure 5.7 - Relationship between the Days in Waiting List and Operating Room time.

Figure 5.8 presents the plot of the surgeon's mean based on procedure code, a new variable that was generated, "Cirur_Ato", with the OR time. It is possible to observe that the variables presented a strong linear relation, also indicated by the high Pearson Coefficient, of 0.865, **Table 5.1**. This is beneficial, particularly for the MLR model since this new variable presents a higher linear relation with the model output.

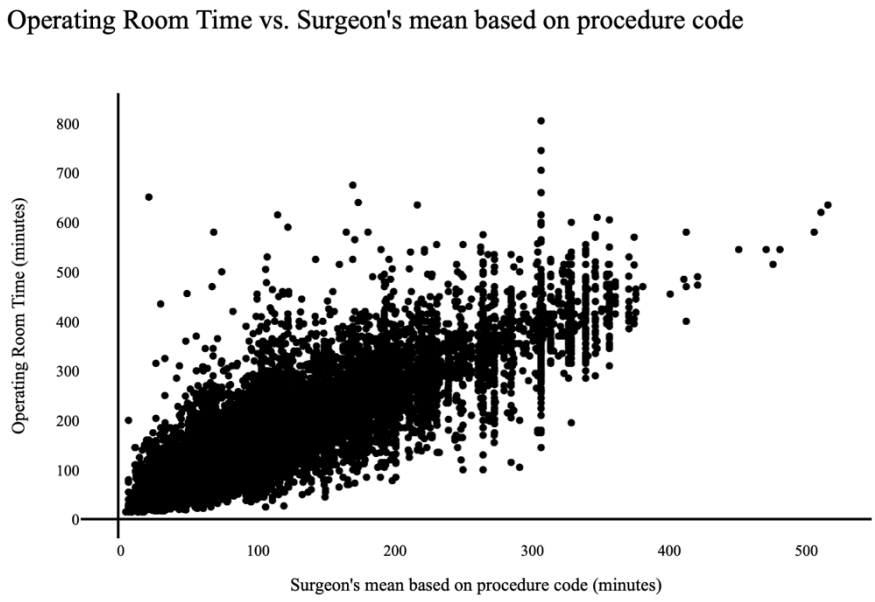

Figure 5.8- Relationship between the Surgeon's mean based on procedure and Operating Room time.

Figure 5.9 contains a box plot relative to the ACT distribution depending on the ASA classification of the patient. Excluding ASA 5, the higher the risk presented in the CHULC database, the higher patient's ACT will be, which leads to higher OR times.

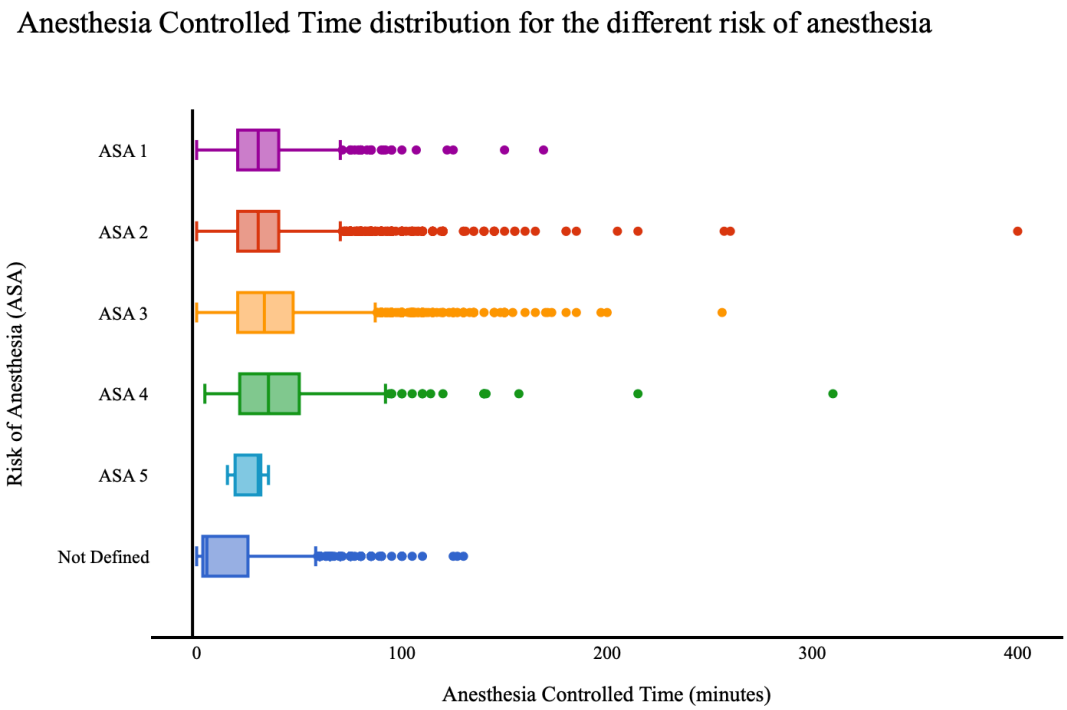

Figure 5.9 - Anesthesia Controlled Time distribution for the different classification of anesthesia risk based on American Society of Anesthesiologists.

Figure 5.10 addresses the OR time distribution across different types of interventions. This figure suggests that the basic type of scheduling conduct to a higher OR time.

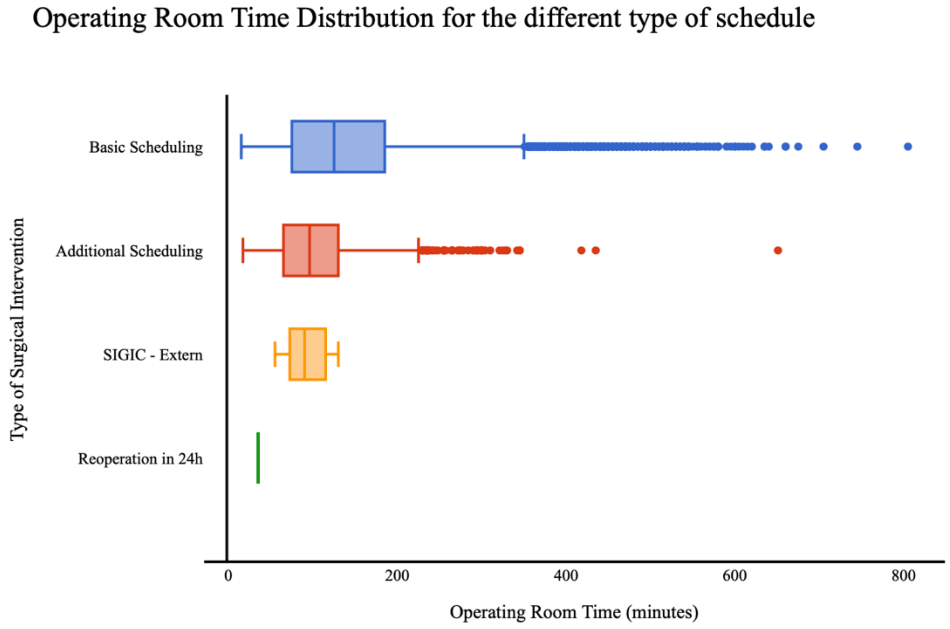

Figure 5.10- Operating Room time distribution of different type of schedule.

The last analysis in this section is represented in **Figure 5.11** for ambulatory and not ambulatory surgeries. Broadly, ambulatory surgeries (red color), i.e., surgeries where the patient does not stay overnight in the hospital (outpatient), lead to lower OR times, conversely to the non-ambulatory surgeries (inpatient) where the patient stays overnight in the hospital.

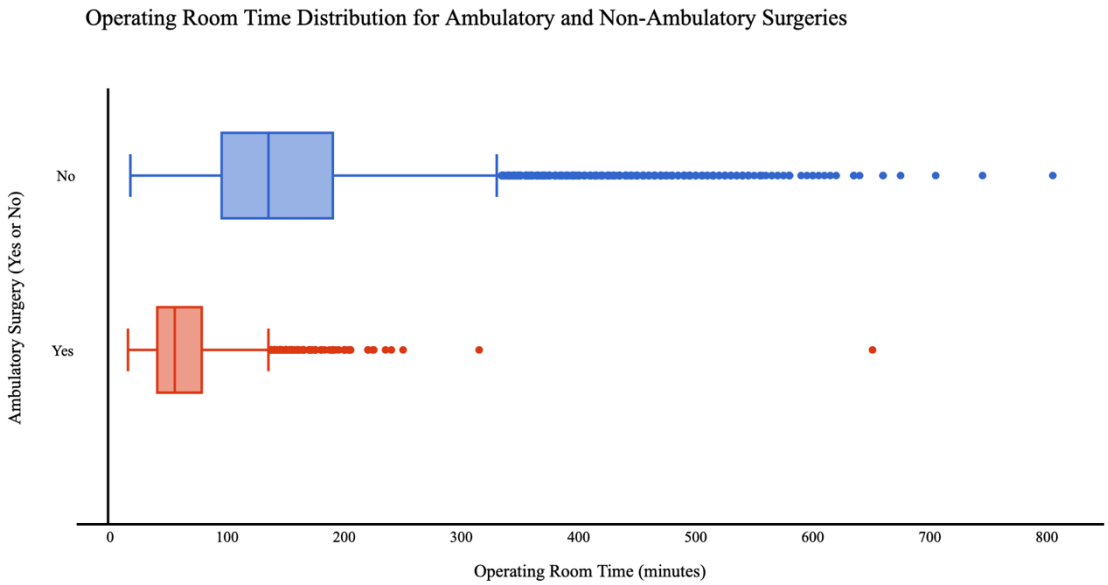

Figure 5.11 - Operating Room time distribution of ambulatory surgeries (red) and non-ambulatory surgeries (blue).

5.5 Hyperparameter Tuning

Before presenting the eight models results for all specialties, and at a specialty level, the results of the random and grid search with CV for the hyperparameter tuning results for the RF specialty level models are presented.

Table 5.3 summarizes the best set of hyperparameters found for each specialty-level RF algorithm along with run time and MAE, which was defined as the criteria selection parameter.

Table 5.3- Summary of the best hyperparameters after hyperparameter tuning phase for Urology, General Surgery and Orthopedics Random Forest models.

Specialty	MAE (minutes)	Number of trees	Maximum Depth of Three	Nb of variable sampled at each split	Run Time (seconds)
Urology	20.24	800	10	4	20.6
General Surgery	25.8	200	10	4	21.0
Orthopedics without knee and hip	26.8	500	10	4	21.4
Orthopedics	26.0	700	30	4	70.4

5.6 Model's Results

Each subsection of this chapter corresponds to a different approach (All specialty and individual specialty model) and contains the same structure: a scatter plot of the actual OR time versus the respective model prediction and a table with the summarized evaluation metrics for the RF and MLR model. The red line in all graphs represents the perfect hypothetical relationship where the model's prediction corresponds to the real OR time. The MLR model results do not include the non-significant variables that were reported in section 5.7.

The log-transformation in the output variable, OR time, does not lead to an improvement in any evaluation metrics. Therefore, this section does not expose the results relative to this transformation.

