TÉCNICO LISBOA

Clinical Prognosis and Risk Prediction of Postoperative Complications in Cancer Patients

Daniel Mateus Gonçalves

Thesis to obtain the Master of Science Degree in

Information Systems and Computer Engineering

Supervisors: Prof. Rafael Sousa Costa
Prof. Rui Miguel Carrasqueiro Henriques

Examination Committee

Chairperson: Prof. José Luís Brinquete Borbinha
Supervisor: Prof. Rafael Sousa Costa
Member of the Committee: Prof. Maria da Conceição Esperança Amado

November 2020

Acknowledgments

After a long year of work, my master's thesis has come to an end, although this project is still under development and I will continue to support it during the coming phase. Looking back, I remember the people without whom this achievement would not be possible, and to whom a mention is due.

Firstly, I would like to thank IPO Porto for providing me with the privilege of working on this project. Solving real-world problems in an area that is so dear to me. Especially, I would like to thank Dr. Lúcio Santos and his team, for their availability to help us, and also for sharing their knowledge and vision of what the future of prognostication might look like.

Secondly, I would like to express my deep gratitude to both my supervisors, Prof. Rafael Costa and Prof. Rui Henriques, for their knowledge and support. In the past year, almost every week, they spared their time to help me with whatever they could. From answering my questions and doubts, to reviewing my reports and articles. With both of you, I have learned more than I could think of, about data science and the world of scientific investigation. Could not have asked for better.

My family has also played a very important role in this entire process. Their support was critical to my well being and academic success. I cannot repay them for all they gave me, so I will always reserve some of my time as a token of my eternal appreciation.

My good old friends are also deserving of an honorable mention. Their words were uplifting in times of discouragement. We might not see each other for extended periods, but I know that my friends will be there for me, as I will do for them.

My girlfriend Ana has also shared some of my pains and joys relating to this work. Through the days, she was always open to hear me out and try to understand and support my work. I appreciate you for taking care of me and supporting my dreams, no matter what.

Lastly, to everyone that contributed to the realization of this work, and I might be forgetting to mention: Thank you!

This work was supported by FCT, through IDMEC, under LAETA, project UIDB/50022/2020 and project IPOscore DSAIPA/DS/0042/2018, and INESC-ID pluriannual (UIDB/50021/2020).

Abstract

Postoperative complications of cancer surgery are still hard to predict, although there are risk scores intended to make such predictions. They vary with regards to their outcome, surgical cohort, or type of predictive model. The differences among studies, contribute for the creation of highly specialized tools, with poor reusability in foreign contexts. Adaptability to different surgical domains and populations can add to larger errors, since often these studies are developed in carefully selected surgical cohorts. Today, new techniques have been proposed to create potentially more powerful and accurate predictors, capable of modeling high dimensional data and its inherent complexities. This thesis aims to study and predict postoperative complications risk for cancer patients, offering two major contributions. First, to develop a risk calculator using machine learning models, with 4 outcomes of interest: existence of postoperative complications, severity level of said complications, death probability within 1 year, and number of days spent in the intermediate care unit. Second, to support the study of this disease with relevant findings and improve the interpretability of predictive models, especially associative models by extending tree representations to capture measures of generalization ability. As a result, we provide a set of models with reliable guarantees of predictive performance and offer new perspectives and insights into the decision process. Postoperative complications can be predicted with 68% accuracy, complications' severity can be predicted with MAE = 1.56, the days in the ICU can be predicted with MAE = 1.04, and 1 year death can be predicted with 75% accuracy. The proposed predictive models yield statistically significant improvements against their respective baseline models (p-value < 0.01).

Keywords

postoperative complications; risk prediction; cancer; machine learning; clinical data modeling.

Resumo

As complicações pós-operatórias decorrentes de cirurgias oncológicas são difíceis de prever, embora existam calculadoras de risco para o efeito. Estas variam no objetivo, área cirúrgica ou tipo de modelo preditivo. As diferenças entre os estudos, contribuem para a criação de ferramentas altamente especializadas, mas com pouca reusabilidade. A adaptação a diferentes domínios cirúrgicos e populações aumenta o erro, já que muitas vezes esses estudos são desenvolvidos em grupos cirúrgicos limitados. Hoje, existem novas técnicas para criar modelos mais poderosos, capazes de modelar dados de elevada dimensionalidade e as suas complexidades inerentes. Esta tese tem como objetivo estudar e prever o risco de complicações pós-operatórias em pacientes com cancro, oferecendo duas contribuições principais. A primeira, uma calculadora de risco que utiliza modelos de aprendizagem automática, para prever 4 objetivos: existência de complicações pós-operatórias, grau da sua severidade, probabilidade de morte no espaço de 1 ano e o número de dias de internamento na unidade de cuidados intermédios. Em segundo lugar, apoiar o estudo desta doença e melhorar a interpretabilidade dos modelos preditivos. Como resultado, é fornecido um conjunto de modelos com garantias desempenho e novas perspectivas quanto ao processo de decisão. As complicações pós-operatórias podem ser previstas com uma precisão de 68%, a gravidade das complicações pode ser prevista com MAE = 1.56, os dias na UCI podem ser previstos com MAE = 1.04 e a morte no espaço de 1 ano pode ser prevista com precisão de 75%. Os modelos preditivos propostos produzem resultados com significância estatística em relação aos seus respectivos baselines (p-value < 0.01).

Palavras Chave

complicações pós-operatórias; previsão de risco; cancro; aprendizagem automática; modelação de dados clínicos.

Contents

1 Introduction **2**

1.1 Project Introduction . . . 3

1.2 Contributions . . . 5

1.3 Document Organization . . . 6

2 Background **7**

2.1 Cancer . . . 8

2.2 Prognostication . . . 8

2.3 Predictive Models . . . 8

2.3.1 Assessment and Validation . . . 9

3 Related Work **10**

3.1 Traditional Statistical Studies . . . 11

3.1.1 Traditional Models . . . 13

3.2 Machine Learning Studies . . . 14

3.2.1 Machine Learning Models . . . 16

3.3 Cohort Data Preprocessing . . . 19

3.3.1 Missing Values . . . 20

3.3.2 Outcome Class Imbalance . . . 20

3.3.3 High Dimensionality . . . 21

3.4 Assessment . . . 21

3.4.1 Error metrics . . . 23

3.5 Validation . . . 24

4 Methodology **25**

4.1 The Dataset . . . 26

4.1.1 Data Profiling . . . 26

4.1.2 Data Exploration . . . 27

4.2 Preprocessing . . . 28

4.2.1 Missing Values . . . 28

4.2.2 Categorical Variable Encoding . . . 29
4.2.3 Imbalanced Variables . . . 29
4.2.4 Resampling . . . 29
4.2.5 Feature Scaling . . . 30
4.2.6 Feature Selection . . . 30
4.3 Prediction Models . . . 32
4.3.1 Prediction of Postoperative Complication . . . 33
4.3.2 Prediction of Complication Severity . . . 33
4.3.3 Prediction of Death Probability Within 1 Year . . . 34
4.3.4 Prediction of Days Spent in the ICU . . . 34
4.3.5 Model Tuning: Hyperparameter Optimization . . . 35
4.4 Assessment and Validation . . . 36
4.4.1 Classification Evaluation Metrics . . . 36
4.4.2 Regression Evaluation Metrics . . . 38
4.4.3 Model Validation . . . 39
4.4.4 Model Comparison . . . 40
4.5 Implementation Details . . . 41

5 Results **42**

5.1 Exploration and Profiling . . . 43
5.1.1 Data Profiling . . . 43
5.1.2 Discriminative Factors of Postoperative Complications . . . 45
5.2 Model Results . . . 47
5.2.1 Preliminary Study . . . 48
5.2.2 Resampling . . . 51
5.2.3 Feature Scaling . . . 53
5.2.4 Hyperparameter Optimization & Feature Selection . . . 57
5.2.4.1 No Feature Selection . . . 58
5.2.4.2 Feature Selection with p-value of 0.1 . . . 60
5.2.4.3 Feature Selection with p-value of 0.0001 . . . 64
5.3 Associative Models - In Depth Analysis . . . 67
5.3.1 Feature Importance in Associative Models . . . 69

6 Discussion **71**

6.1 Existence of Postoperative Complications . . . 72
6.2 Severity of Complications . . . 74
6.2.1 Classification Approach . . . 74

6.2.2 Regression Approach . . . 75
6.2.3 Approach Comparison . . . 77
6.3 Days Spent in the ICU . . . 78
6.4 Death Probability Within 1 Year . . . 79
7 Conclusion **81**
7.1 Conclusions . . . 82
7.2 Limitations and Future Work . . . 82
7.3 Scientific Communication . . . 83
A Appendix - Selected Features **88**
A.1 Full Set of Variables . . . 88
A.2 First Feature Selection Stage (p-value = 0.1) . . . 89
A.2.0.1 Feature ranking for the output “Days in the ICU”: . . . 89
A.2.0.2 Feature ranking for the output “Existence of Postoperative Complications”: 89
A.2.0.3 Feature ranking for the output “Complications Severity”: . . . 90
A.2.0.4 Feature ranking for the output “1 Year Death”: . . . 90
A.3 Second Feature Selection Stage (p-value = 0.0001) . . . 90
A.3.0.1 Feature ranking for the output “Days in the ICU”: . . . 90
A.3.0.2 Feature ranking for the output “Existence of Postoperative Complications”: 91
A.3.0.3 Feature ranking for the output “Complications’ Severity”: . . . 91
A.3.0.4 Feature ranking for the output “1 Year Death”: . . . 91
B Appendix - Models’ Hyperparameters **92**
B.1 Optimized Set - No Feature Selection . . . 92
B.1.1 Classification - Existence of Complications . . . 92
B.1.2 Classification - Complications’ Severity . . . 93
B.1.3 Classification - 1 Year Death . . . 93
B.1.4 Regression - Complications’ Severity . . . 94
B.1.5 Regression - Days in the ICU . . . 94
B.2 Optimized Set - p-value = 0.0001 Feature Selection . . . 95
B.2.1 Classification - Existence of Complications . . . 95
B.2.2 Classification - Complications’ Severity . . . 96
B.2.3 Classification - 1 Year Death . . . 96
B.2.4 Regression - Complications’ Severity . . . 97
B.2.5 Regression - Days in the ICU . . . 97
B.3 Optimized Set - p-value = 0.1 Feature Selection . . . 98
B.3.1 Classification - Existence of Complications . . . 98

B.3.2 Classification - Complications' Severity 98
B.3.3 Classification - 1 Year Death 99
B.3.4 Regression - Complications' Severity 99
B.3.5 Regression - Days in the ICU 100

List of Figures

1.1 Search results[2]for the keywords “machine learning postoperative complication cancer” . . 4
1.2 Project’s global workflow 5

2.1 Tabular dataset example - general case 9

4.1 Feature selection methods, according to variable and output type 31
4.2 Error/confusion matrix representation with labels 37
4.3 Receiver Operating Characteristic (ROC) chart - prediction complications - (Support Vector Machine (SVM) 5th stage) 38

5.1 Numeric feature analysis sample 44
5.2 Value distribution analysis for severity level 44
5.3 Presence of connective tissue disease variable distribution 45
5.4 Surgical specialty variable distribution 45
5.5 Gender variable distribution according to the existence of complications 46
5.6 Percentage of complications by gender 46
5.7 Emergency variable distribution according to the existence of complications 46
5.8 Percentage of complications for emergencies vs. non emergencies 46
5.9 BMI variable distribution according to the existence of complications 47
5.10 Percentage of complications by BMI level 47
5.11 The ten most common cancer types ranked by postoperative complication rate 47
5.12 The 6 stages of the model development pipeline 48
5.13 Graph with primary results for the Existence of Complications prediction 49
5.14 Graph with primary results for the Severity of Complications prediction 49
5.15 Graph with primary results for the Severity of Complications (regression) prediction 50
5.16 Graph with primary results for the Severity of Complications (discretized) prediction 50
5.17 Graph with primary results for Death Within 1 Year prediction 51
5.18 Graph with primary results for Days in the ICU prediction 51

5.19 Improvement after resampling for the prediction of complications 52
5.20 Improvement after resampling for complications' severity prediction 53
5.21 Improvement after resampling for the 1 year death prediction 53
5.22 Improvement after normalization for complications prediction 54
5.23 Improvement after normalization for the prediction of complications' severity 55
5.24 Improvement after normalization for the complications's severity (regression) prediction . 55
5.25 Improvement after normalization for the complication's severity (discretized) prediction . . 56
5.26 Improvement after normalization for 1 year death prediction 56
5.27 Improvement after normalization for the days in the ICU prediction 57
5.28 Improvement after hyperparametrization for complications prediction 58
5.29 Improvement after hyperparametrization for complications' severity prediction 58
5.30 Improvement after hyperparametrization for complications' severity (regression) prediction 59
5.31 Improvement after hyperparametrization for complications' severity (discretized) prediction 59
5.32 Improvement after hyperparametrization for 1 year death prediction 60
5.33 Improvement after hyperparametrization for days in the ICU prediction 60
5.34 Improvement after feature selection (p-value=0.1) and hyperparameter optimization for complications prediction 61
5.35 Improvement after feature selection (p-value=0.1) and hyperparameter optimization for complications' severity prediction 61
5.36 Improvement after feature selection (p-value=0.1) and hyperparameter optimization for complications' severity (regression) prediction 62
5.37 Improvement after feature selection (p-value=0.1) and hyperparameter optimization for complications' severity (discretized) prediction 62
5.38 Improvement after feature selection (p-value=0.1) and hyperparameter optimization for 1 year death prediction 63
5.39 Improvement after feature selection (p-value=0.1) and hyperparameter optimization for the days in the ICU prediction 63
5.40 Improvement after feature selection (p-value=0.0001) and hyperparameter optimization for complications prediction 64
5.41 Improvement after feature selection (p-value=0.0001) and hyperparameter optimization for complications' severity prediction 65
5.42 Improvement after feature selection (p-value=0.0001) and hyperparameter optimization for complications' severity (regression) prediction 65
5.43 Improvement after feature selection (p-value=0.0001) and hyperparameter optimization for complications' severity (discretized) prediction 66

5.44 Improvement after feature selection (p-value=0.0001) and hyperparameter optimization for 1 year death prediction . . . 66
5.45 Improvement after feature selection (p-value=0.0001) and hyperparameter optimization for the days in the ICU prediction . . . 67
5.46 Color scheme used for leaf error representation . . . 68
5.47 Tree graph for the DT from stage 6 used to classify complications' severity . . . 68
5.48 Tree graph for the DT from stage 6 used for complications' severity regression . . . 69
5.49 Feature importance for the Severity of Complications . . . 70
5.50 Feature importance for the Severity of Complications (regression) . . . 70

6.1 Existence of Complications - Global Accuracy . . . 73
6.2 Existence of Complications - Global AUC . . . 73
6.3 Existence of Complications - Global Recall . . . 73
6.4 Existence of Complications - Global Kappa . . . 73
6.5 Severity of Complications - Global Accuracy . . . 74
6.6 Severity of Complications - Global AUC . . . 74
6.7 Severity of Complications - Global Recall . . . 74
6.8 Severity of Complications - Global Kappa . . . 74
6.9 Severity of Complications - Global MAE . . . 75
6.10 Severity of Complications - Global RMSE . . . 75
6.11 Severity of Complications - Global R^2 . . . 76
6.12 Severity of Complications - Global Accuracy . . . 76
6.13 Severity of Complications - Global Recall . . . 76
6.14 Severity of Complications - Global Kappa . . . 76
6.15 Days in the ICU - Global MAE . . . 78
6.16 Days in the ICU - Global RMSE . . . 78
6.17 Days in the ICU - Global R^2 . . . 78
6.18 Death Within 1 Year - Global Accuracy . . . 80
6.19 Death Within 1 Year - Global AUC . . . 80
6.20 Death Within 1 Year - Global Recall . . . 80
6.21 Death Within 1 Year - Global Kappa . . . 80

List of Tables

3.1 Review of traditional statistical studies in postoperative prognostics (chronological order) . 12
3.2 Review of machine learning studies in postoperative prognostics (chronological order) . . 15

5.1 Primary results for the Existence of Complications prediction 49
5.2 Primary results for the Severity of Complications prediction 49
5.3 Primary results for the Severity of Complications (regression) prediction 50
5.4 Primary results for the Severity of Complications (discretized) prediction 50
5.5 Primary results for Death Within 1 Year prediction 51
5.6 Primary results for Days in the ICU prediction 51
5.7 Results after resampling for the prediction of complications 52
5.8 Results after resampling for complications' severity prediction 53
5.9 Results after resampling for the 1 year death prediction 53
5.10 Results after normalization for complications prediction 54
5.11 Results after normalization for the prediction of complications' severity 55
5.12 Results after normalization for the complications's severity (regression) prediction 55
5.13 Results after normalization for the complication's severity (discretized) prediction 56
5.14 Results after normalization for 1 year death prediction 56
5.15 Results after normalization for the days in the ICU prediction 57
5.16 Results after hyperparametrization for complications prediction 58
5.17 Results after hyperparametrization for complications' severity prediction 58
5.18 Results after hyperparametrization for complications' severity (regression) prediction . . . 59
5.19 Results after hyperparametrization for complications' severity (discretized) prediction . . . 59
5.20 Results after hyperparametrization for 1 year death prediction 60
5.21 Results after hyperparametrization for days in the ICU prediction 60
5.22 Results after feature selection (p-value=0.1) and hyperparameter optimization for complications prediction 61

5.23 Results after feature selection (p-value=0.1) and hyperparameter optimization for complications' severity prediction . . . 61
5.24 Results after feature selection (p-value=0.1) and hyperparameter optimization for complications' severity (regression) prediction . . . 62
5.25 Results after feature selection (p-value=0.1) and hyperparameter optimization for complications' severity (discretized) prediction . . . 62
5.26 Results after feature selection (p-value=0.1) and hyperparameter optimization for 1 year death prediction . . . 63
5.27 Results after feature selection (p-value=0.1) and hyperparameter optimization for the days in the ICU prediction . . . 63
5.28 Results after feature selection (p-value=0.0001) and hyperparameter optimization for complications prediction . . . 64
5.29 Results after feature selection (p-value=0.0001) and hyperparameter optimization for complications' severity prediction . . . 65
5.30 Results after feature selection (p-value=0.0001) and hyperparameter optimization for complications' severity (regression) prediction . . . 65
5.31 Results after feature selection (p-value=0.0001) and hyperparameter optimization for complications' severity (discretized) prediction . . . 66
5.32 Results after feature selection (p-value=0.0001) and hyperparameter optimization for 1 year death prediction . . . 66
5.33 Results after feature selection (p-value=0.0001) and hyperparameter optimization for the days in the ICU prediction . . . 67

6.1 5 best models for the existence of complications prediction . . . 73
6.2 5 best models for complication's severity prediction . . . 75
6.3 5 best models for complication's severity (regression) prediction . . . 77
6.4 5 best models for severity prediction (approach comparison) . . . 77
6.5 5 best models for days spent in the ICU prediction . . . 79
6.6 5 best models for 1 year death prediction . . . 80

Acronyms

IPO	Instituto Português de Oncologia
ICU	Intermediate Care Unit
ML	Machine Learning
NN	Neural Network
DT	Decision Tree
kNN	k-Nearest Neighbours
RF	Random Forest
NB	Naive Bayes
SVM	Support Vector Machine
XGB	Extreme Gradient Boosting
LR	Logistic Regression
MLP	Multilayer Perceptron
MLPR	Multilayer Perceptron Regressor
PLS	Partial Least Squares
SVR	Support Vector Regression
RRF	Reciprocal Rank Fusion
AUC	Area Under the Curve
MAE	Mean Absolute Error
RMSE	Root Mean Squared Error
SMOTEENN	Synthetic Minority Oversampling Technique and Edited Nearest Neighbour
BMI	Body Mass Index
ANOVA	Analysis of Variance
ACS	American College of Surgeons
NSQIP	National Surgical Quality Improvement Program
ARISCAT	Assess Respiratory Risk in Surgical Patients in Catalonia
ROC	Receiver Operating Characteristic
TPR	True Positive Rate
FPR	False Positive Rate
POSSUM	Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity
CARE	Cardiac Anesthesia Risk Evaluation

1

Introduction

Contents

1.1 Project Introduction 3

1.2 Contributions 5

1.3 Document Organization 6

Cancer is a major health problem worldwide and it's among the leading death causes of the 21st century. In the United States, during 2019, the number of new cases of cancer is estimated to be close to 1,800,000 and the deaths due to cancer should hit close to 600,000. The survival rate within 5 years for patients with cancer is currently around 65% and has steadily been improving over the last century [53].

In the Mortality Dashboard[1] by DGS, cancer is the main cause of premature death (under 70 years) in Portugal. In 2018, 10,022 people died prematurely due to cancer, in a year with 23,267 total deaths.

1.1 Project Introduction

There are at least two battlefronts in trying to reduce deaths associated to cancer, which can be a result from direct consequences of the disease, or occur due to operative and postoperative complications resulting from surgery for cancer treatment. These complications contribute to lower survival probability and, in certain types of cancer, to aggravate the recurrence rate [40, 1, 9, 42]. The outcome of such surgeries is still widely unpredictable due to the huge number of factors involved. Postoperative risk assessment tools are already available, not only for cancer patients but for surgery in general, with the aim of reducing mortality and morbidity rates [63].

With the advancements of technology and areas like data science, new techniques and better resources are available, while big clinical data is also growing. In the recent years there has been an increasing amount of studies aimed at identifying the main factors for postoperative complications and, considering these factors, developing risk assessment calculators [63]. The predictions given by these calculators help doctors and patients in surgery decision-making. From a clinical perspective, the risk scores are determinant in choosing the course of actions, such as additional testing, prehabilitation or supportive measures, to be taken during the preoperative, intraoperative and postoperative periods [63].

The primary objective of this project is to develop a risk score that is able to predict 4 outcomes of interest: existence of postoperative complications, the severity of said complications, the number of days spent in the ICU and the probability of death within 1 year after surgery, in cancer patients. Secondly, this project also aims to support the study of this disease and surgical prognostication, either by finding relevant variables, or improving the interpretability of these models. Being a typical data science project, the dataset in use becomes the centerpiece of all work. In this case, a clinical dataset with more than 800 patients and 100 attributes is available.

This project is being developed with IPO Porto (Dr. Lúcio L. Santos), which is a Portuguese hospital and research center specialized in the oncology branch of medicine. For decades, patients with cancer have been doing their treatments and surgery at IPO Porto and there has always been a need to make predictions about their postoperative state. For that reason, there is currently a joint effort between

[1] https://www.sns.gov.pt/noticias/2019/11/11/dgs-plataforma-da-mortalidade/

IPO Porto and IDMEC to develop a new postoperative risk scoring system specific for the Portuguese population, using machine learning models.