5.6.1 All Specialties Models

The All Specialties Model was the first approach under analysis. This model aimed to build a generalized model that captures the OR patterns of the different specialties and then built specific specialty models. **Figures 5.12** and **5.13** refer to the comparison of the model's prediction and the actual OR time for MLR and RF, respectively. **Table 5.4** summarizes the evaluation metrics. The MLR presented an R-squared of 0.780 compared with the 0.682 of the RF model, but higher MAE and RMSE and similar over/under-estimation and within percentages.

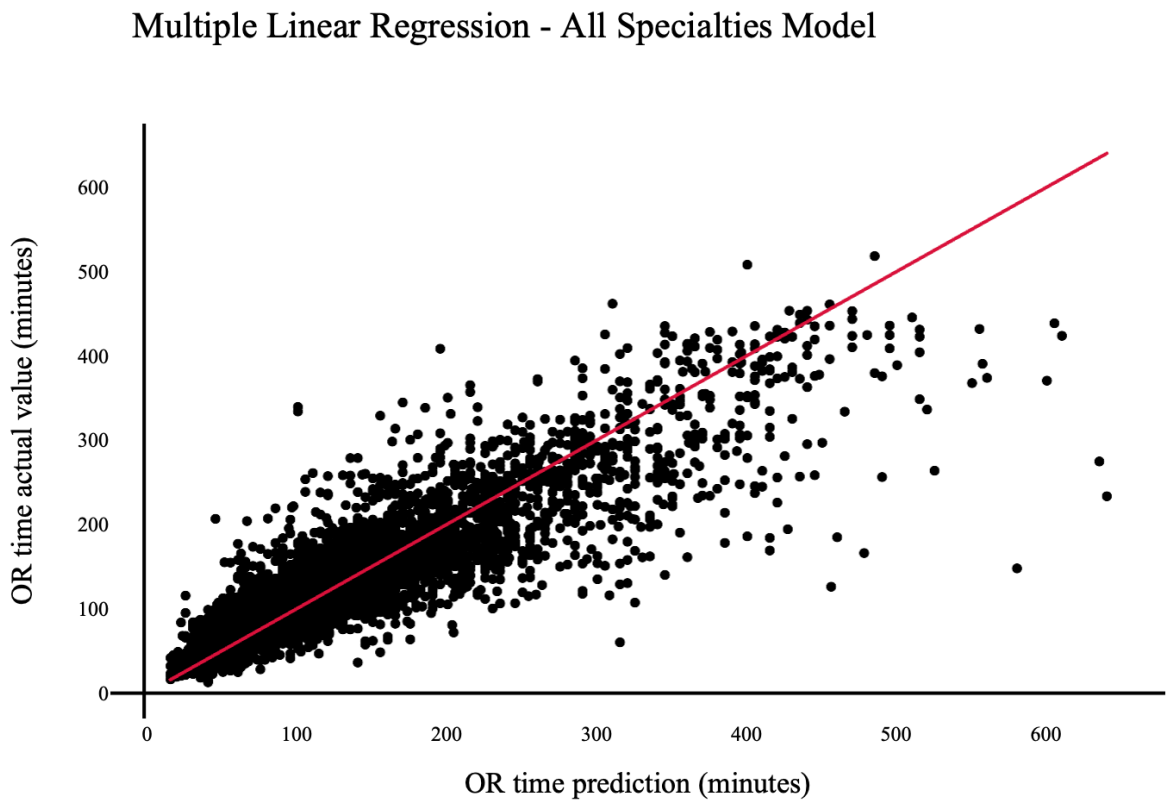

Figure 5.12- Scatter plot of actual Operating Room time duration (Y-axis) versus Multiple Linear Regression Operating Room time predictions (X-axis) for All Specialties. The red line represents the hypothetical relationship where the model's prediction corresponds to the real Operating Room time.

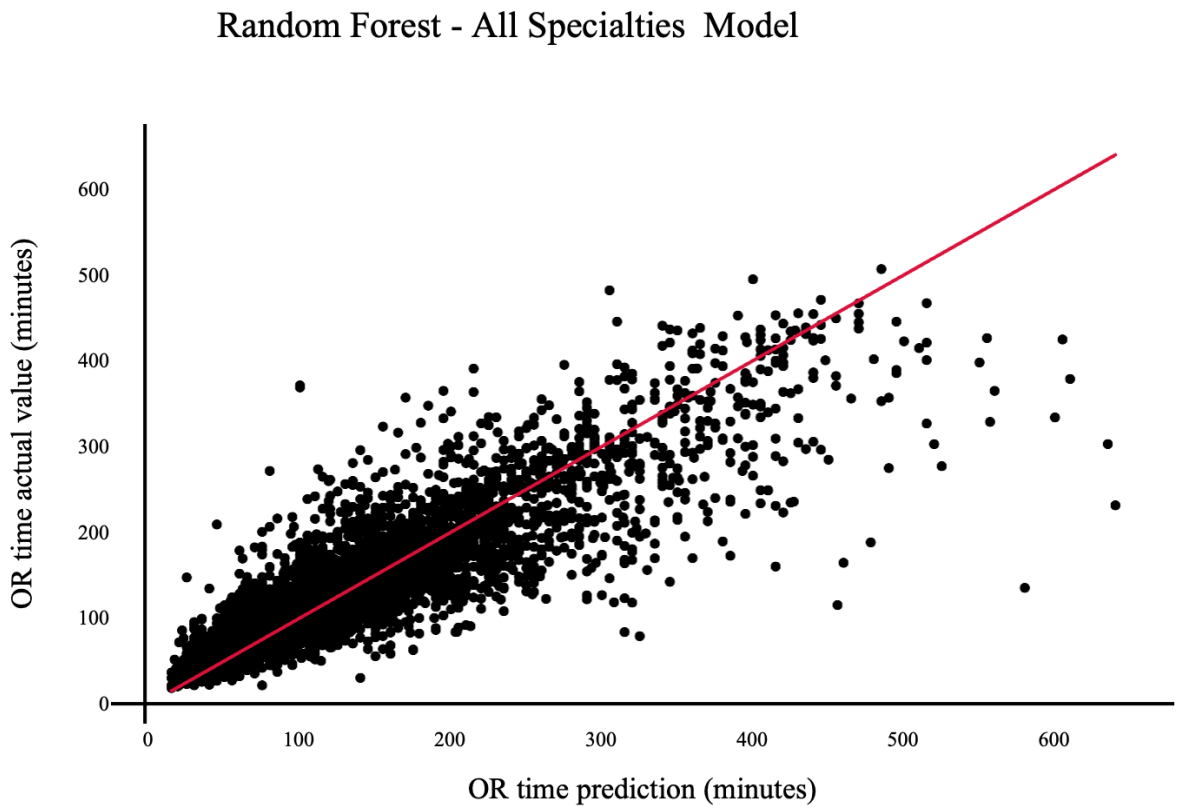

Figure 5.13- Scatter plot of actual Operating Room time duration (Y-axis) versus Random Forest Operating Room time predictions (X-axis) for All Specialties. The red line represents the hypothetical relationship where the model's prediction corresponds to the real Operating Room time.

Table 5.4- Evaluation Metrics for the All Specialties Multiple Linear Regression and Random Forest models.

All Specialties Model	MAE (minutes)	RMSE (minutes)	R-squared (Train Phase)	R-squared (Test Phase)	Overesti-mation (%)	Underesti-mation (%)	Within (%)
Multiple Linear Regression	26.9	41.5	0.797	0.780	37	30	33
Random Forest	26.0	40.5	0.723	0.682	37	29	34

5.6.2 Urology Model

The Urology Model corresponds to the model with the lowest data volume, and to the lowest OR time median (70.0 minutes), **Figure 5.4**. **Figures 5.14** and **5.15** establish the comparison of the model's predictions and the actual OR time followed by the summary of the evaluation metrics, **Table 5.5**. The RF model presented a higher R-squared (0.831) and lower MAE (20.9 minutes) when compared with MLR (R-squared=0.822, MAE=21.7 minutes), with equal percentages for the over, under and within estimations, and a higher RMSE (RMSE (RF)= 35.0 minutes vs. RMSE (MLR)= 32.7 minutes).

Table 5.5- Evaluation Metrics for the Urology Multiple Linear Regression and Random Forest models.

Urology Model	MAE (minutes)	RMSE (minutes)	R-squared (Train Phase)	R-squared (Test Phase)	Overesti-mation (%)	Underesti-mation (%)	Within (%)
Multiple Lin-ear Regression	21.7	32.7	0.834	0.822	40	30	30
Random Forest	20.9	35.0	0.893	0.831	40	30	30

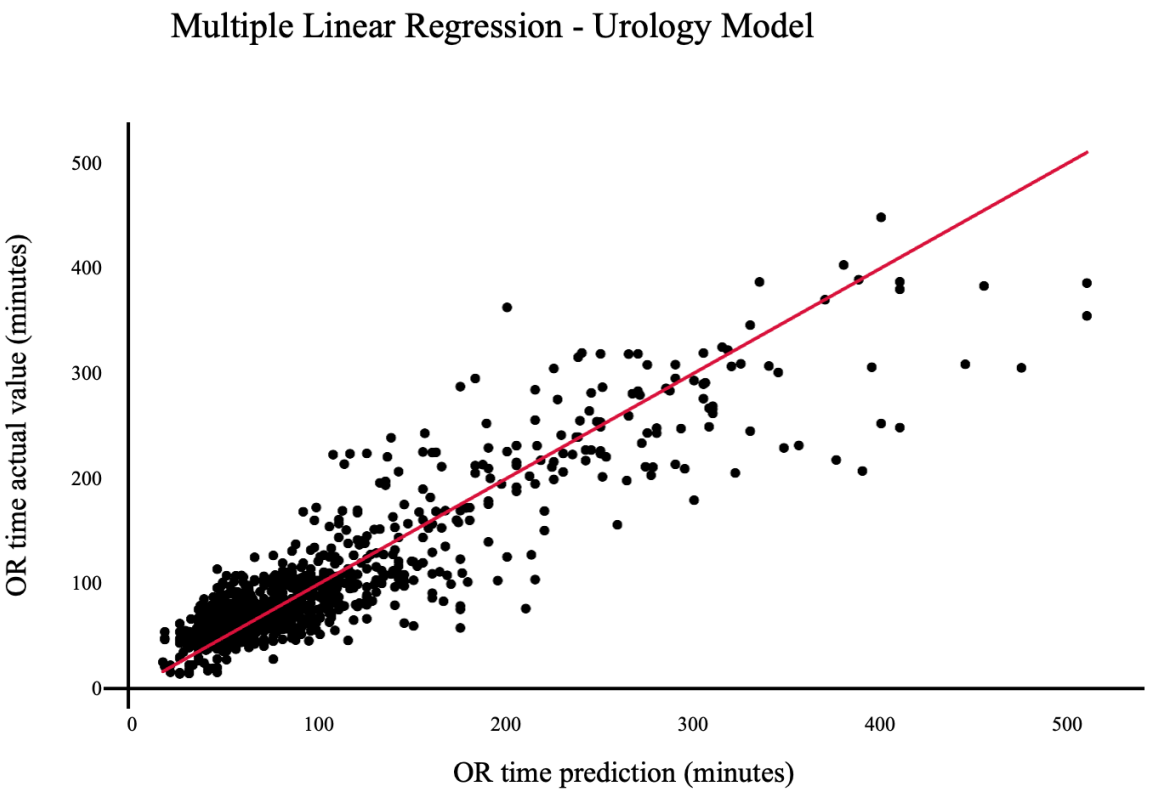

Figure 5.14- Scatter plot of actual Operating Room time duration (Y-axis) versus Multiple Linear Regression Operating Room time predictions (X-axis) for Urology. The red line represents the hypothetical relationship where the model's prediction corresponds to the real Operating Room time.

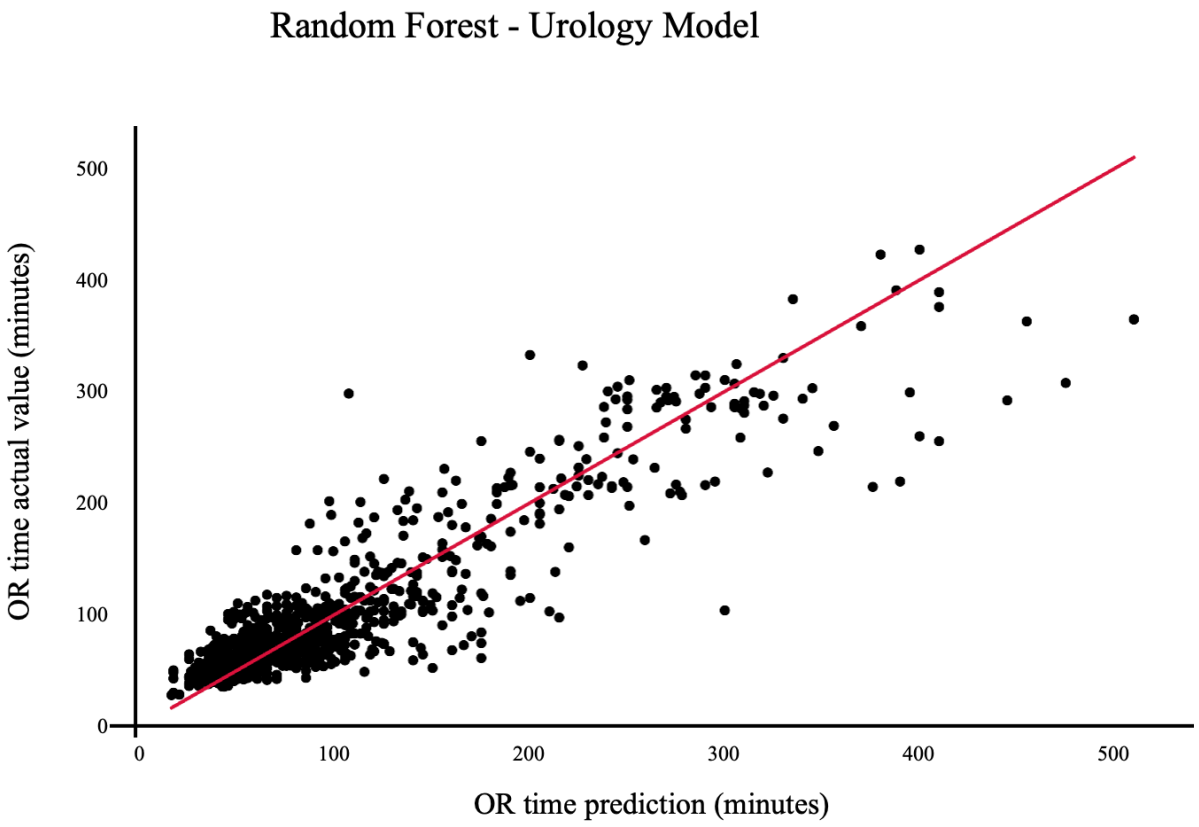

Figure 5.15- Scatter plot of actual Operating Room time duration (Y-axis) versus Random Forest Operating Room time predictions (X-axis) for Urology. The red line represents the hypothetical relationship where the model's prediction corresponds to the real Operating Room time.

5.6.3 General Surgery Model

Similarly, to the All Specialties and Urology Models, **Figures 5.16** and **5.17** represent the comparison of the model's predictions and the actual OR time for MLR and RF, respectively, and **Table 5.6** a summary of the evaluation metrics. All evaluation metrics were similar for both models, as it is possible to observe in **Table 5.6.**

Table 5.6- Evaluation Metrics for the General Surgery Multiple Linear Regression and Random Forest models.

General Surgery Model	MAE (minutes)	RMSE (minutes)	R-squared (Train Phase)	R-squared (Test Phase)	Overestimation (%)	Underestimation (%)	Within (%)
Multiple Linear Regression	26.2	40.2	0.828	0.826	37	29	34
Random Forest	26.1	40.4	0.854	0.825	37	28	35

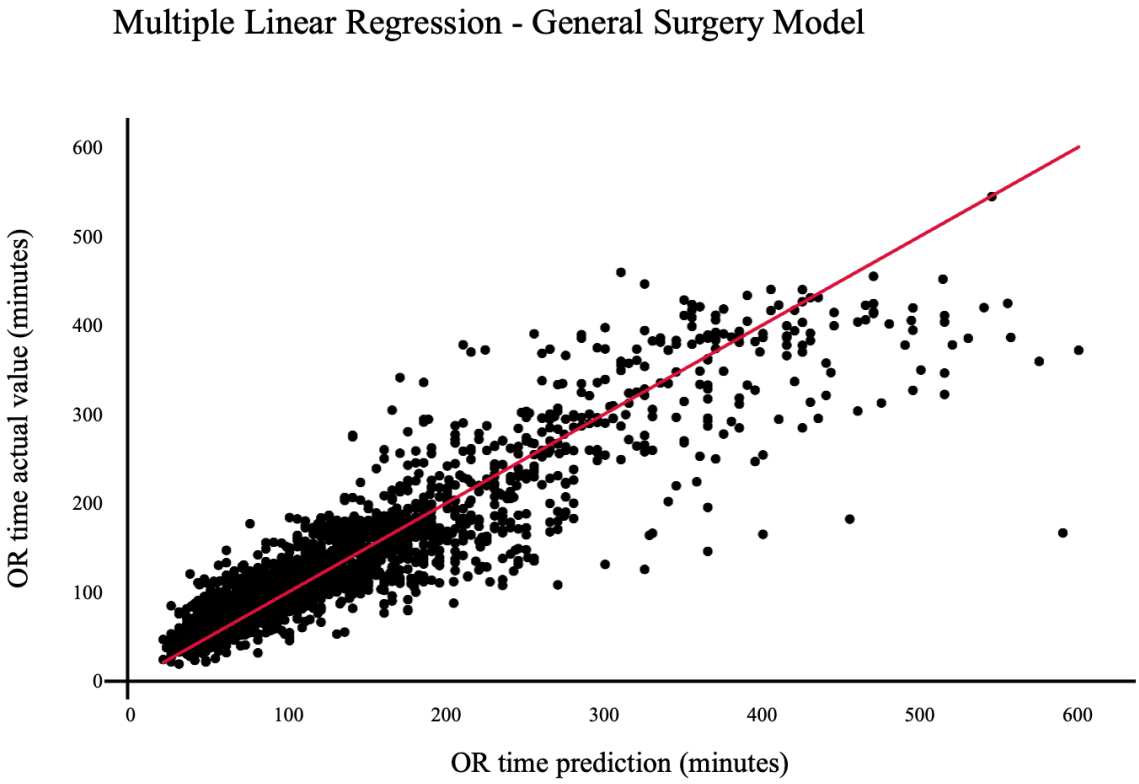

Figure 5.16- Scatter plot of actual Operating Room time duration (Y-axis) versus Multiple Linear Regression Operating Room time predictions (X-axis) for General Surgery. The red line represents the hypothetical relationship where the model's prediction corresponds to the real Operating Room time.

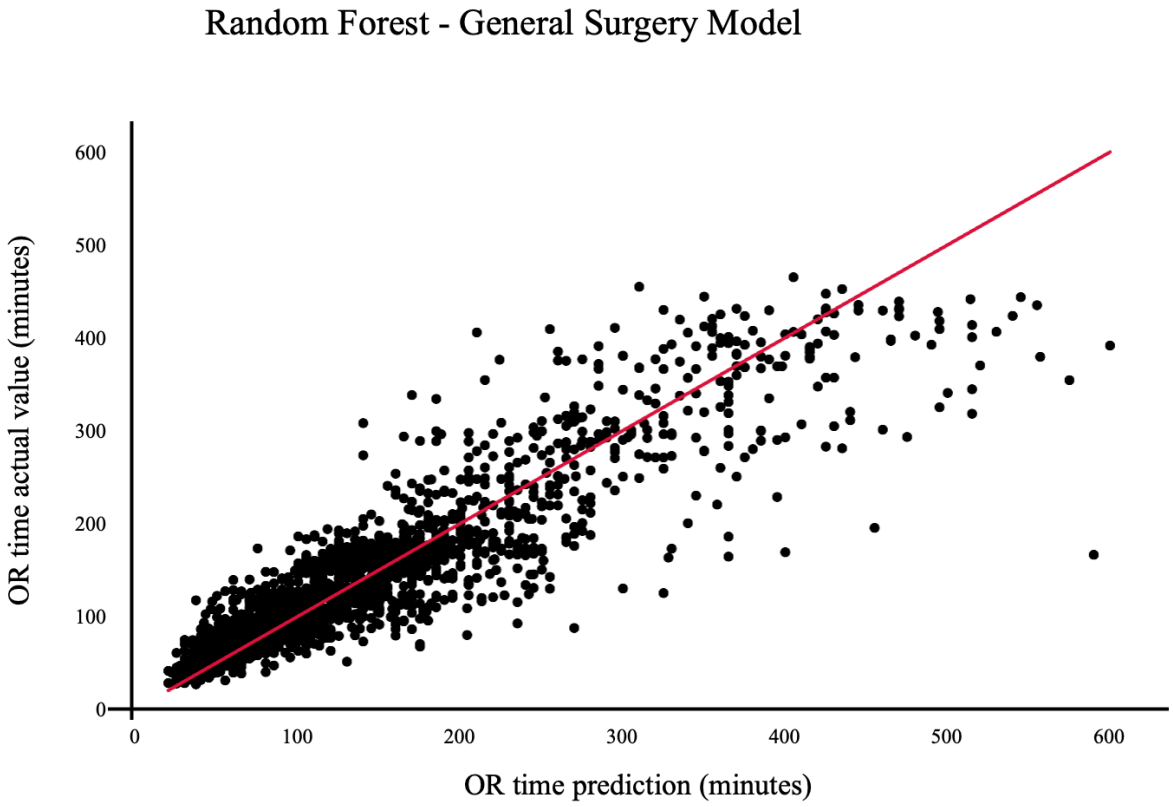

Figure 5.17- Scatter plot of actual Operating Room time duration (Y-axis) versus Random Forest Operating Room time predictions (X-axis) for General Surgery. The red line represents the hypothetical relationship where the model's prediction corresponds to the real Operating Room time.