The existing solution at IPO Porto relies on already available and well known scores and calculators [12, 49, 10, 5] that could predict such information to a certain degree, but the results don't match reality and are sometimes discordant among each other, according to professionals. Generally, risk scores are specific for certain types of complications or surgical area, forcing the need for multiple specialized tools. All of these scores, reviewed in the next section, were also developed in international context and are commonly biased, not accounting for geographic variations in sociodemographic and lifestyle factors, medical settings, and cancer screening behaviors, which are all factors that are very hard to track and register. The models can only be so good as the datasets available and at the moment trying to adapt a foreign risk score will most certainly introduce some error in the predictions [20].

Despite the inherent flaws in data and current risk assessment tools, advances in machine learning and increasingly available data are radically changing the world's social landscape, including medicine and clinical research. Big Clinical Data is a growing phenomenon but for now there is still little public availability due to confidentiality reasons [16]. The data is also scattered among its owners, although things are starting to change and there are governments investing in large national clinical databases [57]. The use of machine learning models is also an approach that is still getting traction. Traditional statistical models have been delivering good results since the last century and the majority of the models reviewed in this project are making use of such methods. The use of machine learning models for cancer surgery prognostication is slowly starting to be tested. The number of papers related to the matter has steadily been increasing, as seen in Fig.1.1.

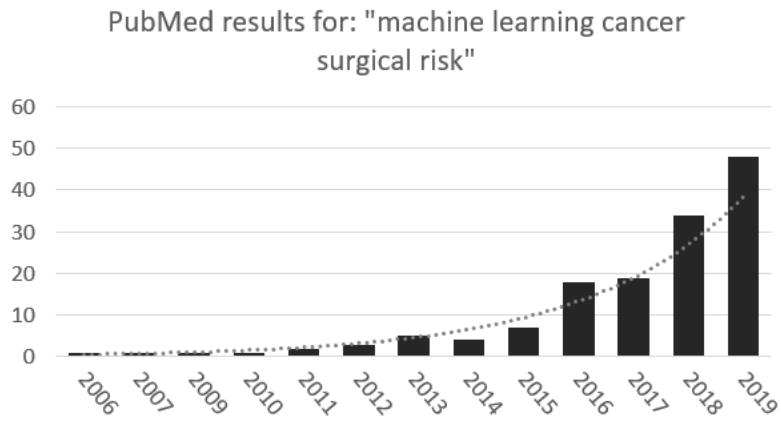


Figure 1.1: Search results[2] for the keywords "machine learning postoperative complication cancer"

[2] Obtained from PubMed, as of December 15th 2019:
https://pubmed.ncbi.nlm.nih.gov/?term=machine+learning+postoperative+complication+cancer

1.2 Contributions

In the context of this project, multiple contributions were proposed along the various phases of a data science challenge. The following list covers the steps and contributions of this project.

- Comprehensive **literature review** on the topic of cancer prognostication using predictive tools (including models, types of data, preprocessing, inputs, outputs, assessment and validation);
- **Dataset exploration** to get familiar with the data, which included detailed data profiling tasks;
- Definition of **target outcomes**, which are the primary goal. The outcomes include:
  - Existence of postoperative complications (y/n);
  - Severity level of said complications (according to the Clavien-Dindo scale);
  - Death probability within 1 year;
  - Number of days spent in the Intermediate Care Unit (ICU).
- **Data preprocessing** was applied to solve issues like missing values, manual inputs and other types of inconsistencies;
- **Application** of the predictive machine learning models;
- **Feature Selection**, in order to reduce training time, overfitting and maintain a relevant set of features;
- Model **optimization** process, by tuning the respective hyperparameters;
- Graphical representation of associative models, to study their decision process and error, improving interpretability;
- Finally, **assessment and validation** of the results, to evaluate the reliability and quality of the solution.

Figure 1.2: Project's global workflow

1.3 Document Organization

This thesis is is organized as follows:

Chapter 1 makes an introduction to cancer surgery prognostication and contextualizes the project backing this thesis.

Then, chapter 2 follows, offering quick access to some of the major concepts and terms relating to this project, including: cancer, surgical prognostication, prediction models.

Chapter 3 consists on a literature review, mainly covering scientific papers resulting from other scores created for the same objective, some are even in use at Instituto Português de Oncologia (IPO)-Porto. This review includes aspects like the outcomes, study cohorts, the models, the evaluation and validation process.

Chapter 4 presents the solution chosen to tackle the challenges of this project. From the data processing used, going through the model development and optimization processes, ending with the evaluation methods.

The results are shown in chapter 5. This chapter is divided in several sections, each containing the results for each stage of the project development. This approach was followed to make tracking progress easier and also to help establish comparisons in the next chapter.

Chapter 6 gives a global review over the entire progress achieved, while also commenting on the final results. As an indication, the 5 top models (according to a specific rank) for each prediction of interest are also shown.

Lastly, chapter 7 presents the final considerations about this project, as well as a list of limitations and aspects which could be further developed in future iterations.

2

Background

Contents

2.1 Cancer 8

2.2 Prognostication 8

2.3 Predictive Models 8

2.1 Cancer

During the last decades **cancer** has become one of the most devastating diseases worldwide [53]. Cancer is caused by genetic and non-genetic changes induced by environmental factors that lead to the activation or inactivation of specific genes leading to neoplastic transformations, or abnormal cell growth. Cancer is therefore a generic term used to classify a group of diseases that occur when an abnormal cell growth develops in one or more organs. Cancer can develop at any stage in life, in any organ, and no two cases are exactly alike. The diagnosis and treatment vary significantly with the type of cancer and patient. For solid tumours, in a relatively early stage, surgery is the standard option for treatment, often combined with radiation therapy or systemic therapy. The last two therapies become prevalent in situations where cancer has evolved into a metastatic stage, affecting not only the primary site but also secondary ones across the body of the patient, reducing the effectiveness of surgery [36].

2.2 Prognostication

The course of the preoperative, intraoperative and postoperative periods can all be determinant in the **postoperative complications** that a patient might suffer. These complications are proven to be related with inferior survival rate and higher recurrence rates and for that reason it is very important to have information that can somehow predict such complications [42, 9, 1].

Surgical prognosis has for a long time been a subject of investigation. Being able to predict what will be the state of a patient after a certain procedure has always been important and the methods used to do so have been improving [52]. From using medical intuition into using **decision support systems** relying on risk scores that can calculate the probability of complications, their type, severity and other relevant metrics.

2.3 Predictive Models

The task of predicting outcomes is commonly associated to some type of **data modeling**, which based on previous data tries to learn a trend, a decision boundary, or rules that can later be used to make predictions about new instances. Nowadays **risk score** models make use of extensive, very detailed datasets and are capable of modeling the dimensionality and complexity of problems that were previously addressed by more rudimentary solutions. **Machine Learning** has offered the possibility of modeling much more complex problems and also the ability to learn and improve through time [35].

A **model** can only be so good as the dataset used. Datasets are sets of data registries, that can be more or less organized, that are used by models to train and learn to predict a certain outcome. In the

context of this project, a **tabular dataset** (simple multivariate data) is being used. The data is stored in a tabular data structure composed of multiple observations, each having a set of values corresponding to the variables (e.g. variables a, b, z from Fig.2.1), also called features.

$$\begin{gathered}((a_1, b_1, ..., z_1),\\ (a_2, b_2, ..., z_2),\\ \vdots\\ (a_n, b_n, ..., z_n))\end{gathered}$$

Figure 2.1: Tabular dataset example - general case

There are close to 850 patients and 130 attributes. The dataset includes records of more than 30 types of cancer, with some being more representative than others. There's also the possibility of a future extension to the dataset including molecular, nutritional and physiological data.

There are usually two types of approaches to data modeling continuous or discrete. **Classification** is the prediction of a well defined discrete set of outcomes, it is the process of labeling a certain unknown entry. For instance one can predict the existence of complications (yes or no), or the type of cancer (e.g. intestine, lung or brain), using a classification model.

For problems where a continuous model is needed to predict a numeric outcome, **regression** models are used, generally relying on statistical processes to estimate the relationship between a set of dependent variables and one or more independent variables.

2.3.1 Assessment and Validation

Result assessment is a process that measures the predictive performance of the models. Different metrics can be used to make these measurements, and some might have pitfalls. So it is not unusual to employ multiple metrics simultaneously to evaluate a model, offering different perspectives to the errors of the predictions. These results can also be used to compare implementations and establish objective functions for performance optimization.

The process of **validating a model** evaluates its performance stability across a range of different samples, or in other terms the ability of the model to generalize its knowledge for unseen data. This process can be done by cross validation using the same dataset, split into training and testing set. Not just once, but multiple times so the model can repeat the process in different subsets of the data, assuring reliability. Another option, is using a new set of data, foreign to the dataset.

3

Related Work

Contents

3.1 Traditional Statistical Studies 11
3.2 Machine Learning Studies 14
3.3 Cohort Data Preprocessing 19
3.4 Assessment 21
3.5 Validation 24

Note: This chapter is based on the review article "Predicting postoperative complications in cancer patients: a survey bridging classical and machine learning contributions to post-surgical risk analysis" (submitted) - more details in section 7.3.

Prognostication tools are in constant improvement. The first studies date back to the 1940's and since then many publications have been made. In this section, we'll be focusing on two main predictions of interest in order to get prognostic information, morbidity and mortality, which are strongly correlated between each other and the existence of postoperative complications.

3.1 Traditional Statistical Studies

One of the most important factors when deciding what surgical treatment is viable for a patient is the risk of possible postoperative complications as well as the chances of survival for a certain patient, which in the end might be strongly connected after all. Being able to predict such outcomes is of crucial importance, creating opportunity for the consideration of alternative therapies or procedures, adequate intensive care, or even assisted life ending options.

Specifically in this area of prediction, there are a lot of studies which have made their way into professional clinical use, and were adopted by hospitals to support medical decision. Most of these clinically adopted scores, indexes and calculators are based on statistical methods, which so far have been reliable and don't suffer of the same degree of distrust that machine learning methods are still struggling with even today, due to the unfamiliarity and black-box character associated to them. Table 3.1 presents the list of all the traditional statistical studies for postoperative prognostic, reviewed in this section.

Cohort-outcome relationship - The monitored population is a determinant factor of each conducted study. The cohort is many times associated to the context of creation of the score and should be closely tied with the outcome predictions. For example, the Physiological and Operative Severity Score for the enUmeration of Mortality and morbidity (POSSUM) score was created from a general surgery cohort [17]. As such, it is very broad and it is highly acceptable that the model is well capable of roughly predicting mortality risk in a general surgery context. In the same line of thought, the Cardiac Anesthesia Risk Evaluation (CARE) score was developed in a cardiac surgery cohort [23]. Being more specific, it makes sense that its predictions for in-hospital death and morbidity are also more adequate to be applied in patients from the same context. Although extrapolation is possible, further testing of the results is advised. Generalizing the predictions for other clinical areas would be unwise, since the tool would be used in unknown grounds where the results are not proven to be reliable.

Often, the focused outcome is a requirement, but it is also strongly related to the dataset used to develop the score models. There are studies which rely on immense datasets contemplating millions

Table 3.1: Review of traditional statistical studies in postoperative prognostics (chronological order)

Study	Surgical cohort	Model	Data type	Data size	Validation	Outcome
Saklad [52]	General	N/A	N/A	N/A	N/A	Morbidity, Mortality
Knaus et al. [39]	General	LR	Clinical	5,815	Yes	In-Hospital Death
Charlson et al. [12]	General	WI	Clinical	559	Yes	1-Year Mortality
Copeland et al. [17]	General	LR	Clinical	1,372	N/A	Morbidity, Mortality
Marcantonio et al. [41]	Noncardiac	LR	Clinical	876	Yes	Postoperative Delirium
Whiteley et al. [62]	General	LR	Clinical	10,000	Yes	Morbidity, Mortality
Roques et al. [51]	Cardiac	LR	Clinical	19,030	N/A	Mortality
Dupuis et al. [23]	Cardiac	LR	Clinical	3,548	N/A	Morbidity, Mortality
Arozullah et al. [2]	Noncardiac	LR	Clinical	160,805	Yes	Postoperative Pneumonia
Sutton et al. [56]	General	LR	Clinical	3,144	Yes	Morbidity
Donati et al. [22]	Cardiac	LR	Clinical	1,936	Yes	Mortality
Gawande et al. [26]	General	PS	Clinical	303	Yes	Morbidity, Mortality
Canet et al. [10]	General	LR	Clinical	2,464	Yes	Pulmonary Complications
Gupta et al. [28]	General	LR	Clinical, Demographic	211,410	Yes	Cardiac Complications
Vaid et al. [59]	General	LR	Clinical, Demographic	202,741	Yes	Mortality
Bilimoria et al. [5]	General	LR	Clinical, Demographic	1,414,006	Yes	Morbidity, Mortality

LR = Logistic Regression; PS = Point System; WI = Weighted Index; N/A = Not Available

of people from different medical cohorts and multiple hospitals, like the American College of Surgeons (ACS) National Surgical Quality Improvement Program (NSQIP), which makes use of data collected from 393 American hospitals, totaling almost 1,500,000 patients [5]. Studies with such extensive datasets are able not only to have more accurate results, since the models have more samples of similar cases, but also to have more than one prediction target like ACS does, with 8 outcomes, one dedicated to mortality, one to morbidity and 6 "secondary" other dedicated to general complications. Each one predicted by its own regression model.

On the other hand, there are scores using datasets no larger than a few hundreds of records, which seem oddly common. The Surgical Apgar Score used only 303 patients for the training phase of the model [26]. It's important to note that, at the same time, the study only considers 3 variables to make the predictions. There is a ratio of 100 records for each variable. So how can the results stay relevant in smaller studies? Apparently, as long as the number of records is enough for the dimensionality of the dataset in hands and the output classes are actually well represented, there should be no performance difference in the validation set, provided that the validation set is somewhat related to the training set used to create the models.

All this goes to show that the surgical cohort available at the time of research and development is a crucial factor, that can limit the final outcome. Broad datasets contribute for a larger populational applicability and for a greater number of possible predictions. Not only the extension of the dataset should be analyzed but also the sparsity within the cases in record. There should be enough cases of the sort to predict, for relevant and reliable results.

Data type - The data throughout the vast majority of the reviewed traditional statistical studies is limited to clinical or clinicopathological data. Very seldom did the studies include socioeconomic or demographic data, important variables that could make the international applicability of each study much broader. One of the few is the ACS NSQIP Surgical Risk Calculator [5], which accounts for demographic data, collected from over 393 hospitals all across America, having a solid and proven national applicability.

3.1.1 Traditional Models

The novelty behind each one of the reviewed scores seems to be related either to the type of model used, the dataset extension, or the set of specific features used to train the model.

Point systems - There are models ranging from simple scoring point systems, based on a number of factors, to slightly more elegant regression models. Charlson Comorbidity Index [12] or the Surgical Apgar Score [26], used to classify disease severity and also predict in-hospital death, are good examples of point systems that sum the results or apply the points in some type of formula in order to get the output. This kind of methods are somewhat basic, lacking the adaptability and complex modeling capabilities

that machine learning models can easily attain nowadays. The tools can be manually tuned, based on a number of factors which were studied and proven to have impact on a certain outcome.

Statistical Models - Other scores in the list make use of more complex models to make their prediction, in fact, this is the case with the majority of the reviewed scores, as seen in Table 3.1. The difference between regression and point systems or weighted indexes, in practice is very small, and resides solely on the way in which the weights of each factor are approximated to fit the data. The most used model is multivariate logistic regression which seems to be a real work horse among the rest of the tools under analysis. Logistic regression is a special case of linear regression, generally used when the target variable is of binary nature. This type of regression is essentially obtained by the application of a sigmoid function to linear regression. The use of approximation methods based on Ordinary Least Squares or Maximum Likelyhood Estimation are viable approaches to determine the parameters of the regression models [6].

3.2 Machine Learning Studies

More recently, machine learning has also stepped into the field, and the studies using this type of models, specifically for the prediction of postoperative complications, have also been increasing, as shown in Fig.1.1. In a primary analysis these studies bring new prediction models to the table, with high dimensional modeling capabilities, each having its own advantages. Table 3.2 provides the complete list of reviewed machine learning studies for postoperative prognostic.

From statistics to machine learning - A key aspect of machine learning studies is the fact that their application is more recent when compared with traditional statistics studies. The median publication year of the traditional statistics studies corresponds to the year 2001, while Machine Learning (ML) studies correspond to the year 2015. In these fourteen years technology has evolved a lot, and now, more than ever, the available hardware allows for feasible application of very complex methods, as some of the ML models presented in this paper can be.

Not only has the hardware improved in pure processing power, but big clinical data is also a growing phenomenon. Because of this increased data availability we can now see, when comparing Table 3.1 and Table 3.2, the difference in data type. More recent ML models are making use of genomics, biological, physiological, radiomics, demographic and socio-economic data. By these means, ML models not only have more advanced prediction capabilities but they also have a diversity of types of data that adds to the adaptability and reusability in different clinical and surgical areas.

Another characteristic differentiating ML and traditional statistic studies is the already mentioned adoption discrepancy. Many statistical studies were actually developed by doctors or with doctors participating in the study development. ML approaches seem less connected to the professional medical

Table 3.2: Review of machine learning studies in postoperative prognostics (chronological order)

Study	Surgical Cohort	Model	Data Type	Data Size	Validation	Outcome
Khan et al. [37]	Breast cancer	Fuzzy DT	Clinical, Biological	162,500	Yes	5-Year Mortality
Chang et al. [11]	Oral cancer	NN, Fuzzy NN, SVM, LR	Clinical, histopathological, genetic	31	Yes	3-Year Mortality
Zikeba et al. [65]	Lung cancer	Boosted SVM	Clinical, histopathological	1,200	N/A	1-Year Survival
Danjuma [19]	Lung Cancer	MLP, DT, NB	Clinical	470	Yes	1-Year Mortality
Parmar et al. [44]	Head & Neck cancer	NB, RF, NN	Radiomics	101	Yes	3-Year Mortality
Wang et al. [61]	Bladder cancer	NB, SVM, kNN, NN	Clinical, histopathological	117	Yes	5-Year Mortality
Thottakkara et al. [58]	Major surgery	LR, GAM, SVM, NB	Demographic, socioeconomic, clinical, laboratory	50,318	Yes	Postoperative Sepsis and Kidney Injury
Soguero-Ruiz et al. [54]	Colorectal cancer	SVM	Physiological, clinical	402	Yes	Anastomosis Leakage
Kim et al. [38]	Oral cancer	NN	Clinical, histopathological	255	Yes	5-Year Mortality
Parikh et al. [43]	General oncology	LR, GB, RF	Demographic, laboratory, comorbidities	26,525	Yes	180-Day and 500-Day Mortality

NN = Neural Network; DT = Decision Tree; LR = Logistic Regression; GB = Gradient Boosting; RF = Random Forest; NB = Naive Bayes; GAM = Generalized Additive Model; SVM = Support Vector Machine; kNN = k-Nearest Neighbors; MLP = Multilayer Perceptron

setting, partly because the development of such algorithms is held by artificial intelligence researchers. The use of this type of models seems to occur only at an experimental level. The ML studies reviewed (Table 3.2) seem to be more model focused due to their benchmarking character, and generally aimed at beating statistical models. While traditional statistics studies (Table 3.1) show a greater care for clinical integration and are less model centered.

3.2.1 Machine Learning Models

k-Nearest Neighbors - The k-Nearest Neighbours (kNN) algorithm is one of me most intuitive and simple methods available among the ML bunch. In the studies reviewed it is used only once by Wang et al. [61]. In the context of that study, the kNN model was chosen to take part in a group of relevant ML techniques tested to find the ones which suited the problem better. The implementation was done using the Euclidean distance to calculate the k closest neighbors. The right k was found through several experiments, settling with the k which offered the best results, avoiding the impact of outliers (k too low) and local dominance (k too high).

Naive Bayes - Naive Bayes models are also suggested in situations where lightweight and simplistic solutions are enough to respond to the challenge. The Naive Bayes (NB) model was used in four of the ML studies in review [58, 61, 44, 19]. This method is applied in all studies assuming that all the attributes are conditionally independent. Due to its simplistic ways, it did not score as the best method across all the 4 studies. Yet, according to Danjuma [19], this simplistic method is capable of improved prognostic compared with logistic regression. In Parmar et al. [44], in spite of the fact that it wasn't the best, the results were competitive with that of Support Vector Machine (SVM), Neural Network (NN) and Random Forest (RF).

Decision Trees - A Decision Tree (DT) is a non-parametric supervised learning algorithm used to model non-linear relations between variables and outcomes, suited for mixed data types, numerical and categorical. DTs are popular due to their shorter learning curve and high interpretability, based on a tree like representation. There are different algorithms to implement a DT, but they work in similar manner, by recursively partitioning to asses the impact of specific variables on the outcome [19, 50].

In the papers reviewed, DTs were used in two of them. Danjuma [19] used a DT to predict mortality within 1 year. The results were very good, only slightly surpassed by the Multilayer Perceptron (MLP), a particular type of artificial neural network.

Fuzzy DTs are very similar to a normal DT, the difference resides in the fact that they do not work with crisp classification, meaning an outcome can be associated to various classes with a certain strength. Khan et al. [37] used this type of model and compared its performance to a crisp DT. The results point out to identical performances, but although there are no significant performance improvement, fuzzy logic can bring a different kind of insight to the predictions and model interpretability.

Support Vector Machines - SVMs are another ML model which is frequently used among clinical predictors. In this review, 4 of the papers used SVMs alongside other models, in order to compare results, making this model the 3rd most used one. SVMs are not as understandable and explicable as other methods like DTs or kNN [4]. In simple terms, the algorithm tries to approximate an hyperplane that can set the border between different outcome classes, by maximizing the margin between the hyperplane and the instances of the different classes, setting the decision boundary.

Chang et al. [11] used a linear kernel SVM to make the predictions about 3-year mortality. The results were not very good, but no further investigation was held. One could assume the problem could not be modeled by a linear kernel, meaning that the data was not linearly separable. Although not competitive, the results were good enough to match the performance of Logistic Regression.

Soguero-Ruiz et al. [54] tested linear and non-linear kernel SVMs. Various sets of variables were in use, free-text from clinical records, blood tests and vital signs. The three sets were tested in different combinations to assess what would yield the best results. The non-linear kernels were doing better when heterogeneous types of data were in use, while the linear kernel was better for free-text resulting from the clinical records of patients. In the end, the linear kernel results were still not as good when compared to the non-linear approach.

Thottakkara et al. [58] also used an SVM as one of the options in study. The results were conclusive, a linear SVM was the best model in the study, surpassing the traditional logistic regression. The trade-off identified was the computational complexity, which in an SVM can go as far as $O(n^3)$ for a kernel SVM, compared to $O(n)$ for logistic regression.

Lastly, Wang et al. [61] used a polynomial kernel SVM model to predict 5-year mortality. The best model in test was a NN type of model. The SVM model had slightly inferior performance, with its sensitivity being lower than its specificity, unlike other models in study.

Neural networks - NNs seem to be one of the most popular models currently. Five of the reviewed studies were making use of this type of models, and some of them even test more than one type of NN, varying in learning algorithm, architecture, and other parameters.

Kim et al. [38] used DeepSurv, a Python module, for building deep neural networks. This package was designed specifically to make predictions about survivability and is referenced in the mentioned paper. DeepSurv is a multi-layer feed forward network. The package already provides optimization functions, based on Grid Search to find the best hyperparameters for the model, so the actual structure used is somewhat abstracted by this highly automated package. In terms of performance, DeepSurv was the best model out of Random Forest and Cox Proportional-Hazards, a traditional statistical model used for survivability prediction.

Allied with various feature selection methods Parmar et al. [44] tried to predict 3-year mortality on a small dataset of 101 patients, with high dimensionality, containing 404 features. After feature selec-

tion, only 30 features remained, and out of all the models NNs were getting the best Area Under the Curve (AUC) and stability across testing with different feature selection methods. Unfortunately, the implementation and structure of the NN was not described in this paper.

Chang et al. [11] used two different types of NNs in its study. First, a multi-layered feed forward neural network, which is the most common type of NN. It was trained using the Levenberg-Marquardt algorithm. The structure of the network was 1 hidden layer with 5 neurons and was run for 5 epochs (overall giving the best results for this method). The training stopped when there was no improvement on the mean squared error for the validation set. The other network was a fuzzy classifier, a paradigm contrasting with crisp classification, called adaptive neuro-fuzzy inference system (ANFIS). The rules generated are based on the number of inputs and the number of input membership functions. The rules generated are the output membership functions which will be computed as the summation of the contribution from each rule towards the output. The overall best method was ANFIS and the overall worst was the normal NN. As to why the traditional NN performed so badly, no thorough mitigation was described in the paper.

Danjuma [19] is another publication using NN, specifically a Multilayer Perceptron using back-propagation to adjust the weights during training. Unfortunately, no further explanation about the MLP structure was disclosed, but the results outperformed the other two methods in study, DT and NB.

Lastly, Wang et al. [61] also used various NNs in their set of ML models. The first NN, was trained using the back-propagation algorithm. It had 10 input nodes and 1 output node. The hidden layers were tuned in number and size, in order to obtain the best accuracy while maintaining a good performance, since it is reported that the generalization ability and computational time increased as the number of hidden neurons also increased. The learning rate was also target of optimization through several experiments. Wang et al. [61] also made use of other types of NNs, called Extreme Learning Machines (ELM). A key feature of ELM is that the weights and the bias between the input and the hidden layers are randomly assigned, whereas the weights between the hidden and the output layers are analytically determined by utilizing Moore–Penrose generalized inverse operation of the hidden output matrices. A variation of ELM, regularized ELM (RELM), was also investigated in this study. RELM improves its generalization performance by using the least squares regression method to identify the degree of relevance of the weight linking a hidden node to the output layer. In RELM, the regularization parameter γ is introduced to improve the controllability. In order to reduce the effect of noise, RELM also introduces a weighting factor v_i to weigh the error between the output of RELM and the actual output of the i^{th} input sample. In the end, RELM was the best model, followed by ELM, while the simpler MLP was left behind with results closer to that of kNN.

Ensemble Learning - Ensemble models in machine learning combine the decisions from multiple models to improve the overall performance. By combining the predictive performance of several weak predictors to form a voting system, ensemble methods are able to improve the overall performance,

offering more stability to Machine Learning algorithms [8]. Among the reviewed papers there are three which are making use of such models.

Zikeba et al. [65] proposed a boosted SVM model to solve inner and between-class imbalanced data problems. The problem of uneven data is solved by proposing weighted error function with different misclassification costs, for positive and negative examples respectively. The boosting algorithm used is AdaBoost, which makes use of weak learners (in this case SVMs) to iteratively adjust the data weights in order to increase the significance of misclassified weights, tackling the imbalance dataset problem. The results revealed good performance from the ensemble method, and proved the ability to overcome imbalance induced bias. In the study, there were other model variations allied with resampling techniques, but the AdaBoost powered SVM was the best overall.

Random Forests are a result of the combination of multiple DTs. Each of the trees classifies one instance and they all contribute to the final result by voting what should be the result. Parmar et al. [44] used a RF model among their models set, the implementation is not detailed in the paper, but the results suggest that although not offering the highest accuracy, this model has a competitive performance, but above everything else it proved to be much more stable across the tests.

Parikh et al. [43] used a Random Forest model (RF) and also Gradient Boosting (GB), both tree based ensemble models. The RF was tuned using Grid Search, meaning that an intensive search method was used to find the best hyperparameters for the model, and the same for GB. The obtained parameters, and even the code used to build the models, are both disclosed in a supplement to the paper. GB works in a similar manner to that of the aforementioned AdaBoost. The difference between the two is in how they identify the shortcomings of the weak learners, AdaBoost does so by giving high weights to data points, while GB does conceptually the same by using gradients in the loss function (which can be user specified). Both models showed good results with a positive predictive value superior to that of traditional statistical values. Another conclusion drawn from the application of these ensemble models was that they also helped recognise important predictive variables that were previously ignored by traditional statistical methods, contributing to the future of short term mortality prediction in cancer patients.

3.3 Cohort Data Preprocessing

Before the learning step, an important phase consists on treating the available data to make it proper for the model application. This process is inherent to every study under analysis in this article, but is scarcely documented. Out of all the 26 publications analyzed for this review only 10 actually referred the strategies used to tackle preprocessing challenges. The problems are distinct among each other, the common characteristic they share is just the fact that they have to be addressed in a preprocessing

stage, prior to model application. Data science problems usually obey to simple principles like “rubbish in, rubbish out”, in other terms the data we use to train our models will strongly influence the quality of the obtained predictions [30]. If there are more variables than records this can only create redundant learners. This problem is know as “the curse of dimensionality”, previously described by Bellman [3]. Not only the size of the training set is important but it should also contain representative portions of the outcomes we wish to predict [3]. In the following paragraphs, the preprocessing challenges found among the reviewed literature are presented, together with a set of generic solutions.

3.3.1 Missing Values

Missing Values are the result of unavailable data at the time of registry and can sometimes be a product of human error. Since some of the models from Scikit Learn cannot handle missing values, they have to be either eliminated or replaced by some other meaningful value.

In some cases, it’s possible to just drop all the records containing missing values, provided that losing the data of one patient won’t have a huge input on the model training. But in real life datasets, missing values and other issues are everywhere. For that reason, we could not simply afford to drop one third of our training set because of random missing values which could easily have been replaced at the cost of introducing some error by potentially misreplacing some of them. There are several strategies to perform what’s known as imputation of missing values, resorting to the use of the mean, median or mode of a numeric variable, or by creating a new class like “missing” for categorical variables, as in [58] and [60].

Another solution consists on using methods which create less of a biased impact. If needed, a model like kNN could be used to predict the value with which to impute the missing one, by taking into account the most similar records, maintaining, in theory, a higher fidelity to the real value when compared to previous proposals, as in Bilimoria et al. [5] (using a regression method). Additionally, a missing value might also exist to represent situations where a certain variable was not applicable, therefore imputation should not be performed blindly.

3.3.2 Outcome Class Imbalance

Class imbalance is a common problem in medical decision problems, often resulting from the high rate of successful cases [65]. Due to this inevitable fact, depending on the model used, the predictions can be biased towards the majority class. This situation is potentially dangerous since the minority class is commonly the class representing negative effects like death or some morbidity factor which cannot be neglected.

This problem is frequently addressed by simple methods like resampling. Resampling consists on either increasing the amount of records belonging to the minority class, reducing the amount of records

belonging to the majority class, or a combination of both. The reduction is the simplest method, since it basically consists of dropping records. But information is precious, and these studies are not making use of very extensive datasets to start with, so oversampling through the creation of new synthetic entries belonging to the minority class might solve the bias issue maintaining all of the original data at the cost of some error which might be introduced through the synthetic generation of records.

But not always do we have to directly solve the problem at its root. This preprocessing issue might also be addressed out of the preprocessing stage, by selecting models somewhat immune to the effect of imbalanced data. Zikeba et al. [65] used various ensemble methods based on SVMs which proved to be efficient at dealing with data imbalance.

3.3.3 High Dimensionality

As mentioned previously, one of the problems that can be faced when dealing with high dimensional data, containing an elevated number of features, is lacking the amount of records to go with the variables ending up in the "dimensionality curse" [3]. This issue is usually associated to overfitting, when the results from the test set are worse than the results obtained in training.

To tackle this problem, one possible solution is to use a feature selection technique, in order to pick the most relevant variables for model construction, as in [11, 44] or [43]. This method allows for better model interpretation, reduced training times, avoid the curse of dimensionality and therefore improve the generalization ability, by reducing overfitting, or more formally, reducing variance.

Another less simplistic alternative consists on applying feature extraction techniques. The latter is different from feature selection in the sense that it doesn't deliberately drop variables used for training. The principle behind feature extraction is to project data into a smaller space, reducing the dimensionality, but it makes sure to keep all the original variables, they are just transformed. One particular example is Principle Component Analysis [46], which, as the name says, computes the principle components in data. The components are represented by vectors which are linearly uncorrelated. The objective is to choose the components that have the most variance, as in Thottakkara et al. [58].

3.4 Assessment

Assessing prognostic accuracy is not necessarily straightforward and can seldom be summarized by a single metric [33]. Several characteristics have to be considered in order for the results to be significant. A good risk score should be able to tell the difference between a patient who is very likely to have a negative prognostic and a positive one. In surgical predictions, it is important that the model offers good discriminative power.

Confusion Matrix – The majority of the reviewed publications use metrics based on information extracted from a confusion matrix [47]. These matrixes are a type of contingency table with two dimensions, showing the instances in a predicted class versus the instances in an actual class. From the error/confusion matrix, various metrics can be withdrawn:

- Sensitivity = Recall = True Positive Rate = $\frac{TP}{TP+FN}$ (3.1)

- Specificity = False Positive Rate = $\frac{FP}{FP+TN}$ (3.2)

- Precision = $\frac{TP}{TP+FP}$ (3.3)

Where TP = True Positives, TN = True Negatives, FP = False Positives and FN = False Negatives.

Receiver Operating Characteristic (ROC) Chart – The Receiver Operating Characteristic (ROC) curve can also be used to assess the model performance specifically as a measure of class separability. This curve plots the true positive rate (TPR) against the false positive rate (FPR), where True Positive Rate (TPR) is on y-axis and False Positive Rate (FPR) is on the x-axis. It is most commonly used in binary outcome settings, but can also be used for categorical outcomes with more than two possibilities. In this last case, one ROC curve is plotted per outcome value in order to assess the separation ability, and plotted in overlap for comparison. In order make an objective analysis, a common metric used with this curve is the Area Under the Curve (AUC), which traduces numerically the principles explained before by calculating the area that is under the ROC curve. Bridging back to the reviewed studies, only three of them did not refer using the aforementioned types of metrics to assess their predictions, [26, 12, 62].

Chi Square Test – The Pearson's Chi Square Test is a statistical test applied to sets of categorical data to evaluate how likely it is that any observed difference between the sets arose by chance [45]. After the confusion matrix and ROC chart derived metrics, the chi square is one of the most frequently used metrics to assess the statistical significance of binary predictions. In total, eight of the reviewed studies explicitly made use of this statistical test, [39, 10, 26, 23, 62, 12, 41, 17].

R^2 – In the context of numeric outcomes, $y \in \mathbb{R}$, the Coefficient of Determination, or R^2, traduces the percentage variation for the dependent variable explained by the independent variables, being a strong indicator of the goodness-of-fit. This metric is used in two of the reviewed studies [39, 12]. In equation

3.4, $\hat{y}$ is the predicted value, y is the actual value and $\bar{y}$ is the mean value of y.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{3.4}$$

Hosmer-Lemeshow Test – Much like the coefficient of determination, the Hosmer-Lemeshow Test (HL) is a measure of the goodness of fit, specifically designed for logistic regression models, frequently used in risk prediction tasks [32]. This test is very useful in the calibration phase of the models, since it assesses whether or not the observed event rates match expected event rates in subgroups of the model population. Those subgroups are based on the deciles of fitted risk values. Models for which expected and observed event rates in subgroups are similar, are called well calibrated. In the context of the reviewed papers, this metric was used by Arozullah et al. [2], Canet et al. [10], Donati et al. [22], Bilimoria et al. [5] and Thottakkara et al. [58], all making use of Logistic Regression.

3.4.1 Error metrics

Regression models and classifiers with probabilistic outputs, respectively, return quantity estimates or the posterior probability of the output, prior to dichotomization as in Danjuma [19]. In this context, error metrics can be placed to assess how distant are predictions from true observations:

- Root Mean Squared Error (RMSE) - Root mean squared error is a quadratic scoring rule that also measures the average magnitude of the error. Since the errors are squared before they are averaged RMSE gives a larger weight to larger errors. This characteristic can also be relevant when Mean Absolute Error (MAE) is used, since RMSE can work as an upper and lower bound to MAE;

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \tag{3.5}$$

- MAE - Mean absolute error measures the average magnitude of the errors on a set of predictions without considering their direction. All the individual differences have equal weight. An advantage of using MAE is that it should be more stable than RMSE when the test samples are of different size which is often the case in the real world;

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j| \tag{3.6}$$

- RAE - The relative absolute error is relative to what the results would have been if a simple predictor had been used, which is just the average of the actual values, $\bar{y}$. Thus, the relative absolute error takes the total absolute error and normalizes it, dividing by the total absolute error of the simple

predictor;

$$RAE = \frac{\sum_{j=1}^{n} |y_j - \hat{y}_j|}{\sum_{j=1}^{n} |\overline{y} - \hat{y}_j|} \tag{3.7}$$

- RRSE - The root relative squared error is very similar to the relative absolute error, in the sense that it is also relative to a simple predictor. The difference is that RRSE uses the squared error instead of the absolute error. Thus, the relative squared error takes the total squared error and normalizes it, dividing by the total squared error of the simple predictor. By taking the square root of the relative squared error one reduces the error to the same dimensions as the quantity being predicted.

$$RRSE = \sqrt{\frac{\sum_{j=1}^{n} (y_j - \hat{y}_j)^2}{\sum_{j=1}^{n} (\overline{y} - \hat{y}_j)^2}} \tag{3.8}$$

3.5 Validation

The validation process is critical as it offers an ultimate view of results before real world application. Problems related to poor world wide applicability have been reported in studies [25, 14, 24, 27]. The common conclusions seem to point out that further validation with foreign datasets would be crucial to obtain better reusability. These studies also seem to point out that there are social and economic factors that should be included in the models. The proposed solutions are, more often than not, to choose the creation of a new score that can suit the characteristics of a specific population, spending precious funds and time investigating new specific tools, instead of focusing on the development of more widely applicable studies.

Out of all the studies presented on Tables 3.1 and 3.2, only five out of twenty six do not refer any validation means. Perhaps because of low data availability or highly experimental character.

The ones that indeed use some type of validation, use one of the aforementioned methods, cross validation or an independent validation set. The latter is the most common among the reviewed studies, with only 5 studies not using a separate dataset as their validation means, resorting to cross-validation [54, 11, 61, 19, 58].

4

Methodology

Contents

4.1 The Dataset 26
4.2 Preprocessing 28
4.3 Prediction Models 32
4.4 Assessment and Validation 36
4.5 Implementation Details 41

This project resembles a classical data science problem, with a tabular dataset. Common issues like missing values, manual text input and other types of inconsistencies are relatively common in this dataset. The aim of this study is to predict 4 outcomes of interest: existence of postoperative complications, severity level of said complications, death probability within 1 year and a prediction for the number of days spent in the ICU for a specific patient. These outcomes are different in the nature of the question asked to the model. For instance, the prediction of the existence of complications calls for a classification task, a yes or no question, while the number of days in the ICU has a purely numeric nature.

The dimensionality of the dataset, with 130 variables for approximately 850 patients (observations), allied with very sparse data, in the sense that there are several different types of cancer and surgical procedures, results in imbalanced data problems and underrepresented groups. The presence of missing values, imbalanced data, hidden variable dependencies and an overall heterogeneous population, makes the preprocessing phase harder and presents new challenges in the development and application of the prediction models.

4.1 The Dataset

The dataset was provided by courtesy of the Portuguese Institute of Oncology (IPO), Porto, Portugal. The data derives from a prospective cohort study of cancer patients that have undertaken surgery at IPO-Porto, and were monitored from 2016 to 2018. It is essentially composed of clinical data, containing approximately 850 entries, of different patients that went through cancer related surgeries, and is already anonymized.

4.1.1 Data Profiling

For each patient there are about 130 variables registered. There are 79 categorical variables, out of which 33 are binary, 44 numeric, 4 in date format, and 9 pure text variables. All these variables are spread across different sections of the dataset, containing different types of information, which are:

- Patient Information (4 variables);
- Hospitalization Characteristics (15 variables);
- Surgery Information (7 variables);
- P-Possum Score (22 variables);
- ACS Risk Calculator (50 variables);
- Assess Respiratory Risk in Surgical Patients in Catalonia (ARISCAT) Score (9 variables);
- Charlson Comorbidity Index (17 variables);
- Post-Operative Complications (7 variables);
- Hospital Discharge (5 variables).

The dataset attributes are mainly categorical, each number or textual key is used as a mapping to some type of meaning. Largely due to the fact that the scores used by IPO-Porto already do a good job standardizing input variables, like using yes/no questions such as “The patient was a smoker in the last year?”, like in ACS Risk Calculator [5]. This approach, helps discretizing numeric input into well defined intervals or mapping a number to one specific option. For instance, the numbers from 1 to 3 could describe the diabetes attribute with“1-Not diabetic”,“2-diabetic controlled with oral medication”,“3-diabetic controlled with insulin”.

The rest of the dataset consists mainly of numeric data, requiring eventual imputation and/or normalization. Only a few attributes are in text format, requiring special treatment. Some might be easier to handle since they describe standardized topics that have codes associated to the text. Like the specific procedures that were used in surgery, which can be both analyzed in text form or code form. There are several strategies to handle textual data, based on word counting and/or dummification of the variables in question, for example. But that might contribute to a problematic increase in dimensionality, and because the text is a product of free user input, dealing with this data becomes not trivial. Therefore, this type of data is not used, since it would require a time investment and techniques that could constitute a project of its own. In fact, IPO-Porto already has a team working on processing these textual variables, in a separate ongoing project. Prior to the beginning of the development phase, it was established that dates and textual data (which could not be easily made into categorical data) would not be part of this study. This way, the total number of variables fell from 130 to 117 variables.

Additionally, IPO-Porto is working on extending the dataset with molecular, nutritional, psychological and physiological data. Creating new possibilities for other prediction outcomes, and contributing to the enhancement of the currently proposed ones.

Across the various sections of the dataset, there are variables that are obtained during or after the surgical procedure. These variables cannot be used to build a tool that is intended to make predictions about the future state of the patients prior to the operation. After dropping these variables, the number decreased from 117 to 83 in total.

4.1.2 Data Exploration

In order to get familiar with the dataset and the topic at hands, a data profiling process was carried in a first phase. This process consisted essentially on identifying the type of data in each variable (e.g. numeric, categorical, textual, or dates), and scan for missing values, calculating the rates of missing data for each variable. In the beginning stages of this project, the clinical meaning of each variable was also assessed against the available domain knowledge. A division was made between preoperative (used as input data), intraoperative and postoperative variables (used as targets).

4.2 Preprocessing

4.2.1 Missing Values

Datasets are frequently affected by the existence of missing values, which make the process of applying machine learning models less trivial, since not every model can deal with such situations.

Some variables were nearly absent for the 850 patients, with missing value rates over 90% in some cases. Such variables could be deemed as non-relevant information from the start and maybe even eliminated from the dataset completely. But not all variables should be analyzed in the same way and not all missing data should be treated as such. In some cases, not having a registry might mean something by itself. Take, for example, the attribute "Death Within 1 Year". The percentage of missing values is 82%. At first, one might think that this variable could be eliminated or maybe the values could be imputed, meaning they could be substituted by some number like the mean or median of the variable. In fact, this variable with 82% of missing values, actually has a large volume of information, since the missing values represent that the patient did not die within a year starting from his surgery. Showing that careful analysis of each feature is advised to avoid losing vital information.

Preprocessing the missing values was not as trivial as initially expected. To make the model application possible, high missing value rated features were left out. Among all the registries, there were still random missing values that would raise future problems. The solution in such cases, where the meaning of the missing data was not clear or the data was actually missing, was to impute the values. The common techniques used for imputation are substitution for the mean, median or mode of the variable. Such methods are very rudimentary and are prone to the introduction of error, since the process of imputation is relatively blind in what concerns the available information for a specific patient. In order to keep the introduction of bias to the minimum, alternative solutions were explored.

The alternative consists of using informed methods to make the substitution. The k-Nearest Neighbors algorithm can be used as a lightweight informed imputer, that helps to reduce the error introduced when dealing with missing values.

The only variables to which imputation was not applied were the outcome variables. As this project is focused on predicting postoperative outcomes, the presence of output variables is necessary. The models should not be trained on synthetic labels, which might introduce bias. Also, all the outcome variables had relatively low missing value rates, so the observations corresponding to patients which happened to have a missing value in the target variable, were dropped from the study, without losing much information overall.

As an additional note, more recently, IPO-Porto pointed out that missing values are in their vast majority used to symbolize a situation where a certain measure or classification is not applicable, at least in places where they are represented by the value "n/a".

4.2.2 Categorical Variable Encoding

Categorical variables are commonly represented through a numeric encoding, which may or may not have some type of order implied in the numeric correspondence. This quantitative or ordinal relationship might undesirably slip into the analysis. There are many possible solutions to this problem, but often the simplest way is to use a One-Hot encoder. This solution is fairly simple, it consists on turning the categorical variable into a series of binary ones. One for each value the original variable might take.

In the context of this project, there are 83 total usable variables. After encoding the categorical data this number rises to 371.

4.2.3 Imbalanced Variables

Imbalanced variables can also present issues, even more in a sparse dataset containing registries of more than 30 types of cancer. This characteristic is common in many fields of study but more so in the clinical field. Resampling can be used to tackle these problems. But the problem becomes more complex, since oversampling might generate too much synthetic data and undersampling leaves very few patients of each class for the models to learn, in certain cases less than 15. Associative models, such as decision trees, are recognized for their capacity to overcome imbalanced data problems. And, from the very start of this project, they have shown promising results.

4.2.4 Resampling

Dealing with clinical data is often not trivial because of certain specificities and complications commonly associated to this type of data. One of them is the imbalance between the positive and the negative class, with the positive class often being severely underrepresented. One of the techniques to deal with this problem, and avoid the bias of the classifiers towards the majority class, is resampling.