5.6.4 Orthopedics Model

When the MLR algorithm ran in the dataset containing all surgeries from Orthopedics specialty, it was not capable of modeling the data. This was reflected by the negative R-squared. A negative R-squared might be related with the chosen model, which might not be adequate to model the data. To test this possibility, the RF algorithm was applied to the Orthopedics dataset. If when we changed the ML algorithm it leads to satisfactory evaluation metrics, it suggests that the MLR does not capture the data patterns well indicating that this model fits the orthopedic data poorly and, therefore, is not suitable for the data. The RF algorithm was conducted to a positive R-squared of 0.683, and the predictions can be observed in **Figure 5.20**. Then, the data was separated according to functional units to test where the MLR model fails in fitting the data. Considering that Orthopedics is a specialty with many different types of procedures in different regions of the human body, commonly this specialty is divided according to the region where the surgical procedure is going to take place, the functional units. With the validation of the CHULC's director of Orthopedics these functional units were divided into: spine, arm, leg, knee, hip, hand, and foot. After this Orthopedic dataset division, the MLR algorithm ran in the dataset corresponding to each functional unit, the only negative R-squared was associated with the hip and knee surgical procedures datasets. Therefore, the MLR was run in the Orthopedics dataset excluding the hip and knee surgical procedures, which conducts to an R-squared of 0.685, **Figure 5.18**. To have a comparative-term model, RF was also applied in this dataset and presented an R-squared of 0.702, **Figure 5.19**. Lastly, **Table 5.7** summarizes the evaluation metrics of the three Orthopedic models. The MLR and RF for the data without the knee and hip surgeries presented similar results. Although the RF for all data presented the lowest R-squared, 0.683, the MAE and RMSE were lower and presented a higher percentage of within predictions.

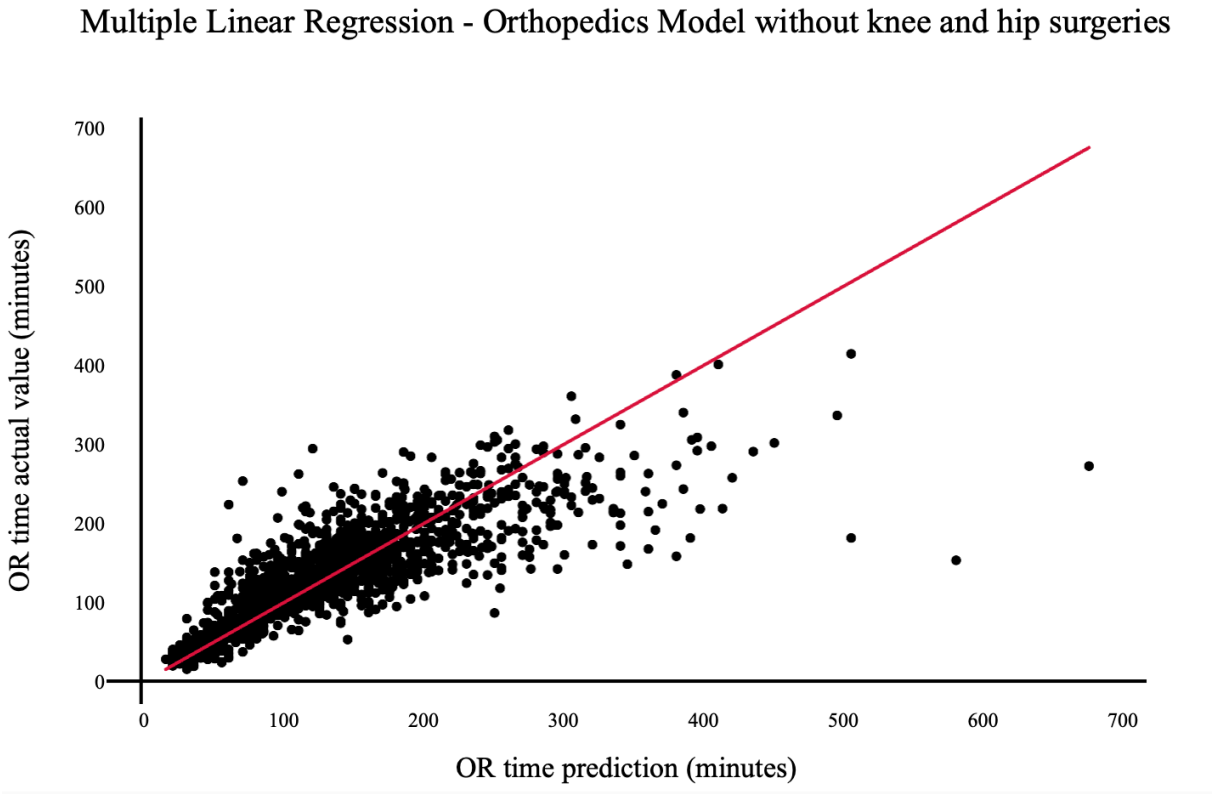

Figure 5.18 - Scatter plot of actual Operating Room time duration (Y-axis) versus Multiple Linear Regression Operating Room time predictions (X-axis) for Orthopedics without knee and hip surgeries. The red line represents the hypothetical relationship where the model's prediction corresponds to the real Operating Room time.

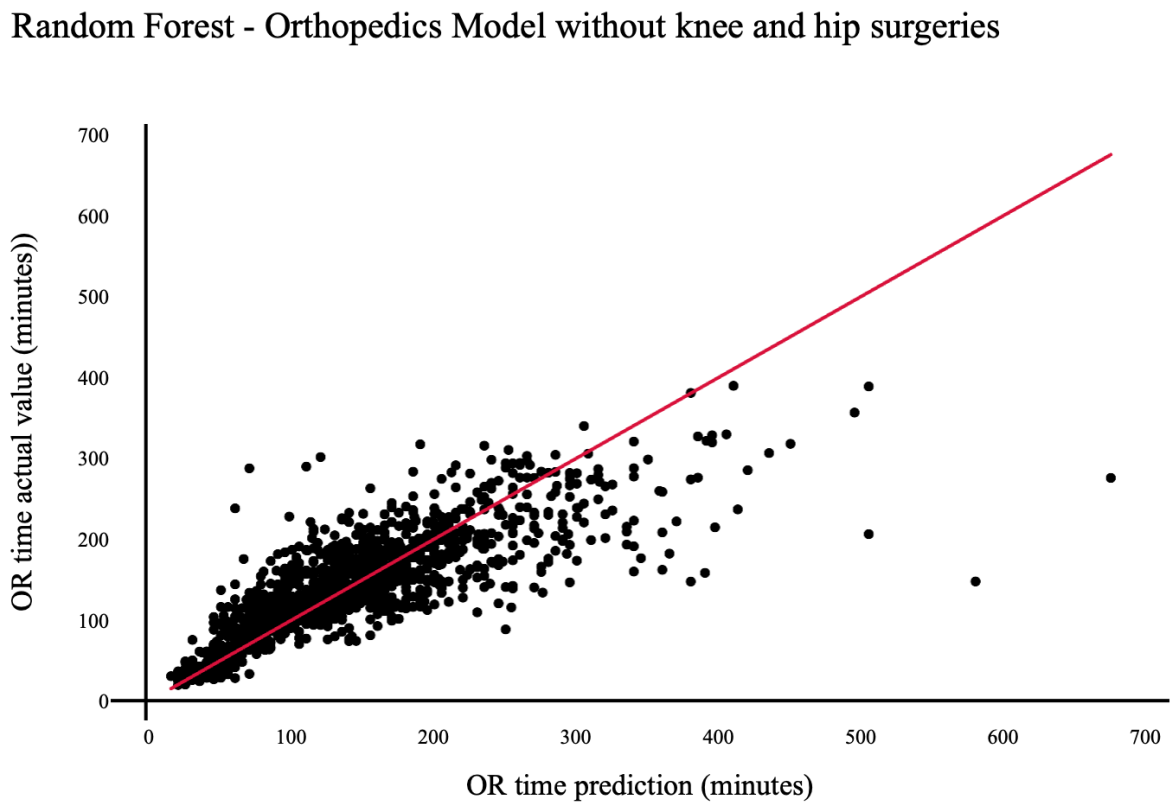

Figure 5.19- Scatter plot of actual Operating Room time duration (Y-axis) versus Random Forest Operating Room time predictions (X-axis) for Orthopedics without knee and hip surgeries. The red line represents the hypothetical relationship where the model's prediction corresponds to the real Operating Room time.

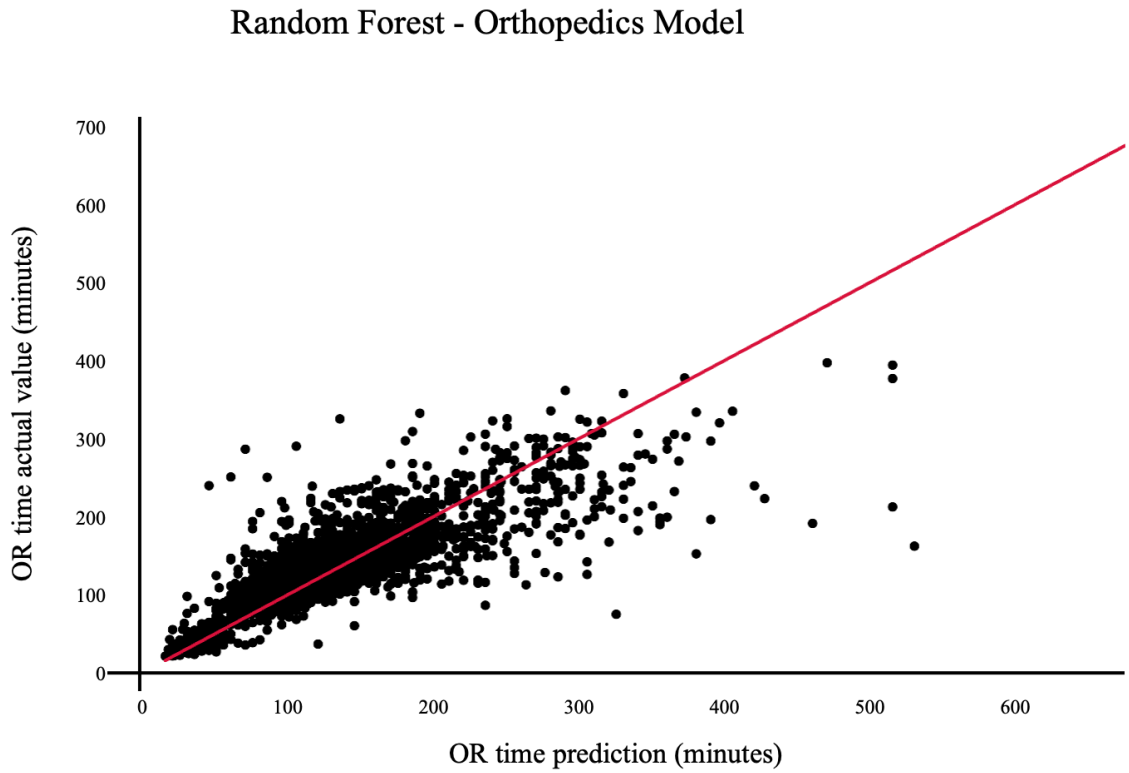

Figure 5.20- Scatter plot of actual Operating Room time duration (Y-axis) versus Random Forest Operating Room time prediction (X-axis) for Orthopedics. The red line represents the hypothetical relationship where the model's prediction corresponds to the real Operating Room time.