Three main strategies were considered: 1) undersampling, 2) oversampling, and 3) mixed strategies. The first one aims to reduce the high amount of instances from the majority class, there are several strategies, one being random elimination of instances. The second one is aimed at increasing the number of records belonging to the minority class. This can be achieved through the duplications of certain records, or, in more sophisticated versions, through the creation of new synthetic entries based on the existing ones. The objective of both these strategies is to have a similar representativity from both (or multiple) classes. The third strategy uses a mix of both the aforementioned ones, to achieve the same goal. Sometimes, it's is critical that we maintain a minimum number of inputs, making us avoid undersampling. Or, in other situations, the introduction of repetitions and synthetic information might contribute to a negative impact of model performance, more even in extreme situations where a lot of new records need to be added to achieve an equilibrium.

In this project, a mixed strategy is used combining synthetic oversampling with k-Nearest Neighbors informed undersampling, as proposed by He and Garcia [31].

4.2.5 Feature Scaling

Numeric data is often available in a wide variety of magnitudes and ranges. Given this undeniable fact, there are algorithms, specially distance based ones, that might give more importance to a variable with values in the ranges of millions than in range of mere decimals. This uneven importance, might end up accounting to neglect variables that could otherwise be critical to the outcome in study. For that reason, it is important to normalize or standardize data.

Gradient Descent based algorithms, such as linear regression, logistic regression, or neural networks, require feature scaling. Having features on a similar scale can help the gradient descent converge faster towards the minima.

Distance based algorithms, such as kNN, or SVMs, are also severely affected by data which is not scaled, since they essentially use the distance between points to make their decisions. As a result of poor scaling, these algorithms have a chance of attributing a greater weightage to high magnitude variables.

Finally, tree based algorithms are fairly insensitive to feature scaling. Magnitudes or ranges should not influence the decision, since there is only one variable being considered at each node.

4.2.6 Feature Selection

In this project's dataset there over 100 variables that can be used as inputs to the models. Just like there are variables that have such high missing value rates that it renders them useless, there are other variables that although not suffering from the same defect, are not relevant to the prediction of certain outcomes.

In data science projects it is very common to select a restricted number of variables that will actually be used to build the prediction models. There are **embedded methods** that can be used for feature selection like regularization in some of the regression algorithms used. For instance, Lasso, Elastic Net and Ridge Regression all use penalization methods that introduce additional constraints into the optimization of the model that bias the model toward lower complexity, resulting in fewer variables. There are also **wrapper methods** that make use of the model to test several sets of variables and find the best match. But these methods are often too blind and require the training and testing of the model, possibly several times, making them computationally costly. And finally, there are **filter methods**, that use methods completely external to the model. They use statistical measures to attribute a score to each variable, often using only one of the independent variables and the dependent variable to make

the tests.

In this project's context, all 3 methods are used. Embedded methods are implicit in the implementation of the algorithms provided in the Scikit-Learn library[48]. Feature selection through filter methods is also used, calculating the relevance and/or correlation of a certain variable for a certain target outcome, using appropriate metrics. The latter method also offers a *p*-value representing the probability that a variable is not correlated to an outcome. For that *p*-value, a threshold value can be defined in order to make the variables selected more or less relevant. That *p*-value is chosen through testing of the models performance, using the wrapper approach.

Regarding filter methods, 3 measures are used, depending on the type of variable being studied and also the type of outcome variable, as shown in fig.4.1.

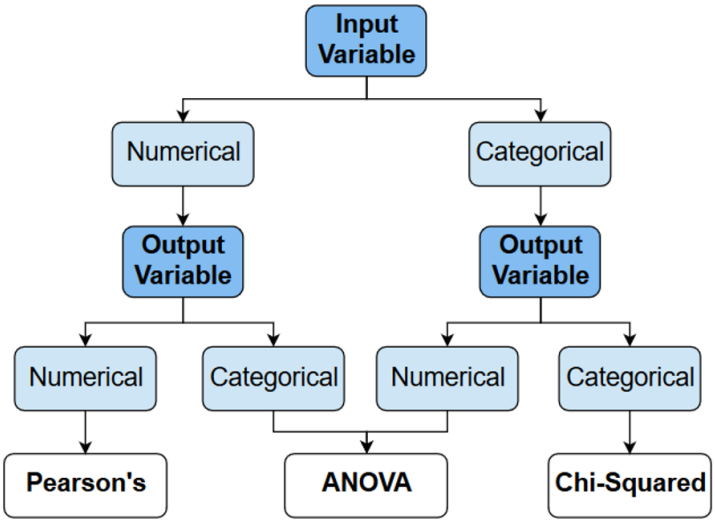


Figure 4.1: Feature selection methods, according to variable and output type

The Chi-Squared test is used to measure correlation for categorical variables, when the output is also categorical. The Analysis of Variance (ANOVA) correlation coefficient is used to measure the correlation between categorical and numeric variables (it is not relevant which one is the dependent variable). And Pearson's correlation coefficient is used when both the independent and the dependent variables were numeric. All 3 of these measures have *p*-values associated to them. The *p*-value expresses the probability that the 2 variables in study are not correlated. Therefore, the feature selection process is done through a choice for the *p*-value threshold, for example, 0.0001. Making the probability of no correlation extremely low, selecting only the most relevant variables.

4.3 Prediction Models

In data science, specifically using supervised learning methods there are essentially two strategies that can be followed: linear and discrete models. In certain cases, one does not invalidate the other, so both approaches should be tested.

The same goes in relation to the models chosen. There are plenty of different models and respective variations. As seen in chapter 3, there is certainly not one model that outperforms all the other options. It depends on a number of factors, and since the dataset available for this project is unique so should be the strategy used to create or choose the prediction model. Therefore, the choice is not obvious, various models will have to be tested for each one of the 4 outcomes. Using a group of state-of-the-art algorithms, the following were the options chosen to make the predictions, distinguishing between classification and regression models.

Classifier algorithms[29]:

- Naive Bayes (NB);
- k-Nearest Neighbours (kNN);
- Decision Trees (DT);
- Random Forests (RF);
- Support Vector Machines (SVM);
- Logistic Regression (LR);
- Multilayer Perceptron (MLP);
- Extreme Gradient Boosting (XGB) [13].

Regression algorithms[29]:

- Linear Regression;
- Ridge Regression;
- Lasso Regression;
- Support Vector Regression (SVR);
- Elastic Regression;
- k-Nearest Neighbours Regressor (kNN);
- Decision Tree Regressor (DT);
- Random Forest Regressor (RF);
- XGBoost Regressor (XGB) [13];
- Partial Least Squares (PLS) Regression;
- Multilayer Perceptron Regressor (MLPR).

In the next section, the various prediction challenges will be addressed as well as the strategy followed for each of the 4 initially proposed tasks.

4.3.1 Prediction of Postoperative Complication

As a first challenge, one broad question could be asked: Is a patient going to have postoperative complications? Since the outcome is binary, “yes” or “no” (1 or 0, respectively), this can be approached as a typical classification problem, with a discrete and well defined set of labels to attribute to a certain patient. Various models could be used in this case. Since there was no clear pick, multiple algorithms for classification were implemented and applied in order to compare results.

The overall theme with this type of project is to test approaches and iteratively improve the solution. At first no feature selection or hyperparameters optimization was applied. Only further ahead in the development of the solution did such tuning processes occur.

Some of these models already use inner mechanisms capable of selecting the most relevant features for their predictions, like XGBoost [13] does. For the next phase this is a crucial step allowing for better accuracies and noise reduction, by eliminating irrelevant attributes, and also reducing the dimensionality of the problem.

4.3.2 Prediction of Complication Severity

After the first challenge as a prospecting study was finished, another one of the outcomes was then focused. There are models capable of predicting the existence of complications, but how severe will those complications be? This question can be answered by one of the variables of the dataset relative to the Clavien-Dindo Classification [21], which is a classification used to standardize in 8 grades the type of therapy needed after a certain surgery. This classification was developed for general surgery and was internationally validated.

The Clavien-Dindo Classification is commonly used as a grade traducing the severity of a postoperative complication. In its lower grades it describes low severity complications, where none to a light therapy is required to deal with a situation where there is a deviation from the normal postoperative course. Then, the classification progresses through grades where the patient needs surgery to deal with the complications, and in its maximum grade there’s the death of the patient. So it is reasonable to assume it describes the severity of postoperative complications well.

Having the target prediction settled, there are two approaches that will be followed in our methodology. This challenge can be seen as a classification problem, or it could be seen as a regression problem, using a continuous model that could predict numeric values. Since no clear approach was best at this point, both had to be tested.

This outcome was found to be severely imbalanced in the exploration phase. In this specific case there were grades of the Clavien-Dindo Classification represented by only 15 individuals, contrasting with other grades, lower ones, that were represented by over 250 individuals. Showing a clear tendency

for lower grades of severity. It would then be reasonable to assume that maybe the models were not able to model the high grade situations properly and therefore data resampling was a possible solution to the issue.

4.3.3 Prediction of Death Probability Within 1 Year

The prediction of the probability of death is a relevant indicator to estimate the existence of future complications, and also the viability of surgery for a certain patient. In this case, death might not be the result of postoperative complications exclusively, but rather a combination of factors. The dataset already has information relative to this type of predictions that could reveal useful. There is a binary variable that tells if the patient died within 1 year from the surgery and there is another variable with more details that tells if the patient died within the first 30 days from surgery, from the 30^{th} day to the 90^{th} and from the 90^{th} day until 1 year past surgery. The first variable gives us the solution to the primary problem and the second one might be interesting if one desires to evaluate this outcome in a different time resolution.

This problem is solved by classification with the objective of deciding if the patient was going to die within 1 year or not. Classifiers are able to give an output with a probability associated to the result. Differently from classification, a continuous model could also be used to obtain a probability, but in this case a regression approach would probably not be fit for the task, since the outcome values in the dataset are only binary. For this reason, classifiers are hypothesized to be better candidates to model the decision boundary between one label and another, even more so because these models are able to output the probability associated to the output.

This problem's complexity is high, considering that the outcome variable is very unbalanced, having close to 700 records of people which did not die, and 150 people which did die. This might call for resampling techniques to be applied in order to avoid possible overfitting situations where the models might get too tied to the data and, in this case, too tied to the records where the patients did not die. Contributing to a bias in favor of lower probabilities overall.

4.3.4 Prediction of Days Spent in the ICU

The number of days spent in the ICU represents important information for medical and also financial reasons. A patient that is predicted to spend a large number of days in the ICU will raise suspicion among medical professionals if the medical opinion indicates that he/she is able to leave the hospital after a short postoperative period. Such situations might happen when further testing is required or unpredicted long term postoperative complications are yet to develop. On the financial side it is useful to the hospital to have this time prediction, for administrative and management reasons, but also for patients, who might have to pay for their long stay at the hospital.

This prediction is better solved by a regression model. The number of days spent in the ICU is recorded for each patient, serving as the dependent variable for the model. In this case, there might be multiple approaches delivering good performance, but further testing would be essential in order to decide what model to choose.

Just like in the previous 2 outcomes, the number of days spent in the ICU is also severely imbalanced, since most people spend 1/2 days there. Only rarely someone ends up spending up to 2 weeks or more.

4.3.5 Model Tuning: Hyperparameter Optimization

After the models for the outcomes of interest are implemented and applied in a prospecting exercise, there will be a need for model tuning allied to the assessment of the results in order to optimize behavior, set baselines, evaluate progress and compare solutions.

In a primary study, the models were applied with their default hyperparameters. These parameters are external to the model and the values cannot be estimated from data. Commonly, they are set by the developers to work generically across a range of scenarios. But in many cases these parameters might be far from ideal, requiring customization and tuning to extract the best possible results. Hyperparametrization is the process of tuning the parameters used by the models before the learning process begins. One of the classical approaches is called Grid Search, [34] which is essentially a process where each parameter in a list of values is exhaustively tested, making all the possible combinations to find the parameters that maximize, or minimize, an objective function, which could be the accuracy or other metrics of the model's error.

There are however informed search models that yield better performance and make part of the proposed methodology. Bayesian optimization [34] associates a probability distribution to the hyperparameters. Often the objective function is expensive to calculate, so instead a surrogate objective function is used, which is easier to calculate. This function will guide the choice of the most promising parameters to be tested later with the expensive objective function, based on the probability distribution model. The process of optimization is slow due to the amount of testing involved. The intuition behind using Bayesian optimization is that the search is guided towards the sets of parameters that are more likely to offer good results. Being a trade-off between optimality and execution time.

In the case of our models, there are 2 different objective functions:

- Regression models are optimized in order to minimize their mean absolute error (MAE) which is the more direct way of measuring prediction error. For instance, in complications' severity prediction the objective is to minimize the error along the Clavien-Dindo numeric scale. RMSE could also be used here, but the best MAE might be achieved by a model that has an unstable performance which would be severely penalized by RMSE. The later metric would be better suited if error stability were to be optimized.

- Classification models are optimized to maximize their recall. The sensitivity is calculated for each target class, and is then averaged in a non-weighted formula, because classes generally have similar importance. For instance, predicting the existence of postoperative complication allows for the timely implementation of preventive or curative measures, but knowing when not to take measures can be time and cost saving.

4.4 Assessment and Validation

The evaluation methodologies of this work consists on the assessment and validation of the applied predictive models in accordance with the principles introduced in Background and Related work sections. Right away there is a clear division between the methods used to evaluate regression models and classifiers, which are discussed in the next section.

4.4.1 Classification Evaluation Metrics

Classifiers are used to predict a label within a well defined finite set, which is attributed to a new entry that has not been classified yet. The discrete nature of classifiers allows for simple evaluation, like checking the number of times the classification was correct or not. But the validation cannot be left at the analysis of the accuracy, which represents the number of times the classifier was correct over the total number of predictions. Accuracy can be misleading in situations where the data is imbalanced. Considering a binary variable, like 'yes' or 'no', with a distribution of 50/50, then accuracy can deliver reliable results. But in a situation in which there is a ratio of 99:1, where the majority of the outcomes are 'yes', then a classifier that always guessed 'yes' would have 99% accuracy. In this case the produced classifier would perform badly in a real world context, as soon as a 'no' outcome was expected. Many situations are not as extreme as the last one, but as the distribution varies further from 50/50 for the possible outcomes then accuracy gets ever more misleading.

In order to overcome the weaknesses of the accuracy metric, others are used to complement it, and the majority are based on the information displayed in the confusion matrix. This matrix show us the number of actual values versus the predicted outcome in a detailed way. Fig.4.2 shows an example of a confusion matrix, for a simple problem where the outcome is binary.

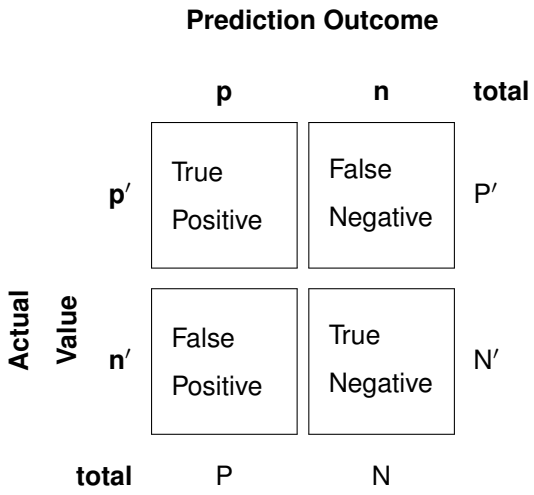


Figure 4.2: Error/confusion matrix representation with labels

From this matrix, various metrics can be withdrawn:

- $TruePositiveRate = TP/(TP + FN) = 1 - FalseNegativeRate$
- $FalsePositiveRate = FP/(FP + TN) = 1 - TrueNegativeRate$
- $Sensitivity/Recall = TruePositiveRate$
- $Specificity = TrueNegativeRate$
- $Precision = TP/(TP + FP)$

The Receiver Operating Characteristic (ROC) curve can also be used to assess the model performance specifically as a measure of class separability. This curve consists of the plot of TPR against the FPR where TPR is on y-axis and FPR is on the x-axis. It is most commonly used in binary outcome settings but can also be used for categorical outcomes with more than two possibilities. In this last case, one ROC curve is plotted per outcome value in order to assess the separation ability, and plotted in overlap for comparison. Fig.4.3 shows an example of a ROC chart.

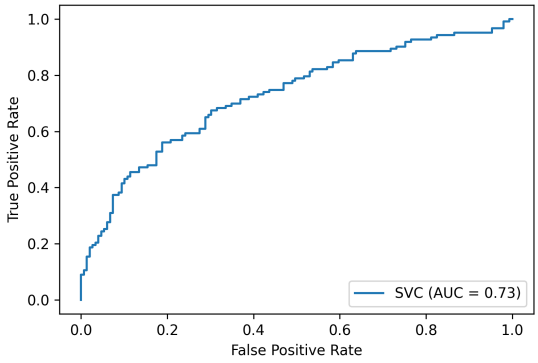


Figure 4.3: ROC chart - prediction complications - (SVM 5th stage)

This plot can be easily analyzed by having in mind 2 key ideas: the dotted transversal line represents a classifier incapable of any type of separation, performing no better than pure chance, and the best classifier would have a curve that would be very near to the 90° angle at the top-left corner. In order to maintain this analysis more objective, a common metric used with this curve is the Area Under the Curve (AUC), which traduces numerically the principles explained before by calculating the area that is under the ROC curve.

These metrics allow for a more precise analysis over each one of the possible outcomes, filling the gaps that the accuracy doesn't cover, complementing its assessment.

Another metric that will be used is the Cohen's Kappa [15], which is a chance corrected standardized measure of agreement between two categorical outputs produced by two raters. In simpler terms, it is a way of comparing the results of two raters also accounting for a chance factor. The result is calculated using the observed agreement (p_o) minus the agreement by chance (p_e), all divided by 1 minus the agreement by chance in order to standardize the result. The formula is presented in equation 4.1.

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{4.1}$$

4.4.2 Regression Evaluation Metrics

While using regression models the results are not on a black and white spectrum like classification. The predictions for regressors are more suitable to be evaluated under mathematical error metrics. There is a plethora of different metrics to use in order to assess model fitment and error. The vastness is explained by the fact that these metrics are very specific in how they put their measures into perspective, on what they measure and how they penalize certain situations, so it's common to use a group of measures that are able to complement each others' weaknesses offering different perspectives:

- RMSE - Root mean squared error is a quadratic scoring rule that also measures the average

magnitude of the error. Since the errors are squared before they are averaged RMSE gives a larger weight to larger errors. This characteristic can also be relevant when MAE is used, since RMSE can work as an upper and lower bound to MAE;

$$RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \tag{4.2}$$

- MAE - Mean absolute error measures the average magnitude of the errors on a set of predictions without considering their direction. All the individual differences have equal weight. If the absolute value is not taken it turns into Mean Bias Error (MBE) instead of MAE. MBE has its own advantages but also strong disadvantages because the positive and negative differences will cancel each other. An advantage of using MAE is that it should be more stable than RMSE when the test samples are of different size which is often the case in the real world;

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j| \tag{4.3}$$

Apart from checking the absolute fitment of the model, the Coefficient of Determination, or R^2, can be used to check the relative fitment of a model, comparing it to a mean model. This metric is based on two simple metrics, the Sum of Square Regression (equation 4.4), which traduces the variation explained by the model, and Sum of Squares Total (equation 4.5), which explains the total variation in data. This coefficient traduces the percentage variation for the dependent variable explained by the independent variables, being a strong indicator of the goodness-of-fit.

$$R^2 = \frac{SSR}{SST} \tag{4.4}$$

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \tag{4.5}$$

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 \tag{4.6}$$

4.4.3 Model Validation

Cross-fold validation offers the possibility to perform a statistical analysis of the results on k folds of the dataset, assessing the ability of the target predictive models to generalize into unseen data. These techniques are used to guarantee that the model isn't overfitting and that it has potential to perform positively when applied in a new validation set or in a real context. The process consists on splitting the dataset into training and test set, not only once but a k number of times, trying to maintain the test set

mutually exclusive between all the splits. Allowing the testing of the model to be performed in simulated independent test sets.

Applying this validation process in a correct manner, requires special care in preprocessing steps. For each fold of the cross validation method, there will always be a training set for the model to learn, and a test set for the model to test it's performance. In each of these folds, an imputer model will be fitted with the training data, so that the training set won't have any influence on the imputation values. The same happens with the normalization module which is fitted with training data only, and applied to the training and test set after. The resampling algorithm behaves in a similar manner, but this time it is only fitted and applied to the training data, leaving the test set raw.

This technique, although useful, cannot fully replace testing in an independent validation set (external validation), in order to assess correctly the generalization capabilities of the models. Right now there are perspectives of having a dataset extension, adding more patient entries and more variables, that can later be used with this purpose.

4.4.4 Model Comparison

The metrics introduced along previous sections are important to assess the final results of this project, assuring reliability and significance, but are also a means of comparing the results for different models. For every outcome of interest several models are developed. When these models reach their peak performance after a process of hyperparametrization, different candidates have to be evaluated and selected either to assume the place of a single predictor or integrate an ensemble algorithm joining multiple models.

A statistical test that might prove useful for model comparison is the Student's *t*-test [55]. This statistic test measures how significant the differences between two normally distributed groups are, in other words, it tells if those differences could have happened by chance. In this case a paired *t*-test should be used, since dependent samples are being compared (the same test set). The null hypothesis is that the pairwise difference between the two test sets is equal. If it proves to be different with a relevant significance level than it is enough to reject the null hypothesis and declare that one is better than the other. This test can be used to compare the performance of different models, against a baseline or even the improvement between development stages. In case the sets are not normally distributed, the Wilcoxon signed-rank test [64] is used as a replacement.

Due to the high number of comparisons, and in order to present a suggestive set of models as the best performing ones in the end, there had to be a system to empirically make these decisions. Reciprocal Rank Fusion (RRF) [18] is recognized as reliable systems to rank different instances according to the values resulting from a group of scores. It's essentially an unsupervised method used to rank the importance of performance estimates, borrowed from the Information Retrieval Systems field. The

formula uses the sum of the inverse of the rank obtained for each of the metrics in use. The rank is affected by a constant, k, to mitigate the effect of performance estimates associated with higher ranks.

$$RRFscore(d) = \sum_{i} \frac{1}{k + r_i(d)} \tag{4.7}$$

4.5 Implementation Details

The proposed methods are implemented using the Python[1] programming language. This language supports a wide variety of libraries specialized in various areas, including Data Science. This project makes use of Scikit-Learn[2] in particular. This package is used for data preprocessing and also for the implementation of predictive models, with the exception of XGBoost[3], which is available in its own package.

Additionally, in order to optimize the hyperparameters of the models, another package called Hyperopt[4] is used. This package offers a Bayesian Optimization implementation which was used over traditional intensive search methods.