Table 5.7 - Evaluation Metrics for the Orthopedics Multiple Linear Regression and Random Forest models.

Orthopedics Model	MAE (minutes)	RMSE (minutes)	R-squared (Train Phase)	R-squared (Test Phase)	Overesti-mation (%)	Underesti-mation (%)	Within (%)
Multiple Linear Regression without knee and hip surgeries	28.9	44.9	0.699	0.685	38	31	31
Random Forest without knee and hip surgeries	27.4	43.6	0.798	0.702	38	29	33
Random Forest	27.1	41.1	0.824	0.683	38	28	34

5.7 Feature Importance and Significance

Table 5.8 and **5.9** summarize the results relative to the feature importance analysis. For the MLR algorithm the features that were not considered significant for the model, i.e., when the null hypothesis is accepted, so there is no relation between the feature and the output ($p>0.05$), were marked, **Table 5.8**, and eliminated from the MLR models. For a better understanding of the RF model, the top 5 feature scores were reported, along with the feature weight in **Table 5.9**.

Table 5.8- Non-significant variables for the Multiple Linear Regression models.

Model	Non-Significant Variables Name for MLR	Predictor Type
All Specialties	DES_PROVENIENCIA	Procedure characteristics
	DES_BLOCO	Procedure characteristics
	AMBULATORIA	Procedure characteristics
Urology	DIAS_EM_LIC	Patient characteristics
	DIAS_PRE_OP	Patient characteristics
	PRIORIDADE	Patient characteristics
	AMBULATORIA	Procedure characteristics
	DES_DESTINO	Procedure characteristics
General Surgery	ANESTESIA	Procedure characteristics
	GRUPO_DIAGNOSTICO	Patient characteristics
	DES_PROVENIENCIA	Procedure characteristics
	AMBULATORIA	Procedure characteristics
	DES_BLOCO	Procedure characteristics
Orthopedics without knee and hip	DES_SALA	Procedure characteristics
	DES_PROVENIENCIA	Procedure characteristics
	AMBULATORIA	Procedure characteristics
	SEXO	Patient characteristics
Orthopedics	----------	----------

Table 5.9- Feature weight for the five variables with the highest predictive power for Random Forest models.

RF Model	Variable Name	Feature Weight (%)	Predictor Type
All Specialties	Cirur_Ato	72.92%	Surgical Team characteristics
	Anestesista_tipo	2.59%	Surgical Team characteristics
	IDADE	2.22%	Patient characteristics
	DIAS_EM_LIC	2.07%	Patient characteristics
	DIAS_PRE_OP	1.25%	Patient characteristics
Urology	Cirur_Ato	75.32%	Surgical Team characteristics
	IDADE	8.3%	Patient characteristics
	Anestesista_tipo	7.35%	Surgical Team characteristics
	DIAS_EM_LIC	1.12%	Patient characteristics
	TIPO_INTERVENCAO	1.07%	Procedure characteristics
General Surgery	Cirur_Ato	74.80%	Surgical Team characteristics
	Anestesista_tipo	12.15%	Surgical Team characteristics
	IDADE	2.16%	Patient characteristics
	DIAS_EM_LIC	1.12%	Patient characteristics
	TIPO_INTERVENCAO	1.09%	Procedure characteristics
Orthopedics without knee and hip	Cirur_Ato	70.64%	Surgical Team characteristics
	Anestesista_tipo	14.81%	Surgical Team characteristics
	IDADE	2.52%	Patient characteristics
	DIAS_PRE_OP	2.14%	Patient characteristics
	DIAS_EM_LIC	1.76%	Patient characteristics
Orthopedics	Cirur_Ato	70.11%	Surgical Team characteristics
	Anestesista_tipo	16.09%	Surgical Team characteristics
	IDADE	2.79%	Patient characteristics
	DIAS_PRE_OP	2.10%	Patient characteristics
	DIAS_EM_LIC	1.82%	Patient characteristics

5.8 Comparison with current Methods

The end of this chapter points out the state of the current CHULC methods to predict OR time and correspond the goal of comparison with the developed models. First, the plot that relates the predicted surgery time by the surgeon for all the surgeries is provided in **Figure 5.21**. The red line represents the hypothetic scenario where the surgeon's predictions correspond to the real OR time. **Table 5.21** contains the over, underestimation and within percentages for the current methods and for the model results, as well as the error evaluation for both surgeon and ML models predictions in the test set.

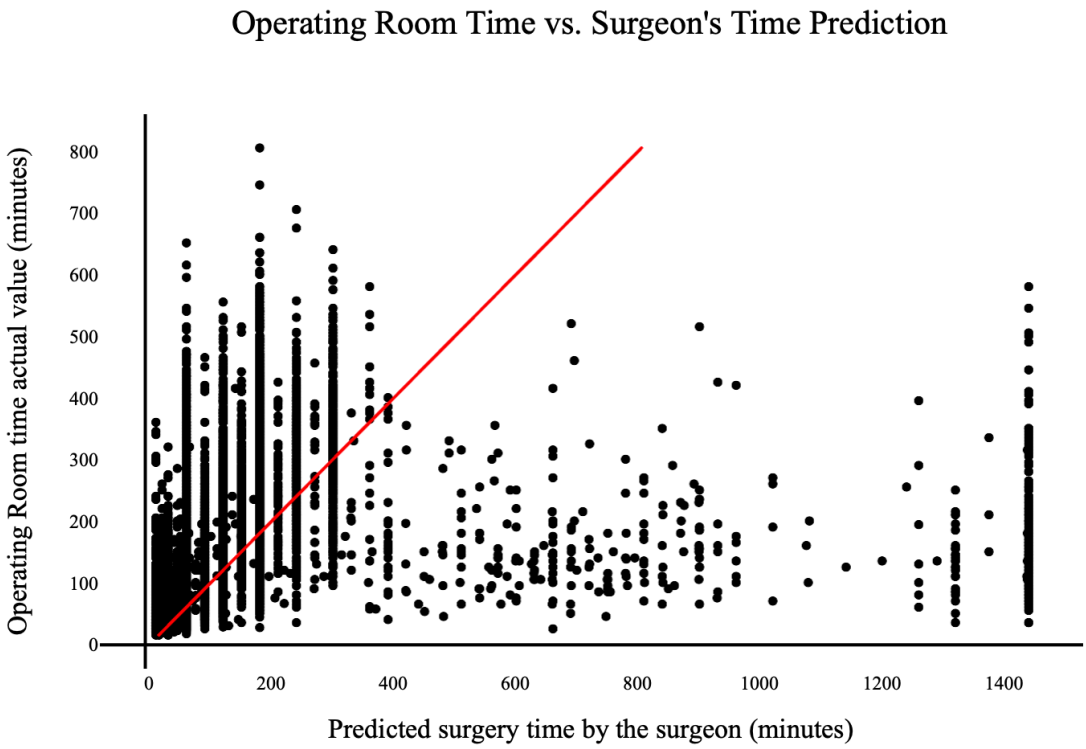

Figure 5.21- Scatter plot of actual Operating Room time duration (Y-axis) versus the predicted surgery time by the surgeon (X-axis). The red line represents the hypothetical relationship where the surgeon's predictions correspond to the real Operating Room time.

Table 5.10- Comparison of within, over- and under-estimation percentages for current methods vs. the ML models.

OR time Prediction Method	Overesti-mation (%)	Underestima-tion (%)	Within (%)	MAE (minutes)	RMSE (minutes)
All Specialties- Current Method	16	74	10	73.5	165.5
All Specialty Model- MLR	37	30	33	26.9	41.5
All Specialty Model- RF	37	29	32	26.0	40.5
Urology- Current Method	30	54	16	34.8	55.4
Urology Model-MLR	40	30	30	21.7	32.7
Urology Model-RF	40	30	30	20.9	35.0
Orthopedics- Current Method	13	77	10	97.5	228.9
Orthopedics Model- RF	38	28	34	27.1	41.1
General Surgery- Current Method	12	79	9	70.3	126.0
General Surgery Model-MLR	37	29	34	26.2	40.2
General Surgery Model-RF	37	28	35	26.1	40.4

6. Discussion

The purpose of this study was to develop ML models based on a linear regression model, the MLR model, and a non-linear DT-based model, the RF model. It was hypothesized that ML techniques might help to enhance OR time predictions, allowing better OR scheduling, efficiency and cost savings. All the models' results improve the percentages of within predictions and reduce the underestimation of OR time prediction, with higher overestimation predictions. All the presented models outperform the current CHULC methods by achieving significant reductions in the error of OR time predictions.

The study's first goal was to comprehend the current OR time prediction in CHULC. After the reunions with the different surgical team experts and professionals involved in the OR environment, it was possible to conclude that this task was only performed by the surgeons and based on their own experience, resulting, most of the time, in underestimated predictions. This fact was confirmed by the high underestimation percentages that can be noticed in **Table 5.10**. Overall, the percentage of underestimation for all the specialties included in the CHULC dataset is 74%, and, when analyzed by specialty, General Surgery presented the highest percentage (79%), followed by Orthopedics (77%), and Urology (54%). The within percentages reported in the current CHULC methodology were low, around 10%. This percentage was only higher for the Urology specialty, which reached 16%. This first analysis of the current CHULC methods enables reporting that the bulk of the surgeries is underestimated, and the accurate estimations are sparse.

The correlation between times was analyzed to confirm the hypothesis that other time factors besides the SCT, particularly the ACT, that do not receive much importance when predicting OR time, can also impact this time. The preparation and final time had a minor impact on OR time, with a Person's Coefficient of 0.234 and 0.191, respectively, **Table 5.2**. The SCT was the time variable with the highest Pearson's Coefficient, 0.966, indicating the strongest relation compared to the remaining time variables. This result highlights that the current methods to predict OR time-based on the SCT is not strictly a bad practice since this is the time with the strongest relationship with the OR time. Although, by doing this, we are not looking into the whole scenario and fail a substantial number of OR time predictions, since the ACT presents a high Pearson's Coefficient with OR time, closer to 0.700. Therefore, should not be discarded from the equation when predicting OR time.

After the first cleaning phase, it was possible to conclude that some fields were not accurately filled, especially in blank or with 0. This fact leads to the elimination of some entries. It diminishes the statistical power of the features along with the significant number of outliers and missing values, making the data less complete and reducing the power of model predictions. The more quality data that a ML model receives, the faster it can learn and improve the predictions. Therefore, it is crucial to reinforce the consciousness of healthcare institutions and professionals to make efforts to produce high-quality data and precision in the data which is reported. Although its negative influence on the model's accuracy, this study still considered a reasonable number of samples, a total of 27051 surgical procedures.

When faced with the cardinality problem for categorical variables, there is a trade-off between the dimensionality reduction and the model interpretability and robustness. On the one hand, dimensionality reduction reduces the computational time and the storage space, but it might eliminate some essential information. With this in mind, instead of eliminating the "COD_DIAGNOSTICO", this variable was converted into the "Grupo_diagnostico" for All Specialties, General Surgery and Orthopedics Models. For the Urology Model, this variable was grouped to reduce the number of levels. By

implementing the “Grupo_diagnostico" variable, the patient diagnosis information was not as detailed as the "COD_DIAGNOSTICO" since this is general information about the patient's condition. For instance, the category “Neoplasm” can point to a variety of different types of diagnosis. This trade-off was also extended to other variables, and the implementation of dimensionality reduction techniques, such as the elimination of the non-significant features based on the t-statistic test for the MLR model, was also conducted to eliminate some information. However, it is important to highlight that if more data is added, the p-values of the statistical test do not remain equal in most of the cases, therefore new tests must be performed. Otherwise, we might not be considering features with statistical significance for the MLR model with the newly added data.

The new generated features had the purpose of converting categorical information regarding the OR staff, into continuous information, with a high correlation with the model output. After the correlation analysis and the elimination of the variables with high collinearity, it was possible to conclude that among the staff variables, the variables relative to the surgeon, “Cirur_Ato”, had the highest correlation with the OR time, given by an elevated Pearson’s coefficient in **Table 5.1**, when compared with the anesthetist variable, with a 0.865 Pearson’s Coefficient for surgeon’s variable versus the 0.535 coefficient for the anesthetist one. This information allows concluding that the surgeon’s variables are strongly correlated with the OR time among the other staff variables.

The data distribution suggests that all the specialties presented a long right-skewed tail. However, it is not possible to assert with certainty that it corresponds to a perfect log-normal distribution due to the non-regular peaks that can be identified in graphics from **Figures 5.3** to **5.6,** especially for the Orthopedics data, **Figure 5.6**. This long right-skewed tail indicates that for OR time, a small number of surgical procedures might take considerably longer than the average since it may take on values from zero to infinity and is skewed with a long right tail.

Regarding the exploratory data analysis, some plots were analyzed to comprehend the relation between the input features and the output, the OR time, for both continuous and categorical variables. Starting with the continuous variables, **Figures 5.7** and **5.8** provide an example of a non-linear and linear relationship with the output, respectively. **Figure 5.7** addresses the relation between the days on the waiting list and the OR time. Surgeries with higher waiting days correspond to a lower priority level for surgery and are relative to less complex procedures, which results in lower OR times. This is not a linear relation, as is possible to notice by the graphic tendency. Another outcome of this figure is the high number of days on the waiting list reported in CHULC, with the highest number reaching 1818 days, corresponding to five years on the waiting list. Although the patient presented a lower priority level, this particular example highlights the necessity to optimize the OR schedule to reduce the number of patients and days on the waiting list and improve the surgical response by the hospital center.

Figure 5.8 describes the relationship between the new variable generated, “Cirur_ato”, the mean of the surgical time based on the surgeon as well as the type of surgical procedure, and the OR time. This variable presented a high linear correlation, also described by the 0.865 value of Pearson’s Coefficient, **Table 5.2**. This strong linear relation provides this variable a strong predictive power in MLR models. Besides the predictive power in MLR models, this variable was also influential for the RF model, as presented in **Table 5.9**, it revealed the highest feature score for the RF across all models.

For the analysis of the categorical variables, **Figure 5.9** addresses the distribution of ACT according to the ASA risk of anesthesia. According to **Table 4.2**, a higher ASA describes a patient with severe systemic disease, a higher BMI, and a higher risk of intra- and postoperative complications, which

will lead to longer ACT and OR times. Excluding ASA 5, **Figure 5.9** reflects higher ACT times for higher ASA classification. This figure induces the idea that an ASA 5 leads to lower ACT times, which is not reflected in clinical practice. This distribution is biased since only three surgeries with patients with ASA 5 in the past five years were performed in CHULC. Therefore, this is a very small sample and should not be compared with the other ASA categories. An ASA 5 represents a life-threatening operation leading to the dilemma for professionals regarding the decision to operate, and therefore there is a reduced number of ASA 5 category operations. When analyzing this distribution, it was possible to notice that 4969 ASA classifications are missing, presented by the category "Not defined". This represents a total of 18% of missing values for this category, which can be explained, most of the time, by the absence of the anesthesia consultation. Consequently, in some of these cases, the anesthetist needs to observe the patient when he enters the OR and, in that moment, decide the type and the risk of anesthesia for the patient to then obtain the anesthetic drugs, which will influence the OR time.

Figure 5.10 states the OR distribution based on the type of schedule. It is possible to identify three main types of surgical scheduling, apart from the cases of reoperation within 24h, the SIGIC-Extern, basic, and additional scheduling. The basic scheduling addresses all elective surgeries scheduled in the hospital by the surgeons during their work period. Additional scheduling is included in SIGIC. The main difference between additional and SIGIC extern is that the operation takes place in the hospital installations in the additional scheduling. In this case, the professionals work outside their working period. SIGIC extern is when the surgical procedure is performed outside the institution when the hospital does not have the capacity to operate the patients within the maximum response time guaranteed. In this situation, the patient receives a surgical voucher to be operated in another hospital institution, and the hospital is responsible for paying these costs. The SIGIC schedule tends to program the less complex surgeries, especially the extern situations, as reflected by the lower OR times. Consequently, the more complex patients and surgical procedures are scheduled on the basic programming, as reflected by the long OR times. This tendency can be explained by the fact that the hospital finances larger monetary amounts for surgeries performed outside the surgeon's work period, therefore this is a strategy to save costs.

The last analysis in the exploratory data analysis is referring to the OR time distribution for ambulatory and non-ambulatory surgeries, **Figure 5.11**. Ambulatory surgeries include surgeries where the patient does not need to stay overnight in the hospital. This type of surgery tends to be simpler, and the patient is a controlled subject without complex comorbidities or other health factors that might complicate the operative and postoperative period. Due to these intrinsic characteristics to classify a patient for ambulatory surgery, the OR time will be shorter when compared to non-ambulatory surgery, where the patient must stay overnight in the hospital.

After the data preparation framework, the data was prepared for the model implementation. Overall, and based on the literature, it was expected that the RF models would outperform the MLR models. However, in this study, this was not observed in all the models. Apart from Orthopedics Model, all presented similar evaluation metrics for both MLR and RF models in each approach. The principal explanation was that the feature with the most significant impact on the RF model was the "Cirur_Ato", **Table 5.9**. This feature also presented a high linear correlation with the output, traduced by a Pearson coefficient of 0.865, **Table 5.1**. Hence, the MLR also modeled this linear relation with high precision and achieved similar results to the RF model. This linearity can also be noticed in **Figure 5.1**. The following paragraphs will focus on the discussion of the specialty models.

The All-Specialty Model was the first model built to understand if the MLR and RF algorithms are suitable to model the data under analysis, if they respond to the research question with good results, and if they are capable of generalizing the results. For a 115.0 median of surgical time, both models presented satisfying performances with an R-squared surrounding 0.700 (MLR=0.780, and RF=0.682), with the MLR presenting a higher R-squared but a higher MAE (MLR=26.9 minutes vs. RF=26.0 minutes), **Table 5.4**. The models were similar when comparing the over/under-estimation and within percentages. The first model analysis made it possible to conclude that the ML models reduce the OR time prediction mean absolute error, in approximately 48 minutes, enhance the within percentages, lower underestimation but with the cost of higher overestimations. Precisely, for the All-Specialties Model, this improvement corresponds to an increase of 23% in within percentages and a decrease in underestimation of 44% approximately, **Table 5.10**.

The Urology Model contained a surgery median of 70.0 minutes with the lowest MAE (MLR=21.7 minutes vs. RF=20.9 minutes), and RMSE (MLR=32.7 minutes vs. RF=35.0 minutes), and an R-squared of 0.822 for the MLR model and 0.831 for RF model, **Table 5.5**. These lower prediction errors might be related to the fact that Urology, when compared with the other two specialties, is the one with more standardized procedures and less variability among the patients, diagnosis, and surgical procedures, accordingly to CHULC professionals. However, this specialty had the highest overestimation percentage by both surgeon's predictions (30%), and by the models (40%).

The General Surgery Model had the highest within percentage of predictions (35%). Similar to the previous models, there were no significant differences between the MLR and RF models, with an R-squared of 0.826 for the MLR model and 0.825 for the RF, **Table 5.6**. The MAE, and RMSE evaluation metrics were also similar, with MLR= 26.2 minutes, RF=26.1 minutes and MLR=40.2 minutes, RF=40.4 minutes for MAE and RMSE, respectively.

For both Urology and General Surgery Models, a decrease in the MAE and RMSE was registered, with an increase of the R-squared compared to the All Specialties Model. In this study, for the Urology and General Surgery specialties models, it was not possible to conclude with certainty which model had the best performance, MLR or RF, due to their similar results in the evaluation metrics as exposed previously. However, a new discussion must be performed if new data or predictors are added.

The last model under evaluation was relative to the Orthopedics data. After the first trial to build a MLR model, it conducted to a negative R-squared, and the possible cause was reported, the selected model fits the data poorly. The R-squared evaluates how well the selected model fits the data when compared with a horizontal straight line (the mean value). Therefore, a negative R-squared occurs when the selected model fits the data worse than a horizontal line. Taking this into account, the RF model was tested. The RF was able to model the Orthopedics data with an R-squared of 0.683 and a MAE of 27.088 minutes, indicating that an RF model suits better for all surgical procedures in this specialty.

In the meeting with the director of the Orthopedics specialty, the doctor stated that this is a specialty with a high degree of unpredictable variability compared to other specialties. Even for surgeons with many years of experience, it is not an easy prediction. Additionally, the conditions of the patients admitted to surgery reported in section 4.1.2 increase the uncertainty of the problem. In this reunion, the doctor also affirmed that specification is necessary for this specialty. Therefore, it is possible to define some functional units, the name given in this hospital center to categorize the Orthopedics surgeries based on the human body anatomy where the surgery will take place. By separate the data and evaluate the MLR model performance, the hip and knee surgeries were removed, and it was possible to achieve an R-squared of 0.685 with a MAE of 28.9 minutes and an RMSE of 44.9 minutes, **Table 5.7**.

The knee and hip surgeries represent a substantial part of this dataset and do not present a linear relationship between the input and the output for most variables. Therefore, the MLR failed to model this relation in this specific type of surgical procedures. The RF model without the knee and hip surgeries served as a comparative set with the MLR model but conduct to similar evaluation metrics. In Orthopedics Model, the R-squared was lower when compared with the other specialty models since this is the specialty with the highest uncertainty. The plausible explanation for this is the fact that for the Orthopedics Model, there are a significant number of surgeries with a high degree of variability, as reported. Therefore, the exercise of predicting the OR time is more complex, as recognized by the CHULC professionals.

Generally, there were no discrepant differences between the R-squared of train and test phases. Since the models performed satisfactory on training and test phases, over- or under- fitting is less likely to exist.

The feature importance and significance analysis was preponderant in two phases: selecting the features for the MLR models and describing which features are relevant to the RF model, which can aid in a better understanding of the resolved problem. Starting with the MLR variable's significance is important to clarify that when the t-test was applied, it does not consider the interaction between variables. If a significant number of variables in the model do not present a linear relation with the output, the MLR is not suitable for data modeling. Whereas some procedure and patient characteristics variables were not significantly correlated with the OR time ($p>0.05$), depending on the model, the surgical team characteristics variables were significant correlated with the OR time in all models ($p<0.05$), indicating that these variables are valuable to perform the OR prediction in these models.