For more details about the implementation, all the code can be consulted in this thesis' Git repository: https://github.com/danielmg97/master-thesis-iposcore

[1] https://www.python.org/
[2] https://scikit-learn.org/
[3] https://xgboost.readthedocs.io/
[4] https://github.com/hyperopt/hyperopt

5

Results

Contents

5.1 Exploration and Profiling 43

5.2 Model Results 47

5.3 Associative Models - In Depth Analysis 67

Note: Due to the impossibility of displaying the entire set of results for every task, summarized versions or particular examples of the results can be used. The detailed versions, along with all code, are stored in this thesis' Git Repository: https://github.com/danielmg97/master-thesis-iposcore

In this chapter, the results gathered at the various stages of the proposed methodology will be presented. Essentially, these results focus on the assessment of the predictive capabilities of the models, according to the incremental applications of improvement strategies. This way, the progress is more easily tracked and the impact of each change can be measured and discussed.

5.1 Exploration and Profiling

The initial exploration phase was determinant to get an insight to the newly acquired data, and also to get familiar with the subject. In this phase, data profiling tasks were performed, as well as a statistical study about the distributions of the variables and their impact on the existence of postoperative complications.

5.1.1 Data Profiling

In this step, the objective was to learn more about the type of variables and their meanings, while also studying their distributions. That task was carried out using a form used by IPO-Porto to collect some of its variables, therefore containing information about their meanings. The rest of the variables, that were not contained in the form, were mainly inputs and/or outputs to risk scores already used at the hospital.

Clinical variables were separated according to their type. Our dataset contains 79 categorical variables, out of which 33 are binary, 44 numeric, 4 in date format, and 9 pure text variable. And, from there, a fine profiling study was carried out using the Data Cleaner[1] tool. Mainly, this tool allowed for a great insight into the distributions of numeric features, with statistical details, and also for other types of data, based on value occurrence counts. As this was an extensive study, not all the results can be presented in this document, but they are available at this project's GitHub under the "Data Exploration" directory. The following are samples drawn from the profiling results.

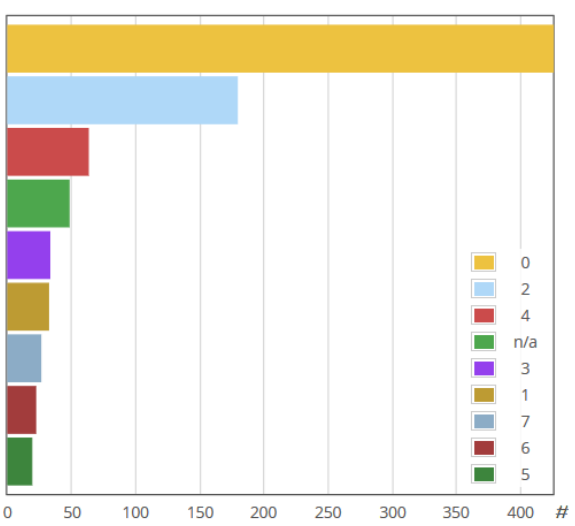
For numeric variables, various statistics were extracted. The number of missing data patients, and several metrics about the value distribution. Fig.5.1 shows us the metrics for 4 of these variables. Several preprocessing issues can be detected by this analysis alone, like the necessity for some type of feature scaling due to range discrepancies between variables, the need for missing value imputation or even to drop a certain variable.

[1] https://datacleaner.org/

	age (number)	days in ICU (number)	days at IPOP (number)	total NAS points (number)
Row count	847	847	847	847
Null count	0	0	7	114
Highest value	97	18	280	1,053.4
Lowest value	1	0.1	1	2.1
Sum	54,819	1,654	16,309	95,376
Mean	64	1	19	130
Geometric mean	63	1	13	98
Standard deviation	13	1	23	120

Figure 5.1: Numeric feature analysis sample

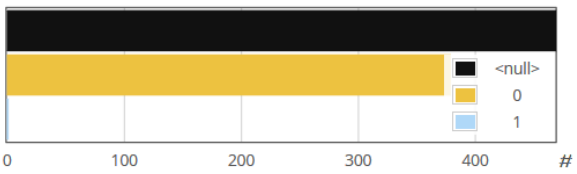
The distribution of the other types of variables, mainly categorical, were were approximated and analyzed. The results also reveal several problems and inherent flaws of the data. For instance, Fig.5.2 makes evident the imbalance problems that would later have to be dealt with. In this specific case, the variable is one of the outputs for the predictions, the Clavien-Dindo severity scale. Out of the 8 grades of this scale, there are lower ones like 0 or 2 that have over hundreds of representing patients. And then there are grades, typically higher ones, like the 6^{th} or 5^{th} grade that do not go over a couple dozens.



0	425
2	179
4	63
n/a	48
3	33
1	32
7	26
6	22
5	19

Figure 5.2: Value distribution analysis for severity level

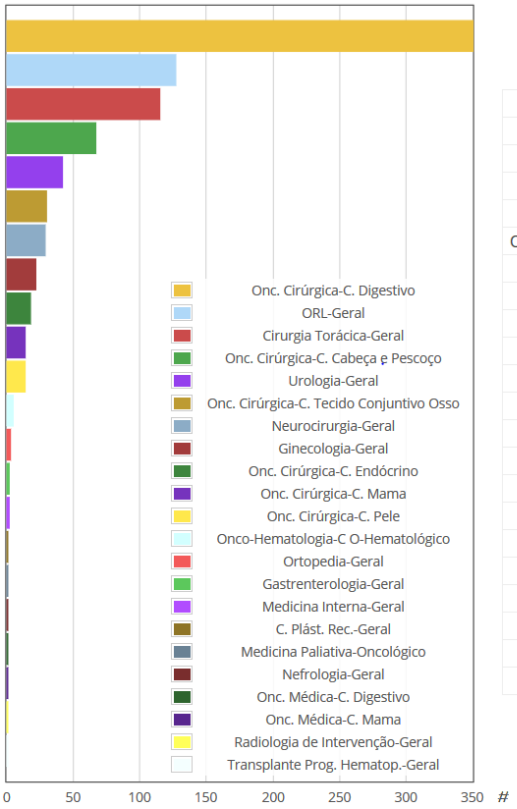
The distribution of the variable “existence of connective tissue diseases”, a binary outcome, exposes a problem. In Fig.5.3, we can clearly see that there are too many missing values for that feature to be useful. Not only that, but also the existing values are totally meaningless since only one patient was found to have connective tissues diseases prior to surgery, and that is obviously not representative. This raises a question, is this a real missing value? Maybe the hospital used null values to represent the absence, or presence, of any disease.



<null>	468
0	378
1	1

Figure 5.3: Presence of connective tissue disease variable distribution

Lastly, Surgical Specialty variable offers a good description of the vast diversity of cancer types or surgical areas. The histogram in Fig.5.4 conveniently shows that the majority of surgeries is related to digestive system cancers. The sparsity of this attribute is also noticeable, strengthening the motivation for future cancer/specialty specific studies.



Onc. Cirúrgica-C. Digestivo	350
ORL-Geral	127
Cirurgia Torácica-Geral	115
Onc. Cirúrgica-C. Cabeça e Pescoço	67
Urologia-Geral	42
Onc. Cirúrgica-C. Tecido Conjuntivo Osso	30
Neurocirurgia-Geral	29
Ginecologia-Geral	22
Onc. Cirúrgica-C. Endócrino	18
Onc. Cirúrgica-C. Mama	14
Onc. Cirúrgica-C. Pele	14
Onco-Hematologia-C O-Hematológico	5
Ortopedia-Geral	3
Gastrenterologia-Geral	2
Medicina Interna-Geral	2
C. Plást. Rec.-Geral	1
Medicina Paliativa-Oncológico	1
Nefrologia-Geral	1
Onc. Médica-C. Digestivo	1
Onc. Médica-C. Mama	1
Radiologia de Intervenção-Geral	1
Transplante Prog. Hematop.-Geral	1

Figure 5.4: Surgical specialty variable distribution

5.1.2 Discriminative Factors of Postoperative Complications

After the profiling phase, an analysis focused on the relation between each variable and the existence of postoperative complications was also performed. Since there are more than 100 features in this dataset, not all the results can be shown in this document. Below, 3 of them are provided as illustrative products of the initial exploration activities. A complete analysis can be found in the “Data Exploration” directory

of this project's GitHub repository.

During this initial analysis phase, the gender attribute seemed to show some useful information, as seen in Fig.5.5 and 5.6. Men were the most representative gender under surgery, representing 64% of the population. Curiously, men also have a higher postoperative complication rate than women by a margin of 7%.

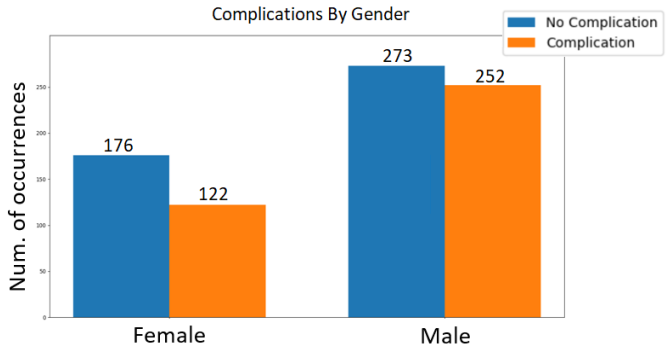


Figure 5.5: Gender variable distribution according to the existence of complications

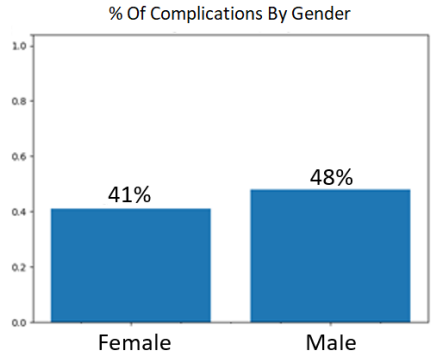


Figure 5.6: Percentage of complications by gender

The emergency character of a surgery also underlies differences in postoperative complications. Despite the vast majority of the surgeries being classified as non urgent, a small portion of all surgeries, about 10%, is urgent and those surgeries practically double the associated risk. Going from 41% of postoperative risk, for a normal surgery, up to 83% for urgent surgeries. Fig.5.7 and 5.8 show the graphic results of the analysis. The percentages are obtained by summing the amount of entries for each class, emergency and non emergency, and then dividing the number of patients who had associated postoperative complications by the previous sum result.

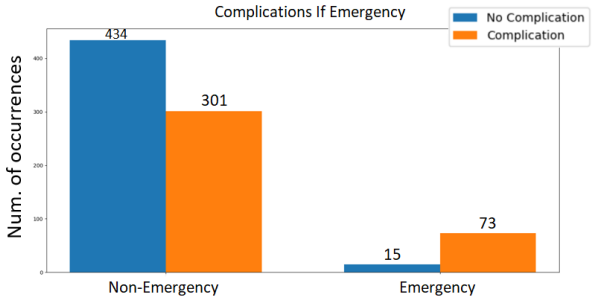


Figure 5.7: Emergency variable distribution according to the existence of complications

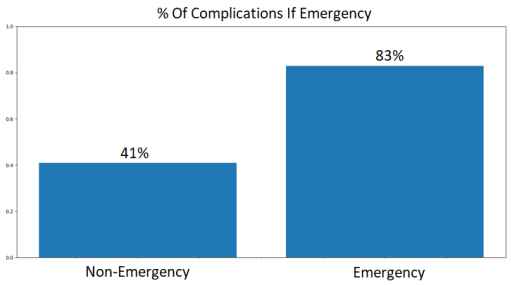


Figure 5.8: Percentage of complications for emergencies vs. non emergencies

As a final example of the results from the exploration phase, Body Mass Index (BMI), a relevant surgery prehabilitation objective, was also analyzed. The first step was to divide the range of values into bins of size 5. starting from the minimum value up until the maximum value. The conclusions, as seen

on Fig.5.9 and 5.10, were relatively unclear. The BMI distribution graphic seems normal, but in terms of correlation to postoperative complications there are no solid clues that it is directly correlated.

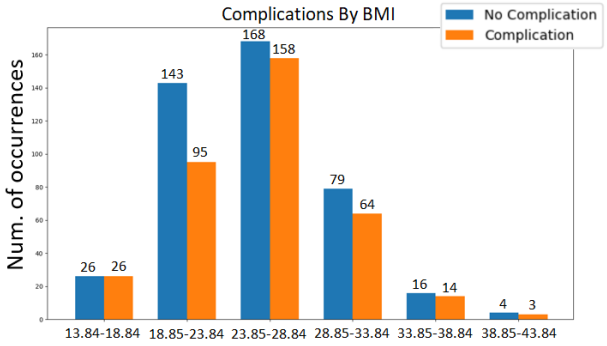


Figure 5.9: BMI variable distribution according to the existence of complications

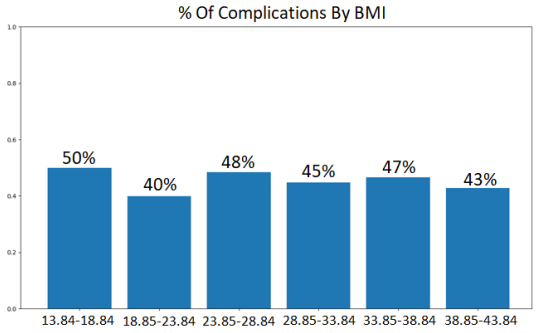


Figure 5.10: Percentage of complications by BMI level

The sparsity of the dataset is also clear upon the inspection of the encompassed cancer locations, there is a wide range of cancer types, more than 30. This variety might prove useful in the future, if specific predictors are developed for representative types of cancer. But it also constitutes a challenge since some of these classes are underrepresented when studied alone, with fewer than 10 individuals. The following chart shows the the principal cancer locations ranked by the rate of postoperative complications for each:

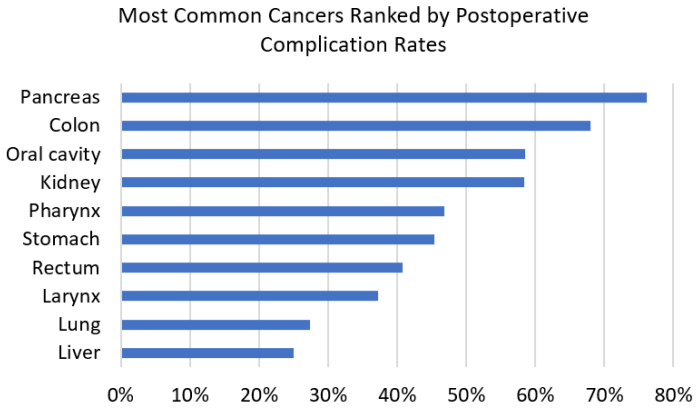


Figure 5.11: The ten most common cancer types ranked by postoperative complication rate

5.2 Model Results

This section will be presented in 6 stages of improvement (Fig.5.12), each representing a modification to the models, or data, that might change the quality of the results. In this context, it is easier to isolate

the performance impact of each step, and tailor an individual solution to the challenges at hand.

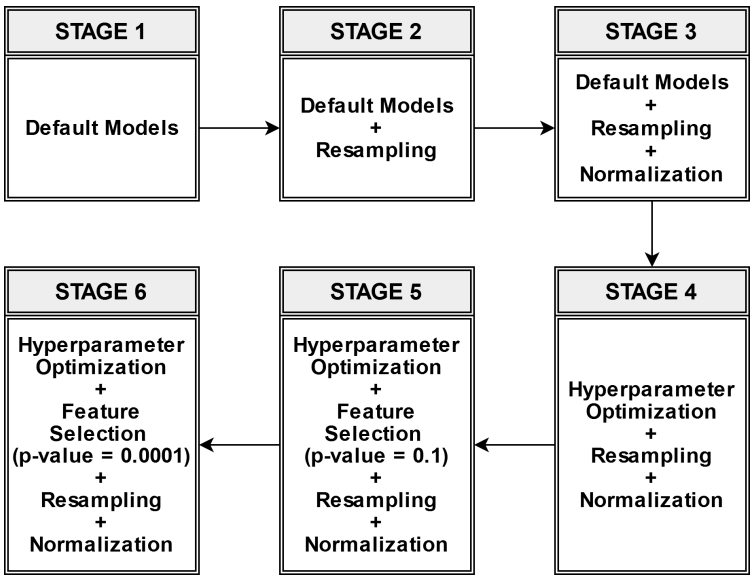


Figure 5.12: The 6 stages of the model development pipeline

For each step, all the models are applied, classification and/or regression, to the 4 outcomes of interest. The results shown are averaged across the 10 fold runs, used for validation purposes. The results of each individual run are used to apply one tail paired *t*-tests, or a Wilcoxon signed-rank test, when the values are not normally distributed. This way it is possible to assess the improvement of each model along the development cycle, and also compare each model to the corresponding baseline at each stage. For classification tasks Naive Bayes is the baseline, and for regression tasks Linear Regression is used. Both chosen precisely due to their simple and naive nature.

On a side note, whenever the recall metric is mentioned it is used as the average predictive capacity for each one of the outcome values. For instance, in a binary problem, recall will represent the average between sensitivity and specificity.

5.2.1 Preliminary Study

In the first study, after the missing values imputation, the set of proposed predictive models are applied using their default settings (as originally defined in the Scikit Learn package) to extract baseline results.

The first outcome under prediction is the existence of postoperative complications. Fig.5.13 shows the average results for various metrics extracted for a 10 fold validation process (according to Table 5.1). In this case the most relevant predictors are RF, LR, XGB, and NB, as the baseline model.

Table 5.1: Primary results for the Existence of Complications prediction

Algorithm	Accuracy	AUC	Recall	Kappa
NB	0.643 ± 0.064	0.710 ± 0.086	0.615 ± 0.068	0.240 ± 0.142
kNN	0.605 ± 0.064	0.636 ± 0.078	0.598 ± 0.064	0.197 ± 0.129
DT	0.609 ± 0.054	0.605 ± 0.055	0.605 ± 0.055	0.210 ± 0.110
RF	0.659 ± 0.061	0.730 ± 0.067	0.649 ± 0.065	0.301 ± 0.130
SVM	0.657 ± 0.043	0.707 ± 0.067	0.637 ± 0.043	0.284 ± 0.089
LR	0.676 ± 0.063	0.705 ± 0.073	0.666 ± 0.063	0.336 ± 0.128
XGB	0.656 ± 0.065	0.708 ± 0.059	0.649 ± 0.069	0.299 ± 0.137
MLP	0.616 ± 0.039	0.672 ± 0.036	0.610 ± 0.032	0.221 ± 0.066

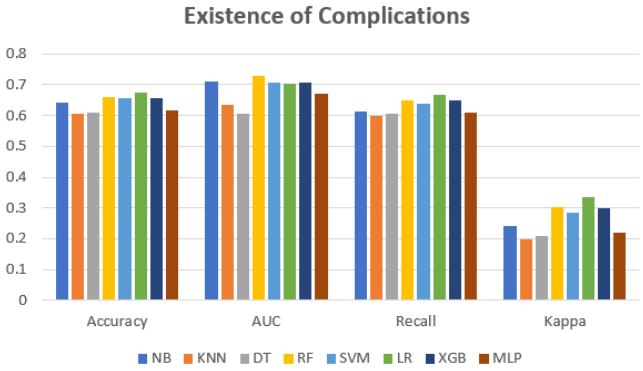


Figure 5.13: Graph with primary results for the Existence of Complications prediction

The second outcome is the severity of the postoperative complications, and in this case both regression and classification could be applied to predict the severity level. Table 5.2 and Fig.5.14 show the performance of the default models for classification.

Table 5.2: Primary results for the Severity of Complications prediction

Algorithm	Accuracy	AUC	Recall	Kappa
NB	0.073 ± 0.013	0.587 ± 0.057	0.197 ± 0.064	0.023 ± 0.016
kNN	0.519 ± 0.033	0.589 ± 0.022	0.167 ± 0.039	0.122 ± 0.054
DT	0.402 ± 0.049	0.530 ± 0.032	0.178 ± 0.057	0.094 ± 0.071
RF	0.539 ± 0.033	0.627 ± 0.033	0.159 ± 0.022	0.145 ± 0.067
SVM	0.533 ± 0.008	0.626 ± 0.037	0.112 ± 0.002	0.021 ± 0.011
LR	0.521 ± 0.022	0.645 ± 0.046	0.154 ± 0.029	0.152 ± 0.045
XGB	0.523 ± 0.025	0.634 ± 0.043	0.168 ± 0.034	0.135 ± 0.054
MLP	0.492 ± 0.040	0.622 ± 0.072	0.179 ± 0.071	0.158 ± 0.058

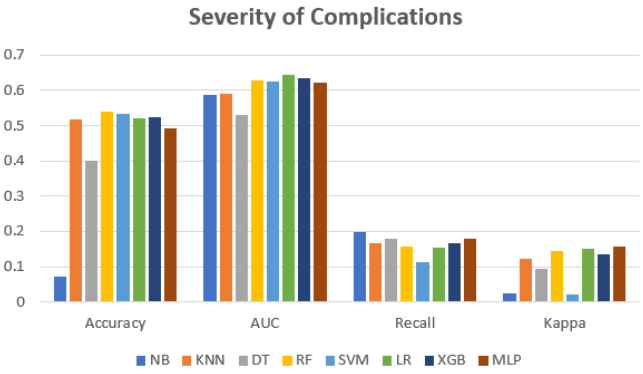


Figure 5.14: Graph with primary results for the Severity of Complications prediction

Table 5.3 and Fig.5.15 provide the results for the regression approach. In this case, in order to enable some type of comparison between the two approaches, the predictions of the regression models are rounded to obtain discrete values. This, allows the output of the regression model to be measured through common classification metrics, like accuracy, kappa statistic and recall. It is important to note that the AUC metric is not included since it needs the class probabilities which only a classifier would be able to retrieve. The results of this type of evaluation are presented in Table 5.4 and Fig.5.16.

It's important to note that the results show a negative R^2. This metric traduces the comparison of the model fit to a model of order 0 (just fitting a constant, usually the mean), both by minimizing a squared loss. Since cross validation leaves out data, it can happen that the mean of the test set is different from the mean of the training set, which alone can induce a higher squared error in the prediction versus just predicting the mean of the test data, resulting in a negative R^2 score.

Table 5.3: Primary results for the Severity of Complications (regression) prediction

Algorithm	MAE	RMSE	R^2
Linear	1.391 ± 0.156	1.799 ± 0.215	0.085 ± 0.142
Ridge	1.363 ± 0.163	1.763 ± 0.228	0.125 ± 0.113
Lasso	1.316 ± 0.138	1.66 ± 0.222	0.221 ± 0.133
SVR	1.195 ± 0.193	1.743 ± 0.305	0.138 ± 0.213
Elastic	1.311 ± 0.141	1.658 ± 0.229	0.222 ± 0.142
kNN	1.340 ± 0.149	1.776 ± 0.280	0.111 ± 0.165
DT	1.521 ± 0.155	2.293 ± 0.186	-0.509 ± 0.294
RF	1.290 ± 0.120	1.673 ± 0.180	0.202 ± 0.154
XGB	1.304 ± 0.146	1.697 ± 0.206	0.185 ± 0.125
PLS	1.289 ± 0.134	1.645 ± 0.208	0.233 ± 0.131
MLPR	1.353 ± 0.149	1.804 ± 0.235	0.076 ± 0.171

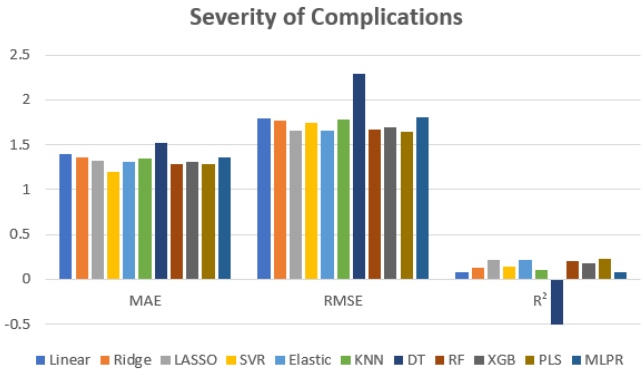


Figure 5.15: Graph with primary results for the Severity of Complications (regression) prediction

Table 5.4: Primary results for the Severity of Complications (discretized) prediction

Algorithm	Accuracy	Kappa	Recall
Linear	0.325 ± 0.051	0.086 ± 0.040	0.149 ± 0.063
Ridge	0.337 ± 0.054	0.101 ± 0.042	0.170 ± 0.078
Lasso	0.295 ± 0.041	0.092 ± 0.055	0.120 ± 0.060
SVR	0.472 ± 0.046	0.121 ± 0.057	0.140 ± 0.060
Elastic	0.297 ± 0.040	0.089 ± 0.058	0.134 ± 0.087
kNN	0.366 ± 0.047	0.113 ± 0.052	0.169 ± 0.076
DT	0.432 ± 0.041	0.105 ± 0.069	0.177 ± 0.041
RF	0.334 ± 0.045	0.113 ± 0.045	0.154 ± 0.080
XGB	0.350 ± 0.043	0.120 ± 0.041	0.151 ± 0.054
PLS	0.342 ± 0.041	0.118 ± 0.051	0.165 ± 0.093
MLPR	0.368 ± 0.044	0.118 ± 0.044	0.191 ± 0.068

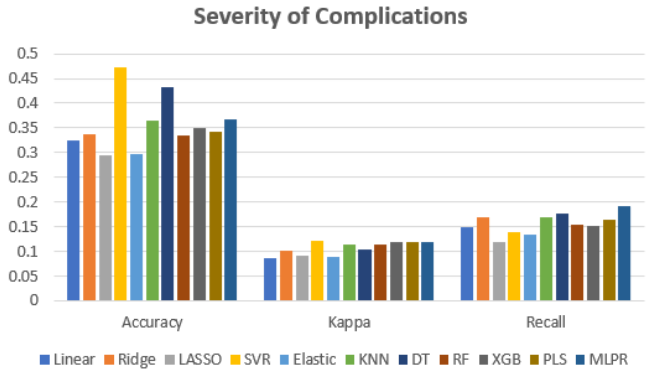


Figure 5.16: Graph with primary results for the Severity of Complications (discretized) prediction

The third outcome is the death probability within a year after surgery. This outcome is not measured though probability values per se, but rather over the binary output of the classifiers. This output can later be compared against the actual data from IPO-Porto to assess model performance. As to the final implementation of the models used for this outcome, actual probabilities can be returned by the models.

Table 5.5: Primary results for Death Within 1 Year prediction

Algorithm	Accuracy	AUC	Recall	Kappa
NB	0.255 ± 0.156	0.515 ± 0.091	0.497 ± 0.059	0.005 ± 0.062
kNN	0.807 ± 0.028	0.681 ± 0.063	0.554 ± 0.038	0.138 ± 0.090
DT	0.748 ± 0.051	0.587 ± 0.064	0.588 ± 0.064	0.173 ± 0.127
RF	0.831 ± 0.028	0.770 ± 0.072	0.561 ± 0.059	0.168 ± 0.160
SVM	0.819 ± 0.004	0.706 ± 0.127	0.500 ± 0	0 ± 0
LR	0.809 ± 0.031	0.734 ± 0.093	0.596 ± 0.058	0.223 ± 0.131
XGB	0.830 ± 0.040	0.729 ± 0.086	0.592 ± 0.063	0.242 ± 0.162
MLP	0.792 ± 0.025	0.719 ± 0.091	0.591 ± 0.052	0.197 ± 0.107

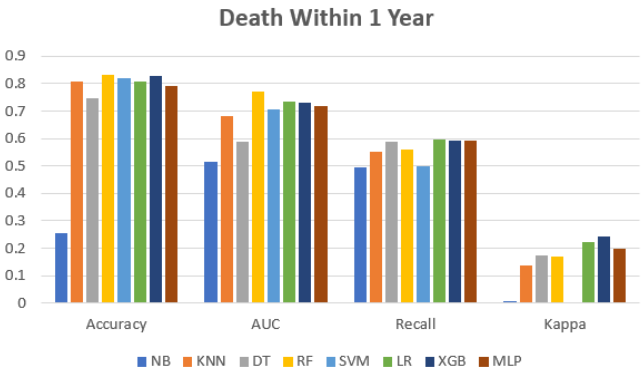


Figure 5.17: Graph with primary results for Death Within 1 Year prediction

The last outcome of interest is the number of days the patient will stay at the intermediate care unit. This outcome is purely numeric and therefore tackled exclusively by regression models. The default models performance is show in Table 5.6 and Fig.5.18.

Table 5.6: Primary results for Days in the ICU prediction

Algorithm	MAE	RMSE	R^2
Linear	1.279 ± 0.162	2.101 ± 0.409	-0.611 ± 1.004
Ridge	1.154 ± 0.142	1.836 ± 0.322	-0.115 ± 0.199
Lasso	1.092 ± 0.171	1.785 ± 0.459	-0.009 ± 0.007
SVR	0.947 ± 0.161	1.760 ± 0.452	0.018 ± 0.050
Elastic	1.092 ± 0.171	1.785 ± 0.459	-0.009 ± 0.007
kNN	1.063 ± 0.142	1.834 ± 0.375	-0.103 ± 0.197
DT	1.409 ± 0.161	2.617 ± 0.331	-1.415 ± 1.025
RF	1.090 ± 0.148	1.810 ± 0.398	-0.088 ± 0.294
XGB	1.078 ± 0.161	1.801 ± 0.411	-0.075 ± 0.302
PLS	1.056 ± 0.141	1.742 ± 0.396	0.026 ± 0.068
MLPR	1.259 ± 0.119	1.970 ± 0.280	-0.322 ± 0.372

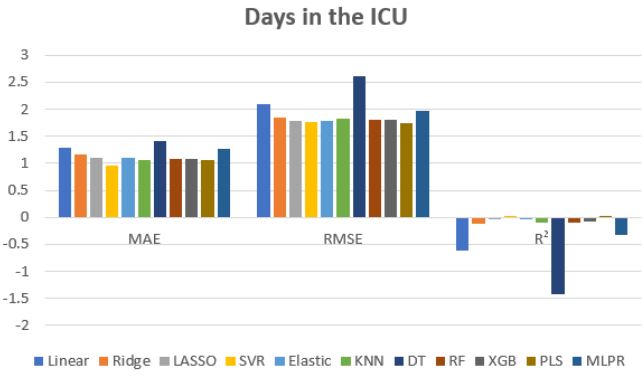


Figure 5.18: Graph with primary results for Days in the ICU prediction

5.2.2 Resampling

Data resampling was applied to the datasets used for classification only, which is the typical approach, although there are preprocessing techniques that are able to deal with data skew in continuous settings [7]. The proposed strategy (section 4.2.4) combines undersampling with oversampling and should help

to clear the bias towards the majority class or at least make it more evident.

The results from the resampled version are compared to the default models through a single tail paired *t*-test. Each algorithm runs 10 times, on the 10 folds across the entire dataset. The results that statistically improved with a *p*-value ≤ 0.05 are in bold.

For the existence of postoperative complications there are no major improvements, since it was a relatively balanced problem from the beginning. With the exception of Naive Bayes, which improved in Recall, Kappa and Accuracy. Table 5.7 shows the results for this step and Fig.5.19 illustrates the improvement since the previous stage.

Table 5.7: Results after resampling for the prediction of complications

Algorithm	Accuracy	AUC	Recall	Kappa
NB	**0.664 ± 0.067**	0.711 ± 0.078	**0.641 ± 0.070**	**0.293 ± 0.144**
kNN	0.601 ± 0.076	0.647 ± 0.087	0.604 ± 0.076	0.205 ± 0.152
DT	0.620 ± 0.060	0.620 ± 0.060	0.620 ± 0.060	0.238 ± 0.118
RF	0.646 ± 0.059	**0.708 ± 0.069**	0.648 ± 0.060	0.293 ± 0.118
SVM	0.646 ± 0.059	**0.695 ± 0.066**	0.641 ± 0.063	0.283 ± 0.124
LR	0.652 ± 0.052	0.685 ± 0.067	0.651 ± 0.053	0.300 ± 0.106
XGB	0.645 ± 0.068	0.705 ± 0.069	0.646 ± 0.070	0.289 ± 0.138
MLP	0.646 ± 0.066	0.698 ± 0.066	0.646 ± 0.064	0.290 ± 0.129

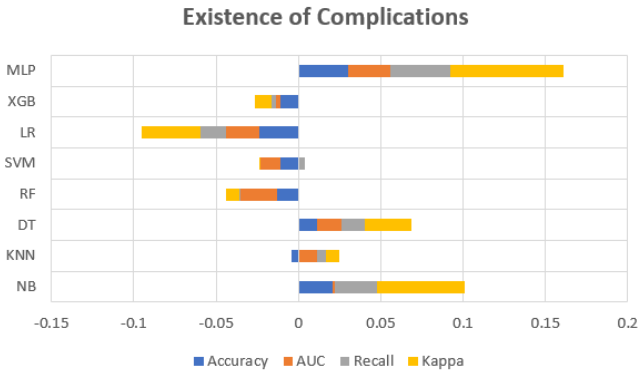


Figure 5.19: Improvement after resampling for the prediction of complications

The severity prediction did suffer some changes after the resampling techniques. Across the entire range of models, the Kappa statistic had improvements with a *p*-value inferior to 0.05. Random Forests, SVM, Logistic regression and XGB improved their Recall. The AUC metric had no significant improvements. And finally, the accuracy fell across the range of models that can longer hide their bias in an imbalanced dataset. Table 5.8 shows the results for this step and Fig.5.20 illustrates the improvement since the previous stage for the severity prediction.

Table 5.8: Results after resampling for complications' severity prediction

Algorithm	Accuracy	AUC	Recall	Kappa
NB	**0.326 ± 0.115**	0.584 ± 0.061	0.197 ± 0.073	**0.096 ± 0.065**
kNN	**0.115 ± 0.024**	**0.562 ± 0.058**	0.187 ± 0.081	**0.026 ± 0.032**
DT	**0.189 ± 0.058**	0.535 ± 0.043	0.186 ± 0.075	0.050 ± 0.063
RF	**0.213 ± 0.039**	0.648 ± 0.057	**0.205 ± 0.074**	**0.073 ± 0.046**
SVM	**0.085 ± 0.044**	0.637 ± 0.066	**0.234 ± 0.068**	**0.038 ± 0.038**
LR	**0.111 ± 0.041**	0.640 ± 0.069	**0.239 ± 0.082**	**0.032 ± 0.036**
XGB	**0.218 ± 0.037**	0.627 ± 0.063	**0.208 ± 0.074**	**0.076 ± 0.040**
MLP	**0.215 ± 0.044**	0.633 ± 0.070	0.222 ± 0.075	**0.069 ± 0.036**

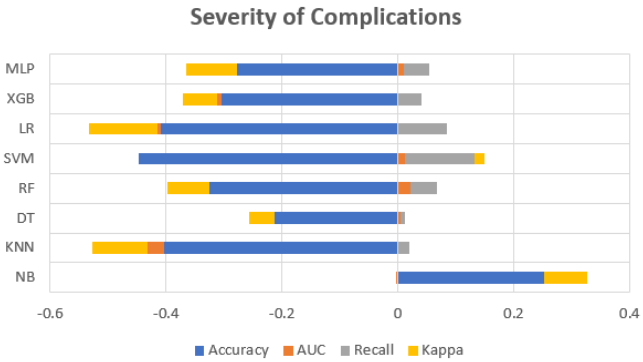


Figure 5.20: Improvement after resampling for complications' severity prediction

For the prediction of death within 1 year, as expected, there were some improvements in recall, except for NB and DT. SVM and RF improved their kappa; NB improved its AUC. Similarly to what happened with severity, the accuracies dropped as a result of the imbalance correction. Table 5.9 shows the results for the resampling step and Fig.5.21 illustrates the improvement since the previous stage for the 1 year death prediction.

Table 5.9: Results after resampling for the 1 year death prediction

Algorithm	Accuracy	AUC	Recall	Kappa
NB	**0.709 ± 0.193**	**0.646 ± 0.117**	0.536 ± 0.090	0.084 ± 0.131
kNN	**0.526 ± 0.041**	0.685 ± 0.045	**0.626 ± 0.058**	0.136 ± 0.063
DT	**0.683 ± 0.046**	0.632 ± 0.085	0.632 ± 0.085	0.195 ± 0.117
RF	**0.742 ± 0.034**	0.755 ± 0.084	**0.671 ± 0.067**	**0.278 ± 0.092**
SVM	**0.457 ± 0.046**	0.709 ± 0.069	**0.628 ± 0.059**	**0.123 ± 0.056**
LR	**0.625 ± 0.048**	0.731 ± 0.076	**0.672 ± 0.069**	0.213 ± 0.087
XGB	**0.763 ± 0.041**	0.752 ± 0.072	**0.676 ± 0.068**	0.303 ± 0.106
MLP	**0.696 ± 0.077**	0.728 ± 0.091	**0.684 ± 0.092**	0.269 ± 0.140

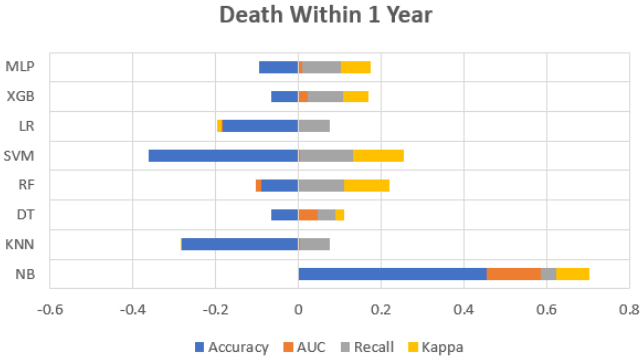


Figure 5.21: Improvement after resampling for the 1 year death prediction

5.2.3 Feature Scaling

After the application of Resampling, clearing the fog of imbalanced data, the process followed on to apply Normalization. Just like before, the results of the Normalized preprocessing are compared to the ones obtained after resampling, using a single tail paired *t*-test.

For the existence of postoperative complications, the performance of NB and RF slightly decreased with regards to accuracy, recall and kappa. kNN improved on recall and kappa. And finally, SVM and LR improved AUC slightly. As for DT, XGB and MLP, the results didn't have any statistically relevant differ-

ence. Table 5.10 shows the results for the normalization step and Fig.5.22 illustrates the improvement since the previous stage for the complications existence prediction.

Table 5.10: Results after normalization for complications prediction

Algorithm	Accuracy	AUC	Recall	Kappa
NB	**0.648 ± 0.072**	0.707 ± 0.079	**0.622 ± 0.075**	**0.254 ± 0.157**
kNN	0.628 ± 0.049	0.688 ± 0.053	**0.636 ± 0.050**	**0.265 ± 0.098**
DT	0.625 ± 0.053	0.626 ± 0.053	0.626 ± 0.053	0.249 ± 0.104
RF	**0.638 ± 0.064**	0.711 ± 0.067	**0.640 ± 0.066**	**0.276 ± 0.129**
SVM	0.662 ± 0.053	**0.720 ± 0.066**	0.659 ± 0.054	0.318 ± 0.108
LR	0.655 ± 0.052	**0.708 ± 0.070**	0.652 ± 0.053	0.304 ± 0.106
XGB	0.645 ± 0.068	0.705 ± 0.069	0.646 ± 0.070	0.289 ± 0.138
MLP	0.660 ± 0.059	0.707 ± 0.063	0.659 ± 0.058	0.316 ± 0.117

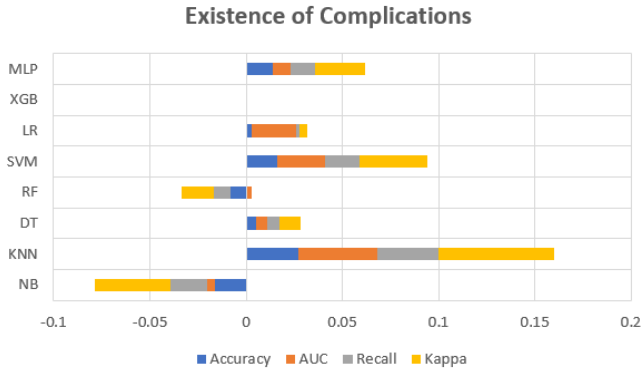


Figure 5.22: Improvement after normalization for complications prediction

On the severity prediction using classification, MLP and NB dropped their performance due to the data normalization process. SVM and LR improved their accuracy, but LR suffered a loss on recall, which shouldn't be a favorable trade-off. Table 5.11 shows the results for this step and Fig.5.23 illustrates the improvement since the previous stage for the severity prediction.

Table 5.11: Results after normalization for the prediction of complications' severity

Algorithm	Accuracy	AUC	Recall	Kappa
NB	**0.316 ± 0.115**	0.58 ± 0.061	0.192 ± 0.073	0.088 ± 0.065
kNN	0.109 ± 0.024	0.555 ± 0.058	0.177 ± 0.081	0.024 ± 0.032
DT	0.194 ± 0.058	0.535 ± 0.043	0.186 ± 0.075	0.051 ± 0.063
RF	0.203 ± 0.039	0.653 ± 0.057	0.210 ± 0.074	0.068 ± 0.046
SVM	**0.174 ± 0.044**	0.62 ± 0.066	0.207 ± 0.068	0.050 ± 0.038
LR	**0.158 ± 0.041**	0.625 ± 0.069	**0.178 ± 0.082**	0.031 ± 0.036
XGB	0.213 ± 0.037	0.627 ± 0.063	0.199 ± 0.074	0.072 ± 0.040
MLP	**0.189 ± 0.044**	0.622 ± 0.070	**0.162 ± 0.075**	**0.031 ± 0.036**

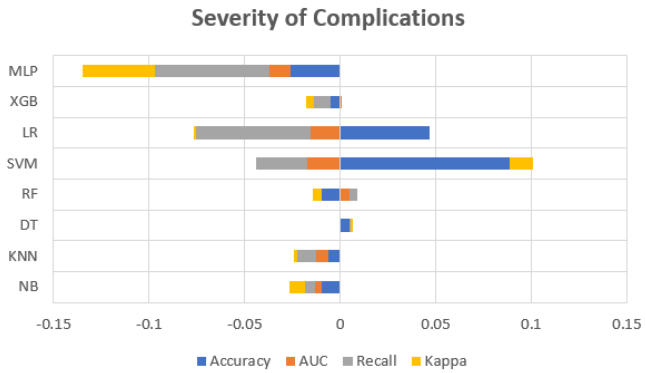


Figure 5.23: Improvement after normalization for the prediction of complications' severity

While evaluating the regression approach, the only significant improvements were on the discrete metrics, which overall tended to improve. The lack of statistical significance for MAE, RMSE and R^2 might be due to the a great range of values across each fold of the validation process. Table 5.12 shows the results for the regression approach, while Table 5.13 shows the results of discrete metrics. Fig.5.24 and Fig.5.25 illustrate its improvement since the previous stages.

Table 5.12: Results after normalization for the complications's severity (regression) prediction

Algorithm	MAE	RMSE	R^2
Linear	1.7E+11 ± 3.2E+11	1.3E+12 ± 2.6E+12	-2.9E+24 ± 7.1E+24
Ridge	1.339 ± 0.151	1.728 ± 0.207	0.158 ± 0.113
LASSO	1.569 ± 0.153	1.909 ± 0.227	-0.021 ± 0.032
SVR	1.18 ± 0.17	1.682 ± 0.258	0.196 ± 0.2
Elastic	1.569 ± 0.153	1.909 ± 0.227	-0.021 ± 0.032
kNN	1.31 ± 0.153	1.759 ± 0.235	0.126 ± 0.147
DT	1.521 ± 0.159	2.29 ± 0.192	-0.506 ± 0.306
RF	1.29 ± 0.119	1.673 ± 0.181	0.202 ± 0.154
XGB	1.305 ± 0.146	1.698 ± 0.206	0.184 ± 0.125
PLS	1.289 ± 0.134	1.645 ± 0.208	0.233 ± 0.131
MLPR	1.457 ± 0.136	1.894 ± 0.176	-0.028 ± 0.213

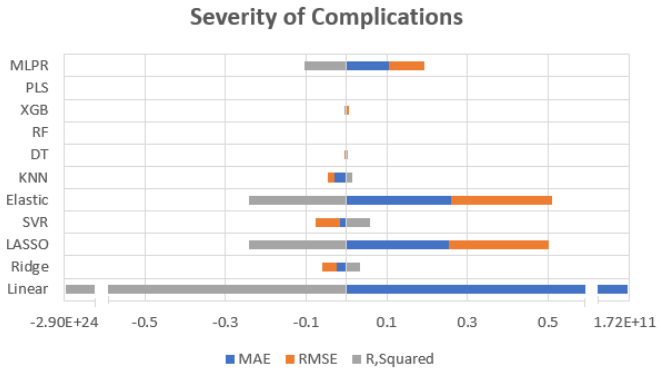


Figure 5.24: Improvement after normalization for the complications's severity (regression) prediction

Table 5.13: Results after normalization for the complication's severity (discretized) prediction

Algorithm	Accuracy	Recall	Kappa
Linear	**0.338 ± 0.040**	**0.178 ± 0.045**	**0.094 ± 0.033**
Ridge	**0.358 ± 0.053**	**0.206 ± 0.072**	**0.121 ± 0.029**
Lasso	**0.040 ± 0.022**	**0.113 ± 0.041**	0 ± 0
SVR	**0.468 ± 0.063**	**0.145 ± 0.057**	**0.154 ± 0.075**
Elastic	**0.040 ± 0.022**	**0.113 ± 0.041**	0 ± 0
kNN	**0.372 ± 0.047**	**0.150 ± 0.078**	**0.067 ± 0.062**
DT	**0.431 ± 0.044**	**0.177 ± 0.042**	**0.105 ± 0.072**
RF	**0.337 ± 0.045**	**0.175 ± 0.081**	**0.115 ± 0.047**
XGB	**0.350 ± 0.043**	**0.170 ± 0.054**	**0.120 ± 0.042**
PLS	**0.342 ± 0.041**	**0.186 ± 0.093**	**0.118 ± 0.051**
MLPR	**0.361 ± 0.054**	**0.19 ± 0.084**	**0.112 ± 0.048**

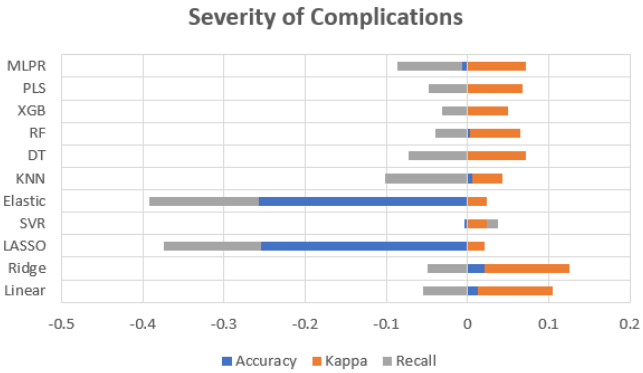


Figure 5.25: Improvement after normalization for the complication's severity (discretized) prediction

Regarding the probability of death, there is only one improvement to highlight with SVMs, more precisely on accuracy and kappa, which is very positive. kNN ended up slightly reducing accuracy, and the MLP reduced its AUC and recall. Table 5.14 shows the results for this normalization step and Fig.5.26 illustrates the improvement since the previous stage for the 1 year death prediction.