Finally, by the analysis of the RF scores, **Table 5.9** it is possible to state the features that most impact the models across the specialties are similar, and the mean of the surgeon based on the type of procedure is the most powerful. This indicates once again that the personnel characteristics, especially the surgeon ones, are essential for this task. Additionally, the procedures characteristics were the less impactful, especially where the operation is going to take place, with no level of significance for the MLR model and with a low feature weight in RF model. Regarding the patient's personal characteristics, age was shown to be the one with more weight for the RF model.

Other variables such as “Anestesista_tipo”, “DIAS_EM_LIC ”, and “DIAS_PRE_OP” also presented a relevant feature weight for RF models, suggesting that anesthesia concerns, the number of days in waiting list, and preoperative days are also important information for OR time prediction task, that must be consider.

The hospital OR time prediction was also included as a comparative variable to objectively compare the variations between the hospital strategy and the ML models. When understanding the current methods in CHULC it was possible to notice that most of the OR predictions were underestimated by the surgeons, about 74% when looking into all the specialties, with a low percentage of within predictions, 10% and with non-significative overestimation percentage, **Table 5.10**. Generally, the models allow us to significantly reduce the underestimation prediction, about 41%, and improve the within percentages in 19%, approximately, and significantly reduce the OR estimation error, but with the cost of having higher overestimation predictions. The goal will always be to improve the within predictions, and the results achieved in this project already represent significant improvements in the accurate predictions. “Any small improvement in the OR time prediction will be valuable for the service” as stated by the director of Urology at CHULC. When compare the MAE and RMSE between the hospital estimation of OR time and the ML models, it is possible to conclude that both RF and MLR outperforms

the current methods, by achieving significantly reduces in both error evaluation metrics, especially in RMSE. Although the positive model's results, it is important to highlight and have a critical sense regarding the OR time surgeon's prediction reported in CHULC database. In some cases, the reported data is not accurately filled in this database, as it was possible to notice in the clinical practice, therefore this error might be biased. However, it was the only available data in this project to compare the model's prediction with the CHULC current methods.

Although the work performed in increase the within OR predictions, it is important to discuss the impact and scenarios of underestimation and overestimation predictions. A model with higher underestimation predictions, will conduct to higher underestimation time, which will delay the succeeding surgeries, increase the percentages of cancellations and working hours for the OR staff. On the other hand, a model with higher overestimation predictions, will increase the free block time that might be used to respond to other hospital requirements. It is outside the scope of this thesis to estimate how the efficiency gain is traduced for both model scenarios, and future work should be developed.

Despite the positive prediction results, this study still had limitations. First, other crucial characteristics that might be retrieved for prediction, include the patient BMI, if the patient was submitted to radiotherapy procedures, the presence of diagnosis that influences the coagulation process, if there were previous surgeries, diabetes, hypertension, etc. are not contemplated in this study, since it was limited to the features of CHULC dataset. These were the standard metrics reported by the healthcare professionals addressed in this project. Although some of these, such as BMI and other comorbidities, might be reflected in the patient's ASA, this is a subjective metric. There is no concordance amongst the CHULC professionals if this is a reliable metric to measure the anesthesia risk. Hence these features would provide a clear report of the patient's condition from some professional's point of view.

7. Conclusion and Future Work

Waiting lists for surgery are one of the challenges for the different health systems. The availability of OR hours is often a gap, and it is not easy or even possible to increase the free OR time to allocate more surgeries. It is imperative to find and develop tools that enhance efficiency in terms of planning, as the one proposed in this project.

This study was a pilot project in CHULC to predict the OR time using a ML approach. In the literature, a lot of studies addressed this problem in healthcare systems such as the United States or outside Europe. In this project, the goal was to investigate if a ML approach was suitable for the problem of OR time prediction in a Portuguese hospital center, CHULC. First, we focus on understanding how this task was performed and the challenges in this hospital center across the three specialties under study. Starting from the challenge mentioned above, a strategy of co-creation with professionals, namely surgeons and anesthesiologists, was approached. Since in the development of these tools that intend to help day-to-day practice, it is essential to ensure the involvement of the users. This task was only possible with an intense fieldwork with an observation component and consultations with the stakeholders of this project, both CHULC surgeons, anesthesiologists, and administers. The involvement of these professionals was a fundamental and a core component in this project since the first moment, to validate the problem, then to draw a suitable framework that correspond to their needs, and simultaneously ensure the adherence and commitment to the project. This way it will be easier to ensure its application and usage, as they followed the development of the models.

The potential predictors were identified based on the literature along with the reunions with the professionals. The literature was also on the basis of model selection, and the importance of ensuring the transparency of the model for this healthcare problem was also a crucial point when the project framework was designed. A MLR and RF models were selected for this task. This project included RAS, which was a novel component, since this type of surgery is not commonly approached in the literature, only by a few studies. With the tendency of increasing RAS surgeries, this is a potential application of this type of studies.

It is conceivable to draw the conclusion that the dissertation's initial goals have been accomplished. The results achieved in this work show that the surgeon’s characteristics are the most impactful for the OR time prediction. Overall, it was not possible to conclude what is the best model for the Urology and General Surgery specialties, due to the similar evaluation metrics. For Orthopedics, the RF model was the only one able to model all the data. Compared with the current methods, the surgeon predictions, the ML models outperform with significantly reductions in the error of OR time estimations in all specialties and more accurate estimations, with a reduction in the underestimation time; however, the model overestimates the OR time compared with the surgeon’s predictions. A future direction of this work that might be helpful for the hospital’s administration is quantifying the costs and benefits of both under- and over- estimations for the OR. This analysis could allow the elaboration of different reality scenarios, and understand, in terms of costs, the best way to allocate the model’s predictions OR times.

The process of increasing OR efficiency is extensive and challenging. The prediction of OR time has been made in this thesis. However, there are still many areas that need more research. Additionally, more work must be performed to ensure the healthcare professionals' adhesion to this type of ML tool. Even though ML models have been proven to be more sophisticated and accurate in predicting OR times in the specialties of Orthopedics, General Surgery, and Urology in CHULC, steps still need

to be taken before applying ML models to a real medical system and ensuring adherence by the clinicians. The first point to ensuring adherence is to build confidence and show the scientific evidence of this type of AI system, construct a clarified discourse, get their involvement, and explain the potential benefits for the hospital and the patients. Then, a simple interface should be integrated into the EHR to automate the process of OR time prediction and support surgical scheduling. By creating this usability layer, the interface, the effective impact on surgical activity obtained through a more effective OR time estimation could be measured.

References

[1] Ayaad, O., Alloubani, A., ALhajaa, E. A., Farhan, M., Abuseif, S., al Hroub, A., & Akhu-Zaheya, L. (2019). The role of electronic medical records in improving the quality of health care services: Comparative study. *International Journal of Medical Informatics*, *127*, 63–67. https://doi.org/10.1016/j.ijmedinf.2019.04.014

[2] World Health Organization. (2021). *Ethics and governance of artificial intelligence for health: WHO guidance.* (pp. 6–15).

[3] Guerriero, F., & Guido, R. (2011). Operational research in the management of the operating theatre: a survey. *Health Care Management Science*, *14*(1), 89–114. https://doi.org/10.1007/s10729-010-9143-6

[4] Dexter, F., & Macario, A. (2004). When to Release Allocated Operating Room Time to Increase Operating Room Efficiency. *Anesthesia & Analgesia*, 758–762. https://doi.org/10.1213/01.ANE.0000100739.03919.26

[5] Lee, D. J., Ding, J., & Guzzo, T. J. (2019). Improving Operating Room Efficiency. *Current Urology Reports*, *20*(6), 28. https://doi.org/10.1007/s11934-019-0895-3

[6] Laskin, D. M., Abubaker, A. O., & Strauss, R. A. (2013). Accuracy of Predicting the Duration of a Surgical Operation. *Journal of Oral and Maxillofacial Surgery*, *71*(2), 446–447. https://doi.org/10.1016/j.joms.2012.10.009

[7] Arne Björnberg. (2012). *Euro Health Consumer Index 2012*. Health Consumer Powerhouse. (pp. 6–28).

[8] Pedro, G., & Luís, L. (2011). The SIGLIC system for improving the access to surgery in Portugal. *Electronic Journal Information Systems Evaluation*, *14*.

[9] Siciliani L, M. Borowitz, & M. Borowitz. (2013). *Waiting Time Policies in the Health Sector: What Works?* OECD Health Policy Studies

[10] Serviço Nacional de Saúde. (2021). *Relatório Anual: Acesso a cuidados de saúde nos estabelecimentos do SNS e entidades convencionadas.*

[11] Magalhães, T. (2022). O Papel da Tecnologia na Era da transformação Digital: Impacto do Conhecimento e da Tecnologia na Área da saúde. In *Transformação Digital em Saúde* (1st ed., pp. 45–55). Almedina.

[12] Aiken, L. H., Sermeus, W., van den Heede, K., Sloane, D. M., Busse, R., McKee, M., Bruyneel, L., Rafferty, A. M., Griffiths, P., Moreno-Casbas, M. T., Tishelman, C., Scott, A., Brzostek, T., Kinnunen, J., Schwendimann, R., Heinen, M., Zikos, D., Sjetne, I. S., Smith, H. L., & Kutney-Lee, A. (2012). Patient safety, satisfaction, and quality of hospital care: cross sectional

surveys of nurses and patients in 12 countries in Europe and the United States. *British Medical Journal*, *344*(2), e1717–e1717. https://doi.org/10.1136/bmj.e1717

[13] L'Heureux, A., Grolinger, K., Elyamany, H. F., & Capretz, M. A. M. (2017). Machine Learning With Big Data: Challenges and Approaches. *Institute of Electrical and Electronics Engineers*, *5*, 7776–7797. https://doi.org/10.1109/ACCESS.2017.2696365

[14] 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

[15] National confidential enquiry into patient outcome and death. (2004). *The NCEPOD classification of interventions*.

[16] Neary, W. D., Prytherch, D., Foy, C., Heather, B. P., & Earnshaw, J. J. (2007). Comparison of different methods of risk stratification in urgent and emergency surgery. *British Journal of Surgery*, *94*(10), 1300–1305. https://doi.org/10.1002/bjs.5809

[17] Clavel, D., Xie, X., Mahulea, C., & Silva, M. (2018). A Three Steps Approach for Surgery Planning of Elective and Urgent Patients. *International Federation of Automatic Control-PapersOnLine*, *51*(7), 243–250. https://doi.org/10.1016/j.ifacol.2018.06.308

[18] Fei, H., Meskens, N., & Chu, C. (2010). A planning and scheduling problem for an operating theatre using an open scheduling strategy. *Computers & Industrial Engineering*, *58*(2), 221–230. https://doi.org/10.1016/j.cie.2009.02.012

[19] Barbagallo, S., Corradi, L., de Ville de Goyet, J., Iannucci, M., Porro, I., Rosso, N., Tanfani, E., & Testi, A. (2015). Optimization and planning of operating theatre activities: an original definition of pathways and process modeling. *BMC Medical Informatics and Decision Making*, *15*(1), 38. https://doi.org/10.1186/s12911-015-0161-7

[20] van Veen-Berkx, E., Bitter, J., Elkhuizen, S. G., Buhre, W. F., Kalkman, C. J., Gooszen, H. G., & Kazemier, G. (2014). The influence of anesthesia-controlled time on operating room scheduling in Dutch university medical centres. *Canadian Journal of Anesthesia/Journal Canadien d'anesthésie*, *61*(6), 524–532. https://doi.org/10.1007/s12630-014-0134-9

[21] MILLS, A. (1990). The economics of hospitals in developing countries. Part I: expenditure patterns. *Health Policy and Planning*, *5*(2), 107–117. https://doi.org/10.1093/heapol/5.2.107

[22] Grupo de trabalho para a avaliação da situação nacional dos blocos operatórios. (2013). Avaliação da Situação Nacional dos Blocos Operatórios. *Ministério Da Saúde, Diário Da República* .