Table 5.14: Results after normalization for 1 year death prediction

Algorithm	Accuracy	AUC	Recall	Kappa
NB	0.668 ± 0.226	0.645 ± 0.116	0.645 ± 0.058	0.030 ± 0.010
kNN	**0.462 ± 0.043**	0.686 ± 0.046	0.603 ± 0.054	0.101 ± 0.054
DT	0.696 ± 0.067	0.648 ± 0.086	0.648 ± 0.086	0.228 ± 0.143
RF	0.739 ± 0.039	0.750 ± 0.071	0.666 ± 0.067	0.271 ± 0.010
SVM	**0.669 ± 0.054**	0.732 ± 0.082	0.670 ± 0.057	**0.235 ± 0.083**
LR	0.634 ± 0.070	0.723 ± 0.107	0.652 ± 0.115	0.197 ± 0.150
XGB	0.763 ± 0.041	0.752 ± 0.072	0.676 ± 0.068	0.303 ± 0.106
MLP	0.700 ± 0.060	**0.703 ± 0.112**	**0.655 ± 0.087**	0.234 ± 0.137

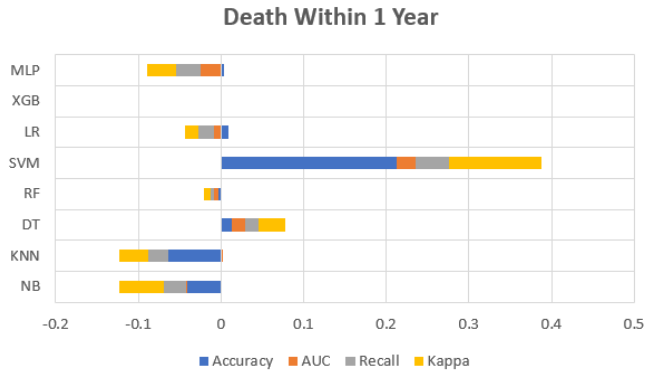


Figure 5.26: Improvement after normalization for 1 year death prediction

Finally, there are no statistically significant changes worth reporting regarding the prediction of days in the ICU, as shown in Table 5.15. The improvement since the previous stage for the days in the ICU prediction is illustrated in Fig.5.27.

Table 5.15: Results after normalization for the days in the ICU prediction

Algorithm	MAE	RMSE	R^2
Linear	1.279 ± 0.162	2.101 ± 0.409	-0.611 ± 1.004
Ridge	1.154 ± 0.142	1.836 ± 0.322	-0.115 ± 0.199
Lasso	1.092 ± 0.171	1.785 ± 0.459	1.785 ± 0.007
SVR	0.947 ± 0.161	1.76 ± 0.452	0.018 ± 0.05
Elastic	1.092 ± 0.171	1.785 ± 0.459	-0.009 ± 0.007
kNN	1.063 ± 0.142	1.834 ± 0.375	-0.103 ± 0.197
DT	1.409 ± 0.161	2.617 ± 0.331	-1.415 ± 1.025
RF	1.09 ± 0.148	1.81 ± 0.398	-0.088 ± 0.294
XGB	1.078 ± 0.161	1.801 ± 0.411	-0.075 ± 0.302
PLS	1.056 ± 0.142	1.742 ± 0.396	0.026 ± 0.068
MLPR	1.254 ± 0.115	1.945 ± 0.263	-0.306 ± 0.426

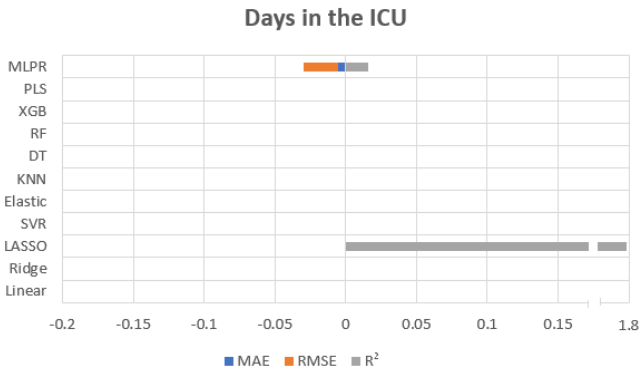


Figure 5.27: Improvement after normalization for the days in the ICU prediction

5.2.4 Hyperparameter Optimization & Feature Selection

After the data preprocessing steps were applied, the development followed to a phase of model optimization. In the context of this project, the models were optimized in relation to the variables used to make the predictions and also in relation to their hyperparameters. Both these aspects are essentially part of the inputs that the algorithms will take to make their predictions.

This stage combines two steps in one. Three sets of input variables are in test, and for each set of selected features, the hyperparameters of the models are optimized to their best possible combination within 100 runs of the Bayesian optimization process.

The strategy adopted consisted on 3 statistical tests to assess dependence or correlation between input and output variables. The tests used are Chi-Squared, ANOVA F-Test and Pearson's Correlation Coefficient, depending on the type of data that is being dealt with.

All these tests are able to return a *p*-value, meaning the probability that there is no association to between two variables in a test instance. In order to retain the most relevant set of variables, this *p*-value was minimized. Two threshold values were chosen, 0.0001 and 0.1. For each prediction outcome there are 3 possible sets of variables that could be used to train the models: the entire set of variables (no selection), the selected set corresponding to the 0.1 *p*-value, and the most restricted set of variables corresponding to 0.0001 *p*-value.

Choosing how restrictive, therefore relevant, the set of input variables should be is not trivial, since part of the available information in the dataset is dropped. Therefore, a wrapper approach was followed. Hyperparameter optimization was performed 3 times for each of the outcomes, using a different set of variables in each time (no selection, 0.1 *p*-value and 0.0001 *p*-value). The inputs to the models, the selected features and the hyperparameters resulting from optimization, are fully disclosed in appendix A and appendix B, respectively.

5.2.4.1 No Feature Selection

The first stage in the optimization phase was to try hyperparameter optimization on models training with the entire set of variables available, 83 variables to be precise.

The prediction of postoperative complications had significant improvements after tweaking the hyperparameters, mainly on DT and SVMs. Although not having proven statistical significance, the entire range of models seemed to perform better on average, as evidenced in Fig.5.28. Table 5.16 shows all the results for this step.

Table 5.16: Results after hyperparametrization for complications prediction

Algorithm	Accuracy	AUC	Recall	Kappa
NB	0.648 ± 0.072	0.707 ± 0.079	0.622 ± 0.075	0.254 ± 0.157
kNN	**0.644 ± 0.052**	0.686 ± 0.049	0.645 ± 0.050	0.287 ± 0.010
DT	**0.657 ± 0.059**	**0.688 ± 0.068**	**0.658 ± 0.58**	**0.313 ± 0.115**
RF	0.650 ± 0.061	**0.696 ± 0.068**	0.654 ± 0.063	0.302 ± 0.124
SVM	**0.680 ± 0.061**	0.716 ± 0.069	**0.677 ± 0.062**	**0.354 ± 0.123**
LR	0.670 ± 0.063	**0.722 ± 0.069**	0.667 ± 0.064	0.334 ± 0.127
XGB	0.651 ± 0.063	0.704 ± 0.068	0.653 ± 0.064	0.302 ± 0.126
MLP	0.660 ± 0.038	0.715 ± 0.064	0.664 ± 0.040	0.323 ± 0.078

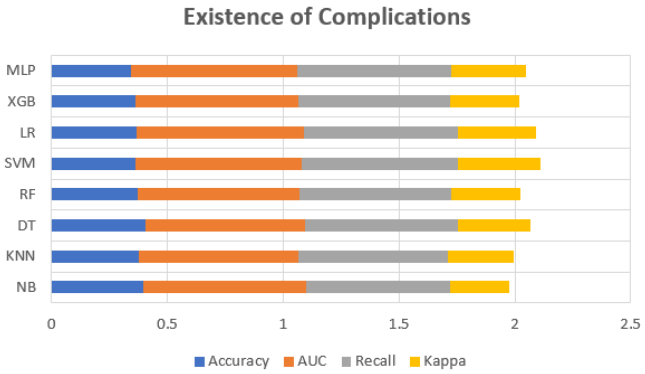


Figure 5.28: Improvement after hyperparametrization for complications prediction

Regarding the severity of postoperative complications, once again the entire range of models had significant improvements. This time many of the models seem to improve their accuracy without the cost of sacrificing other metrics, and therefore reducing or maintaining bias levels. Table 5.17 shows the results for this step and Fig.5.29 illustrates the improvement since the previous stage for the complications' severity prediction.

Table 5.17: Results after hyperparametrization for complications' severity prediction

Algorithm	Accuracy	AUC	Recall	Kappa
NB	0.316 ± 0.115	0.580 ± 0.061	0.192 ± 0.073	0.088 ± 0.065
kNN	**0.086 ± 0.024**	**0.600 ± 0.068**	**0.220 ± 0.084**	0.029 ± 0.027
DT	0.123 ± 0.044	**0.633 ± 0.062**	**0.268 ± 0.096**	**0.051 ± 0.047**
RF	**0.149 ± 0.049**	0.646 ± 0.060	0.221 ± 0.092	0.052 ± 0.049
SVM	**0.089 ± 0.027**	0.637 ± 0.052	0.216 ± 0.091	**0.025 ± 0.026**
LR	**0.134 ± 0.041**	0.635 ± 0.068	0.202 ± 0.098	0.041 ± 0.037
XGB	**0.250 ± 0.037**	0.614 ± 0.062	**0.238 ± 0.071**	0.091 ± 0.039
MLP	0.214 ± 0.070	**0.599 ± 0.070**	**0.202 ± 0.090**	**0.065 ± 0.063**

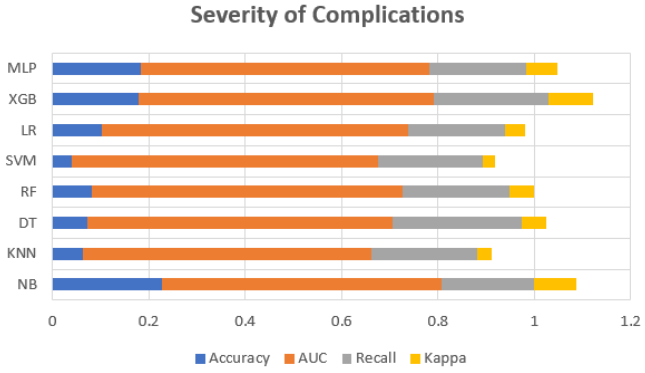


Figure 5.29: Improvement after hyperparametrization for complications' severity prediction

When viewing this problem from a regression perspective, the results are also better since the error metrics (MAE, RMSE, R^2) seem to generally drop, which means the models are closer than they previously were to the real value for the severity scale. Table 5.18 shows the results for the regression approach, while Table 5.19 shows the results of discrete metrics. Fig.5.30 and Fig.5.31 illustrate the improvement since the previous stage.

Table 5.18: Results after hyperparametrization for complications' severity (regression) prediction

Algorithm	MAE	RMSE	R^2
Linear	1.1E+9 ± 3.0E+9	9.3E+9 ± 2.7E+10	-2.6E+20 ± 8.1E+20
Ridge	**1.283 ± 0.138**	**1.631 ± 0.211**	**0.248 ± 0.123**
Lasso	**1.439 ± 0.258**	**2.378 ± 0.318**	**-0.584 ± 0.123**
SVR	1.178 ± 0.166	**1.671 ± 0.252**	**0.204 ± 0.205**
Elastic	**1.439 ± 0.258**	**2.378 ± 0.318**	**-0.584 ± 0.123**
kNN	1.291 ± 0.177	**1.707 ± 0.277**	**0.178 ± 0.166**
DT	**1.170 ± 0.172**	**1.885 ± 0.251**	**-0.005 ± 0.180**
RF	**1.166 ± 0.198**	1.741 ± 0.290	0.139 ± 0.217
XGB	1.279 ± 0.121	1.700 ± 0.202	0.183 ± 0.112
PLS	1.274 ± 0.143	1.629 ± 0.218	0.248 ± 0.144
MLPR	**1.289 ± 0.111**	**1.651 ± 0.184**	**0.225 ± 0.145**

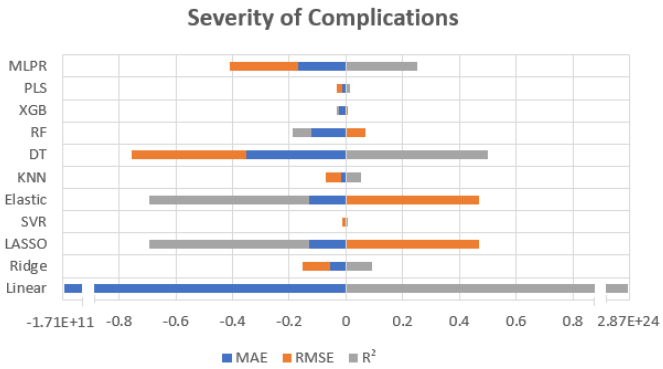


Figure 5.30: Improvement after hyperparametrization for complications' severity (regression) prediction

Table 5.19: Results after hyperparametrization for complications' severity (discretized) prediction

Algorithm	Accuracy	Kappa	Recall
Linear	**0.318 ± 0.040**	0.172 ± 0.054	0.088 ± 0.024
Ridge	**0.324 ± 0.047**	0.184 ± 0.088	**0.112 ± 0.047**
Lasso	**0.532 ± 0.055**	0.125 ± 0.009	0 ± 0
SVR	0.466 ± 0.049	0.155 ± 0.061	0.161 ± 0.077
Elastic	**0.532 ± 0.055**	0.125 ± 0.009	0 ± 0
kNN	0.374 ± 0.054	0.142 ± 0.067	**0.103 ± 0.070**
DT	**0.517 ± 0.042**	0.174 ± 0.024	**0.191 ± 0.065**
RF	**0.487 ± 0.045**	0.138 ± 0.024	**0.142 ± 0.065**
XGB	**0.401 ± 0.037**	0.177 ± 0.067	0.131 ± 0.054
PLS	0.342 ± 0.049	0.187 ± 0.109	0.126 ± 0.068
MLPR	0.335 ± 0.044	0.145 ± 0.058	0.120 ± 0.054

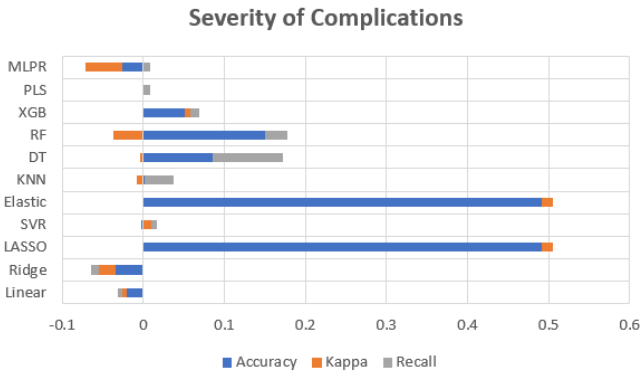


Figure 5.31: Improvement after hyperparametrization for complications' severity (discretized) prediction

In the prediction of death within 1 year, most of the algorithms seem to improve, with the exception of MLP, RF and NB. The differences from the previous stage appear to be residual. Table 5.20 shows the results for this step and Fig.5.32 illustrates the improvement since the previous stage for the 1 year death prediction.

Table 5.20: Results after hyperparametrization for 1 year death prediction

Algorithm	Accuracy	AUC	Recall	Kappa
NB	0.668 ± 0.226	0.645 ± 0.116	0.509 ± 0.058	0.030 ± 0.010
kNN	**0.559 ± 0.029**	0.682 ± 0.045	**0.644 ± 0.048**	**0.161 ± 0.054**
DT	0.668 ± 0.055	0.670 ± 0.077	0.670 ± 0.077	0.230 ± 0.095
RF	**0.700 ± 0.041**	0.752 ± 0.085	0.686 ± 0.064	0.267 ± 0.082
SVM	**0.698 ± 0.051**	0.735 ± 0.085	0.673 ± 0.083	0.250 ± 0.114
LR	0.620 ± 0.052	**0.746 ± 0.087**	0.673 ± 0.082	0.212 ± 0.098
XGB	0.754 ± 0.036	0.759 ± 0.085	0.683 ± 0.059	0.305 ± 0.088
MLP	**0.606 ± 0.096**	**0.733 ± 0.107**	0.647 ± 0.103	0.186 ± 0.136

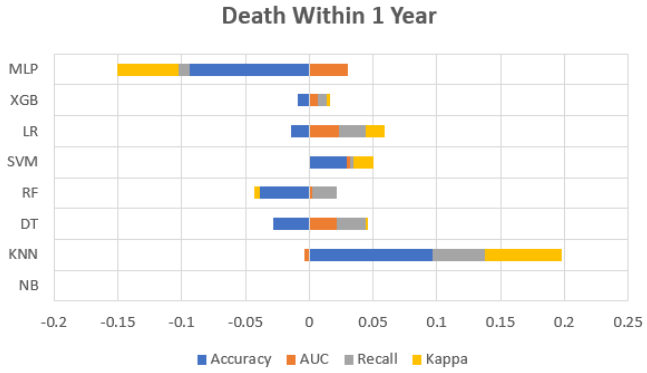


Figure 5.32: Improvement after hyperparametrization for 1 year death prediction

And finally, predicting the days in the ICU presents good results, with MAE and RMSE dropping in most models. While R^2 improves in Ridge, SVR, kNN, DT and MLP, traducing the better fitment of the improved model. Table 5.21 shows the results for this step and Fig.5.33 illustrates the improvement since the previous stage for the days in the ICU prediction.

Table 5.21: Results after hyperparametrization for days in the ICU prediction

Algorithm	MAE	RMSE	R^2
Linear	**1.261 ± 0.150**	2.039 ± 0.315	-0.484 ± 0.720
Ridge	**1.036 ± 0.158**	**1.716 ± 0.412**	**0.060 ± 0.080**
Lasso	1.092 ± 0.171	1.785 ± 0.459	-0.009 ± 0.007
SVR	0.942 ± 0.163	**1.745 ± 0.438**	**0.032 ± 0.060**
Elastic	**1.046 ± 0.162**	1.774 ± 0.459	0.004 ± 0.031
kNN	**1.012 ± 0.132**	**1.738 ± 0.418**	**0.036 ± 0.056**
DT	**0.913 ± 0.154**	**1.813 ± 0.452**	**-0.051 ± 0.113**
RF	**0.926 ± 0.153**	1.760 ± 0.463	0.020 ± 0.059
XGB	**0.977 ± 0.173**	1.803 ± 0.475	-0.032 ± 0.108
PLS	1.056 ± 0.142	1.742 ± 0.396	0.026 ± 0.068
MLPR	**1.056 ± 0.160**	**1.744 ± 0.403**	**0.017 ± 0.161**

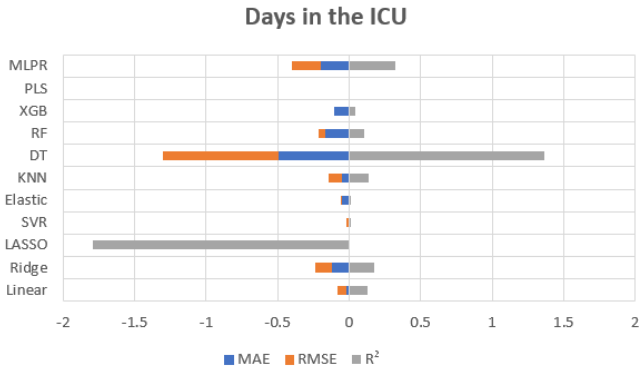


Figure 5.33: Improvement after hyperparametrization for days in the ICU prediction

5.2.4.2 Feature Selection with p-value of 0.1

After trying hyperparameter optimization without any feature selection, the developed followed on to test a wrapper approach to feature selection, reducing dimensionality and noisy variables. More precisely, a *p*-value of 0.1 was imposed on the feature selection process, reducing the number of variables to:

- 41 variables, for the prediction of postoperative complications;
- 41 variables, for the prediction of the severity of complications;

- 33 variables, for the prediction of the probability of death within 1 year;
- 60 variables, for the prediction of days in the ICU.

Starting with the existence of postoperative complications, the results were in the vast majority worse than the previous optimized version, without feature selection. Table 5.22 shows the results for this step and Fig.5.34 illustrates the improvement since the previous stage for complications prediction.