[23] Tuwatananurak, J. P., Zadeh, S., Xu, X., Vacanti, J. A., Fulton, W. R., Ehrenfeld, J. M., & Urman, R. D. (2019). Machine Learning Can Improve Estimation of Surgical Case Duration: A Pilot Study. *Journal of Medical Systems*, *43*(3), 44.

[24] Kour, H., & Gondhi, N. (2020). *Machine Learning Techniques: A Survey* (pp. 266–275). https://doi.org/10.1007/978-3-030-38040-3_31

[25] 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*, 106220. https://doi.org/10.1016/j.cmpb.2021.106220

[26] Jain, A., Patel, H., Nagalapatti, L., Gupta, N., Mehta, S., Guttula, S., Mujumdar, S., Afzal, S., Sharma Mittal, R., & Munigala, V. (2020). Overview and Importance of Data Quality for Machine Learning Tasks. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 3561–3562. https://doi.org/10.1145/3394486.3406477

[27] Condie, T., Mineiro, P., Polyzotis, N., & Weimer, M. (2013). Machine learning on Big Data. *2013 IEEE 29th International Conference on Data Engineering (ICDE)*, 1242–1244. https://doi.org/10.1109/ICDE.2013.6544913

[28] Jason Brownlee. (2020). *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python* (Machine Learning Mastery, Ed.).

[29] Roh, Y., Heo, G., & Whang, S. E. (2021). A Survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective. *IEEE Transactions on Knowledge and Data Engineering*, *33*(4), 1328–1347. https://doi.org/10.1109/TKDE.2019.2946162

[30] Kwak, S. K., & Kim, J. H. (2017). Statistical data preparation: management of missing values and outliers. *Korean Journal of Anesthesiology*, *70*(4), 407. https://doi.org/10.4097/kjae.2017.70.4.407

[31] Kang, M., & Tian, J. (2018). Machine Learning: Data Pre-processing. In *Prognostics and Health Management of Electronics* (pp. 111–130). John Wiley and Sons Ltd. https://doi.org/10.1002/9781119515326.ch5

[32] Ilyas, I. F., & Chu, X. (2019). Data transformation. In *Data Cleaning*. Association for Computing Machinery. https://doi.org/10.1145/3310205.3310210

[33] Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: A new perspective. *Neurocomputing*, *300*, 70–79. https://doi.org/10.1016/j.neucom.2017.11.077

[34] Burton, A. L. (2021). OLS (Linear) Regression. In *The Encyclopedia of Research Methods in Criminology and Criminal Justice* (pp. 509–514). Wiley. https://doi.org/10.1002/9781119111931.ch104

[35] The problem of multicollinearity. (n.d.). In *Understanding Regression Analysis* (pp. 176–180). Springer US. https://doi.org/10.1007/978-0-585-25657-3_37

[36] Bashir, D., Montañez, G. D., Sehra, S., Segura, P. S., & Lauw, J. (2020). *An Information-Theoretic Perspective on Overfitting and Underfitting* (pp. 347–358). https://doi.org/10.1007/978-3-030-64984-5_27

[37] Eberly, L. E. (2007). *Multiple Linear Regression* (pp. 165–187). https://doi.org/10.1007/978-1-59745-530-5_9

[38] Rigatti, S. J. (2017). Random Forest. *Journal of Insurance Medicine*, *47*(1), 31–39. https://doi.org/10.17849/insm-47-01-31-39.1

[39] Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250. https://doi.org/10.5194/gmd-7-1247-2014

[40] Naser, M. Z., & Alavi, A. H. (2021). Error Metrics and Performance Fitness Indicators for Artificial Intelligence and Machine Learning in Engineering and Sciences. *Architecture, Structures and Construction*. https://doi.org/10.1007/s44150-021-00015-8

[41] Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061

[42] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830.

[43] Olivares, M., Terwiesch, C., & Cassorla, L. (2008). Structural Estimation of the Newsvendor Model: An Application to Reserving Operating Room Time. *Management Science*, *54*(1), 41–55. https://doi.org/10.1287/mnsc.1070.0756

[44] Tankard, K., Acciavatti, T. D., Vacanti, J. C., Heydarpour, M., Beutler, S. S., Flanagan, H. L., & Urman, R. D. (2018). Contributors to Operating Room Underutilization and Implications for Hospital Administrators. *The Health Care Manager*, *37*(2), 118–128. https://doi.org/10.1097/HCM.0000000000000214

[45] Dexter, F., & Macario, A. (2004). When to Release Allocated Operating Room Time to Increase Operating Room Efficiency. *Anesthesia & Analgesia*, 758–762. https://doi.org/10.1213/01.ANE.0000100739.03919.26

[46] Devi, S. P., Rao, K. S., & Sangeetha, S. S. (2012). Prediction of Surgery Times and Scheduling of Operation Theaters in Optholmology Department. *Journal of Medical Systems*, *36*(2), 415–430. https://doi.org/10.1007/s10916-010-9486-z

[47] Strum, D. P., May, J. H., & Vargas, L. G. (2000). Modeling the Uncertainty of Surgical Procedure Times. *Anesthesiology*, *92*(4), 1160–1167. https://doi.org/10.1097/00000542-200004000-00035

[48] Edelman, E. R., van Kuijk, S. M. J., Hamaekers, A. E. W., de Korte, M. J. M., van Merode, G. G., & Buhre, W. F. F. A. (2017). Improving the Prediction of Total Surgical Procedure Time Using Linear Regression Modeling. *Frontiers in Medicine*, *4*. https://doi.org/10.3389/fmed.2017.00085

[49] Eijkemans, M. J. C., van Houdenhoven, M., Nguyen, T., Boersma, E., Steyerberg, E. W., & Kazemier, G. (2010). Predicting the Unpredictable. *Anesthesiology*, *112*(1), 41–49. https://doi.org/10.1097/ALN.0b013e3181c294c2

[50] 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

[51] 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. Annual Symposium Proceedings. AMIA Symposium*, *2015*, 640–648

[52] ShahabiKargar, Z., Khanna, S., Good, N., Sattar, A., Lind, J., & O'Dwyer, J. (2014). *Predicting Procedure Duration to Improve Scheduling of Elective Surgery* (pp. 998–1009). https://doi.org/10.1007/978-3-319-13560-1_86

[53] Huang, C.-C., Lai, J., Cho, D.-Y., & Yu, J. (2020). *A Machine Learning Study to Improve Surgical Case Duration Prediction*. https://doi.org/10.1101/2020.06.10.20127910

[54] 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), 346-354e3. https://doi.org/10.1016/j.jamcollsurg.2019.05.029

[55] Zhao, B., Waterman, R. S., Urman, R. D., & Gabriel, R. A. (2019). A Machine Learning Approach to Predicting Case Duration for Robot-Assisted Surgery. *Journal of Medical Systems*, *43*(2), 32. https://doi.org/10.1007/s10916-018-1151-y

[56] Fairley, M., Scheinker, D., & Brandeau, M. L. (2019). Improving the efficiency of the operating room environment with an optimization and machine learning model. *Health Care Management Science*, *22*(4), 756–767. https://doi.org/10.1007/s10729-018-9457-3

.

[57] Soh, K. W., Walker, C., O'Sullivan, M., Wallace, J., & Grayson, D. (2020). Case study of the prediction of elective surgery durations in a New Zealand teaching hospital. *The International Journal of Health Planning and Management*, *35*(6), 1593–1605. https://doi.org/10.1002/hpm.3046

[58] 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.

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

[60] de Cassai, A., Boscolo, A., Tonetti, T., Ban, I., & Ori, C. (2019). Assignment of ASA-physical status relates to anesthesiologists' experience: a survey-based national-study. *Korean Journal of Anesthesiology*, *72*(1), 53–59.

[61] Cartwright, D. J. (2013). ICD-9-CM to ICD-10-CM Codes: What? Why? How? *Advances in Wound Care*, *2*(10), 588–592. https://doi.org/10.1089/wound.2013.0478

[62] Southern, D. A., Norris, C. M., Quan, H., Shrive, F. M., Galbraith, P. D., Humphries, K., Gao, M., Knudtson, M. L., & Ghali, W. A. (2008). An administrative data merging solution for dealing with missing data in a clinical registry: adaptation from ICD-9 to ICD-10. *BMC Medical Research Methodology*, *8*(1), 1. https://doi.org/10.1186/1471-2288-8-1

[63] Ghasemi, A., & Zahediasl, S. (2012). Normality tests for statistical analysis: a guide for non-statisticians. *International Journal of Endocrinology and Metabolism*, *10*(2), 486–489. https://doi.org/10.5812/ijem.3505

[64] Adadi, A., & Berrada, M. (2020). *Explainable AI for Healthcare: From Black Box to Interpretable Models* (pp. 327–337). https://doi.org/10.1007/978-981-15-0947-6_31

Appendix

Table A1- Group of Diagnosis for ICD-9 and ICD-10 codes.

Group of Diagnosis Description	ICD-9 codes	ICD-10 codes
Certain infectious and parasitic diseases	001-139	A00-B99
Neoplasms	140-239	C00-D49
Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	280-289	D50-D89
Endocrine, nutritional and metabolic diseases	240-279	E00-E89
Mental, Behavioral and Neurodevelopmental disorders	290-319	F01-F99
Diseases of the nervous system	320-389	G00-G99
Diseases of the eye and adnexa	_____	H00-H59
Diseases of the ear and mastoid process	_____	H60-H95
Diseases of the circulatory system	390-459	I00-I99
Diseases of the respiratory system	460-519	J00-J99
Diseases of the digestive system	520-579	K00-K95
Diseases of the skin and subcutaneous tissue	680-709	L00-L99
Diseases of the musculoskeletal system and connective tissue	710-739	M00-M99
Diseases of the genitourinary system	580-629	N00-N99
Pregnancy, childbirth and the puerperium	630-679	O00-O9A
Certain conditions originating in the perinatal period	760-779	P00-P96
Congenital malformations, deformations and chromosomal abnormalities	740-759	Q00-Q99
Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	780-799	R00-R99
Injury, poisoning and certain other consequences of external causes	E800-E999	S00-T88
Codes for special purposes	_____	U00-U85
External causes of morbidity	_____	V00-Y99
Factors influencing health status and contact with health services	V01-V89	Z00-Z99