Table 5.22: Results after feature selection (p-value=0.1) and hyperparameter optimization for complications prediction

Algorithm	Accuracy	AUC	Recall	Kappa
NB	0.655 ± 0.067	0.710 ± 0.072	0.628 ± 0.071	0.267 ± 0.148
kNN	0.666 ± 0.034	0.702 ± 0.066	0.649 ± 0.037	0.306 ± 0.074
DT	0.649 ± 0.057	0.696 ± 0.056	0.645 ± 0.062	0.289 ± 0.122
RF	0.648 ± 0.064	**0.714 ± 0.064**	0.647 ± 0.068	0.291 ± 0.133
SVM	0.688 ± 0.068	0.732 ± 0.076	0.682 ± 0.070	0.366 ± 0.141
LR	0.687 ± 0.087	0.729 ± 0.076	0.682 ± 0.089	0.364 ± 0.178
XGB	0.668 ± 0.063	0.711 ± 0.070	0.665 ± 0.063	0.330 ± 0.126
MLP	0.677 ± 0.072	0.704 ± 0.083	0.674 ± 0.072	0.348 ± 0.144

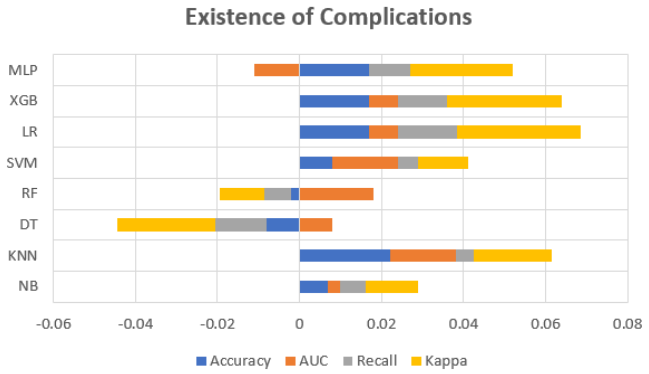


Figure 5.34: Improvement after feature selection (p-value=0.1) and hyperparameter optimization for complications prediction

The prediction of the severity of the complications shows mixed results. kNN and NB have some improvements. While RF, DT, SVM clearly decrease their performance. And there are models such as the MLP, XGB and LR which improve their AUC and recall, but sacrifice their accuracy and kappa statistic. Table 5.23 shows the results for this step and Fig.5.35 illustrates the improvement since the previous stage for complications' severity prediction.

Table 5.23: Results after feature selection (p-value=0.1) and hyperparameter optimization for complications' severity prediction

Algorithm	Accuracy	AUC	Recall	Kappa
NB	**0.407 ± 0.075**	**0.625 ± 0.051**	0.177 ± 0.047	0.112 ± 0.072
kNN	**0.106 ± 0.041**	0.581 ± 0.061	0.206 ± 0.082	0.034 ± 0.034
DT	**0.083 ± 0.022**	0.638 ± 0.058	0.247 ± 0.047	0.038 ± 0.018
RF	**0.066 ± 0.019**	0.652 ± 0.035	0.219 ± 0.069	**0.029 ± 0.020**
SVM	**0.133 ± 0.055**	0.612 ± 0.068	0.239 ± 0.118	0.050 ± 0.054
LR	**0.095 ± 0.031**	0.661 ± 0.066	**0.248 ± 0.081**	0.040 ± 0.030
XGB	**0.095 ± 0.030**	**0.657 ± 0.055**	0.262 ± 0.060	**0.040 ± 0.032**
MLP	**0.118 ± 0.044**	**0.636 ± 0.066**	0.200 ± 0.105	0.034 ± 0.042

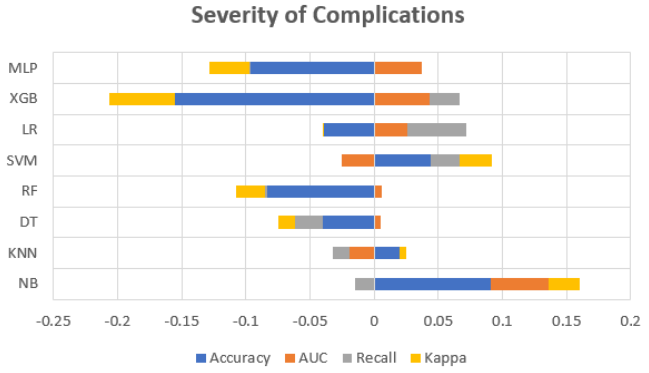


Figure 5.35: Improvement after feature selection (p-value=0.1) and hyperparameter optimization for complications' severity prediction

On the regression approach the picture is similar, with very mixed results, although only a few carry

statistical significance. Elastic regression sees a big improvement across all metrics, except for recall. On the losing side are Ridge, SVM, DT and RF. Table 5.24 shows the results for the regression approach, while Table 5.25 shows the results of discrete metrics. Fig.5.36 and Fig.5.37 illustrate the improvement since the previous stage.

Table 5.24: Results after feature selection (p-value=0.1) and hyperparameter optimization for complications' severity (regression) prediction

Algorithm	MAE	RMSE	R^2
Linear	**1.299 ± 0.143**	**1.673 ± 0.229**	**0.209 ± 0.137**
Ridge	1.268 ± 0.125	1.623 ± 0.192	0.253 ± 0.124
Lasso	1.439 ± 0.258	2.378 ± 0.318	-0.584 ± 0.123
SVR	1.158 ± 0.162	1.707 ± 0.239	0.172 ± 0.182
Elastic	1.391 ± 0.152	**1.763 ± 0.233**	**0.130 ± 0.067**
kNN	1.284 ± 0.150	1.677 ± 0.227	0.208 ± 0.116
DT	1.156 ± 0.181	1.860 ± 0.284	0.017 ± 0.206
RF	1.162 ± 0.202	**1.800 ± 0.307**	**0.082 ± 0.225**
XGB	1.284 ± 0.138	**1.776 ± 0.228**	**0.108 ± 0.141**
PLS	1.268 ± 0.130	1.633 ± 0.198	0.244 ± 0.125
MLPR	**1.255 ± 0.135**	**1.617 ± 0.187**	0.258 ± 0.130

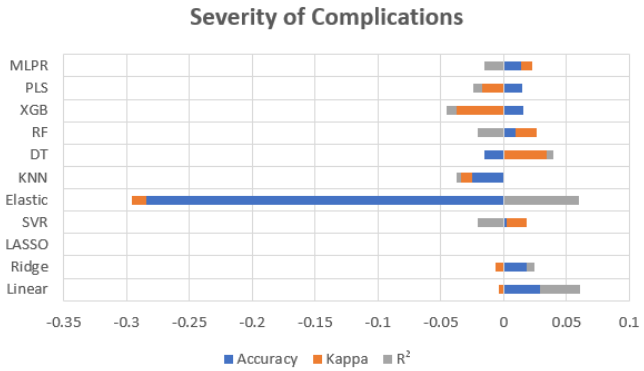


Figure 5.36: Improvement after feature selection (p-value=0.1) and hyperparameter optimization for complications' severity (regression) prediction

Table 5.25: Results after feature selection (p-value=0.1) and hyperparameter optimization for complications' severity (discretized) prediction

Algorithm	Accuracy	Kappa	Recall
Linear	**0.347 ± 0.049**	0.169 ± 0.064	**0.120 ± 0.038**
Ridge	0.342 ± 0.044	0.178 ± 0.106	0.118 ± 0.050
Lasso	0.532 ± 0.055	0.125 ± 0.009	0 ± 0
SVR	**0.468 ± 0.059**	0.172 ± 0.054	**0.140 ± 0.065**
Elastic	0.248 ± 0.040	0.113 ± 0.061	0.060 ± 0.048
kNN	0.349 ± 0.048	**0.133 ± 0.035**	0.099 ± 0.051
DT	0.502 ± 0.035	0.208 ± 0.052	**0.197 ± 0.050**
RF	0.496 ± 0.049	0.155 ± 0.043	0.121 ± 0.056
XGB	0.417 ± 0.028	0.139 ± 0.026	0.123 ± 0.034
PLS	0.357 ± 0.056	0.170 ± 0.087	0.119 ± 0.051
MLPR	0.349 ± 0.057	0.153 ± 0.056	0.105 ± 0.032

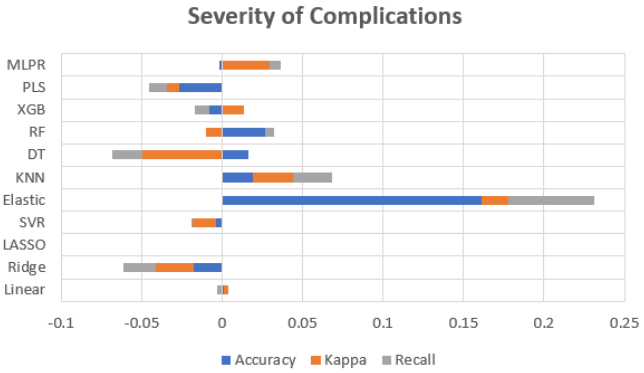


Figure 5.37: Improvement after feature selection (p-value=0.1) and hyperparameter optimization for complications' severity (discretized) prediction

The outcome for probability of Death within a year shows a similar scenario to previous outcomes. DT and NB have statistically significant improvements, specially on recall and kappa. But the rest of the models doesn't seem to benefit from the reduced information, resulting from feature selection. Table 5.26 shows the results for this step and Fig.5.38 illustrates the improvement since the previous stage for 1 year death prediction.

Table 5.26: Results after feature selection (*p*-value=0.1) and hyperparameter optimization for 1 year death prediction

Algorithm	Accuracy	AUC	Recall	Kappa
NB	**0.729 ± 0.192**	**0.694 ± 0.105**	0.546 ± 0.098	0.089 ± 0.156
kNN	**0.657 ± 0.055**	0.721 ± 0.065	0.671 ± 0.083	0.228 ± 0.110
DT	**0.619 ± 0.064**	0.683 ± 0.076	**0.624 ± 0.049**	**0.163 ± 0.065**
RF	**0.669 ± 0.058**	0.762 ± 0.081	0.696 ± 0.068	0.261 ± 0.095
SVM	**0.599 ± 0.040**	0.743 ± 0.077	0.681 ± 0.048	0.210 ± 0.050
LR	0.618 ± 0.107	0.746 ± 0.080	0.697 ± 0.065	0.236 ± 0.081
XGB	**0.738 ± 0.054**	0.744 ± 0.047	0.694 ± 0.075	0.305 ± 0.069
MLP	0.592 ± 0.037	0.739 ± 0.072	0.659 ± 0.053	0.185 ± 0.053

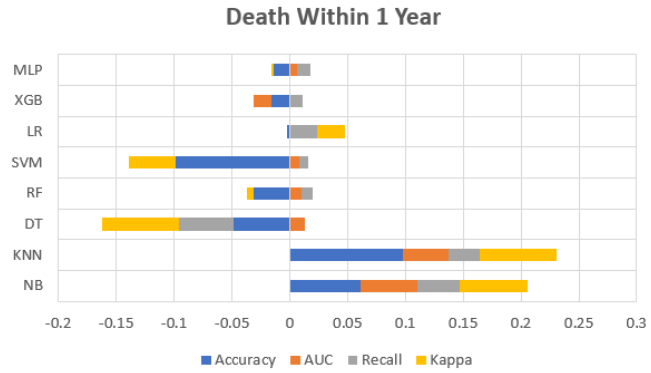


Figure 5.38: Improvement after feature selection (*p*-value=0.1) and hyperparameter optimization for 1 year death prediction

Finally, the prediction of days in the ICU also shows assorted results. PLS and XGB, benefit slightly from the reduction of dimensionality. MLP, RF, Elastic and Linear regression all suffer a negative impact. Table 5.27 shows the results for this step and Fig.5.39 illustrates the improvement since the previous stage for days in the ICU prediction.

Table 5.27: Results after feature selection (*p*-value=0.1) and hyperparameter optimization for the days in the ICU prediction

Algorithm	MAE	RMSE	R^2
Linear	8.5E+7 ± 2.7E+8	7.8E+8 ± 2.5E+9	-2.3E+19 ± 7.2E+18
Ridge	1.043 ± 0.150	1.722 ± 0.404	0.047 ± 0.110
Lasso	1.092 ± 0.171	1.785 ± 0.459	-0.009 ± 0.007
SVR	0.941 ± 0.159	**1.768 ± 0.455**	**0.007 ± 0.066**
Elastic	**1.039 ± 0.161**	**1.779 ± 0.465**	-0.001 ± 0.032
kNN	1.002 ± 0.143	1.73 ± 0.43	0.046 ± 0.070
DT	0.898 ± 0.169	1.782 ± 0.466	-0.012 ± 0.123
RF	0.928 ± 0.154	**1.778 ± 0.464**	**0 ± 0.047**
XGB	0.973 ± 0.166	1.804 ± 0.474	-0.032 ± 0.085
PLS	1.068 ± 0.139	1.745 ± 0.396	0.017 ± 0.118
MLPR	0.017 ± 0.227	1.810 ± 0.486	**-0.047 ± 0.191**

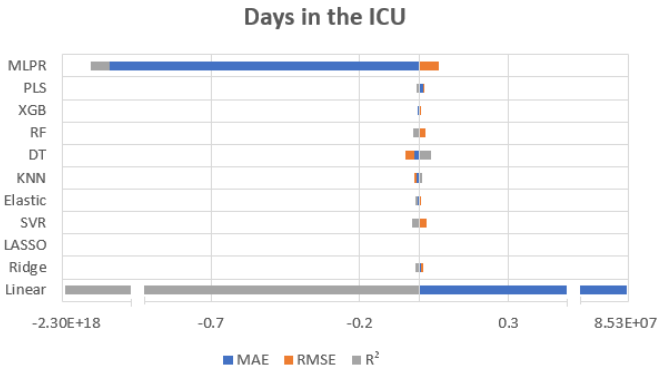


Figure 5.39: Improvement after feature selection (*p*-value=0.1) and hyperparameter optimization for the days in the ICU prediction

5.2.4.3 Feature Selection with p-value of 0.0001

On a last effort to assess the impact of feature selection, which previously showed a mixed set of results, the threshold *p*-value was reduced to 0.0001. This restrictive *p*-value allowed for a more selective process of feature elimination resulting on a reduction to:

- 27 variables, for the prediction of postoperative complications;
- 28 variables, for the prediction of the severity of complications;
- 16 variables, for the prediction of the probability of death within 1 year;
- 27 variables, for the prediction of days in the ICU.

On the prediction of the existence of postoperative complications, the results are mainly worse, accentuating the negative effect of the reduction of information for the models. Table 5.28 shows the results for this step and Fig.5.40 illustrates the improvement since the previous stage for the existence of complications prediction.

Table 5.28: Results after feature selection (*p*-value=0.0001) and hyperparameter optimization for complications prediction

Algorithm	Accuracy	AUC	Recall	Kappa
NB	0.662 ± 0.051	0.703 ± 0.066	0.651 ± 0.056	0.306 ± 0.110
kNN	**0.646 ± 0.049**	0.694 ± 0.074	0.638 ± 0.053	0.279 ± 0.106
DT	0.654 ± 0.063	0.696 ± 0.078	0.657 ± 0.067	0.310 ± 0.131
RF	0.642 ± 0.066	**0.697 ± 0.068**	0.644 ± 0.070	0.283 ± 0.136
SVM	0.661 ± 0.069	**0.702 ± 0.070**	0.654 ± 0.072	0.310 ± 0.144
LR	**0.656 ± 0.065**	**0.705 ± 0.065**	**0.654 ± 0.069**	**0.306 ± 0.136**
XGB	0.645 ± 0.048	0.695 ± 0.070	0.642 ± 0.050	0.282 ± 0.099
MLP	**0.645 ± 0.061**	0.702 ± 0.066	0.651 ± 0.056	0.297 ± 0.113

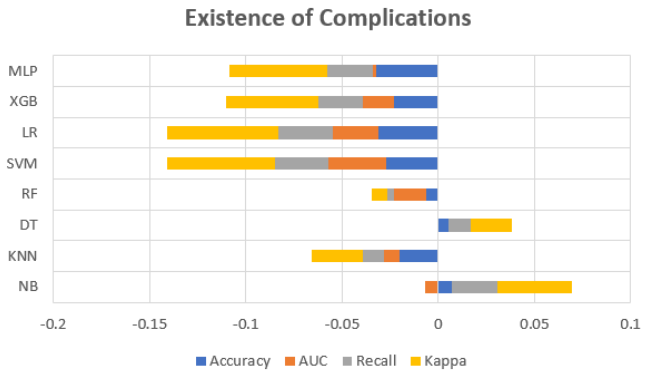


Figure 5.40: Improvement after feature selection (*p*-value=0.0001) and hyperparameter optimization for complications prediction

The prediction of severity shows a mixed scenario. NB improves its recall; LR, XGB, and MLP all improve their accuracy, but at the cost of a reduced AUC and recall. DT, RF register the greatest negative impact, across all metrics. Table 5.29 shows the results for this step and Fig.5.41 illustrates the improvement since the previous stage for the complications' severity prediction.

Table 5.29: Results after feature selection (p-value=0.0001) and hyperparameter optimization for complications' severity prediction

Algorithm	Accuracy	AUC	Recall	Kappa
NB	0.395 ± 0.079	0.645 ± 0.03	**0.232 ± 0.077**	0.148 ± 0.052
kNN	0.123 ± 0.045	0.592 ± 0.079	0.230 ± 0.106	0.045 ± 0.043
DT	**0.046 ± 0.017**	0.604 ± 0.066	**0.161 ± 0.044**	**0.010 ± 0.012**
RF	**0.053 ± 0.018**	0.640 ± 0.048	**0.183 ± 0.073**	**0.015 ± 0.020**
SVM	**0.093 ± 0.041**	0.605 ± 0.077	0.238 ± 0.114	0.040 ± 0.040
LR	**0.145 ± 0.050**	**0.612 ± 0.084**	0.220 ± 0.104	0.052 ± 0.044
XGB	**0.217 ± 0.056**	**0.615 ± 0.069**	0.232 ± 0.081	**0.082 ± 0.054**
MLP	**0.163 ± 0.045**	0.604 ± 0.066	0.189 ± 0.093	0.040 ± 0.041

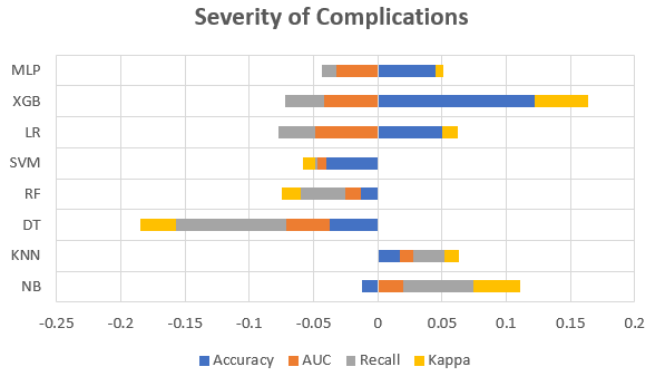


Figure 5.41: Improvement after feature selection (p-value=0.0001) and hyperparameter optimization for complications' severity prediction

On the regression approach for severity, the results are not good for the vast majority of the models. Apart from Elastic regression which actually improved even further with a more restrictive p-value. Table 5.30 shows the results for the regression approach, while Table 5.31 shows the results of discrete metrics. Fig.5.42 and Fig.5.43 illustrate the improvement since the previous stages.

Table 5.30: Results after feature selection (p-value=0.0001) and hyperparameter optimization for complications' severity (regression) prediction

Algorithm	MAE	RMSE	R^2
Linear	1.292 ± 0.137	1.662 ± 0.210	0.219 ± 0.122
Ridge	1.284 ± 0.142	1.646 ± 0.225	0.234 ± 0.131
Lasso	1.439 ± 0.258	2.378 ± 0.318	-0.584 ± 0.123
SVR	**1.187 ± 0.160**	1.747 ± 0.249	0.137 ± 0.159
Elastic	**1.351 ± 0.171**	**1.803 ± 0.269**	**0.091 ± 0.097**
kNN	1.289 ± 0.141	1.675 ± 0.229	0.209 ± 0.120
DT	1.167 ± 0.217	1.865 ± 0.314	0.010 ± 0.260
RF	1.190 ± 0.202	**1.843 ± 0.306**	**0.039 ± 0.220**
XGB	1.300 ± 0.118	1.755 ± 0.208	0.129 ± 0.124
PLS	1.281 ± 0.142	1.645 ± 0.227	0.235 ± 0.135
MLPR	1.278 ± 0.154	1.645 ± 0.228	0.234 ± 0.137

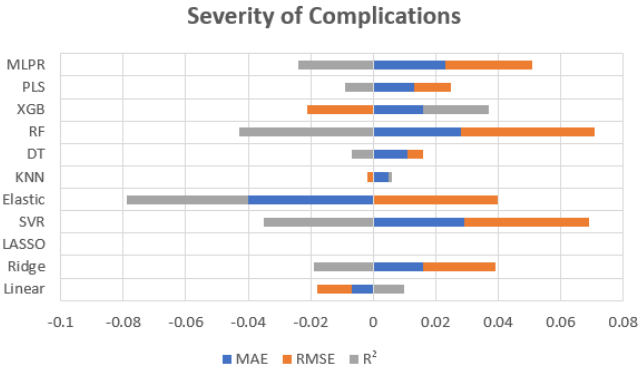


Figure 5.42: Improvement after feature selection (p-value=0.0001) and hyperparameter optimization for complications' severity (regression) prediction

Table 5.31: Results after feature selection (p-value=0.0001) and hyperparameter optimization for complications' severity (discretized) prediction

Algorithm	Accuracy	Kappa	Recall
Linear	0.348 ± 0.057	0.171 ± 0.062	0.117 ± 0.064
Ridge	**0.324 ± 0.047**	0.154 ± 0.073	**0.098 ± 0.061**
Lasso	0.532 ± 0.055	0.125 ± 0.009	0 ± 0
SVR	0.464 ± 0.033	0.157 ± 0.052	0.139 ± 0.044
Elastic	**0.409 ± 0.045**	0.130 ± 0.048	**0.113 ± 0.051**
kNN	0.368 ± 0.057	0.159 ± 0.067	0.123 ± 0.069
DT	0.518 ± 0.058	**0.158 ± 0.023**	0.178 ± 0.049
RF	**0.523 ± 0.053**	0.145 ± 0.036	0.126 ± 0.077
XGB	0.409 ± 0.039	0.153 ± 0.065	0.114 ± 0.069
PLS	0.330 ± 0.043	0.162 ± 0.076	0.108 ± 0.055
MLPR	0.347 ± 0.072	0.183 ± 0.092	0.112 ± 0.064

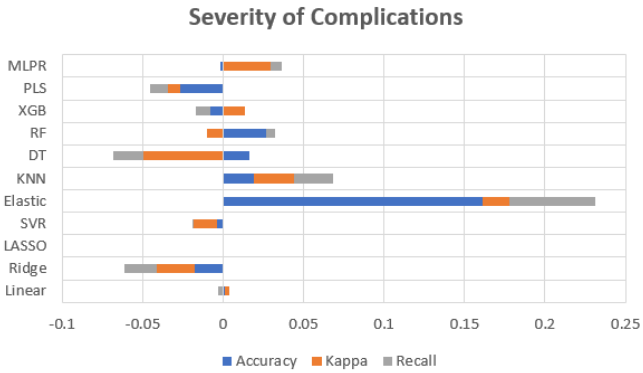


Figure 5.43: Improvement after feature selection (p-value=0.0001) and hyperparameter optimization for complications' severity (discretized) prediction

The probability of death within 1 year is also registering a negative impact in most models. Only DT and NB, as in the previous stage, held performance improvements through a more restrictive feature selection process. Table 5.32 shows the results for this step and Fig.5.44 illustrates the improvement since the previous stage for the 1 year death prediction.

Table 5.32: Results after feature selection (p-value=0.0001) and hyperparameter optimization for 1 year death prediction

Algorithm	Accuracy	AUC	Recall	Kappa
NB	0.728 ± 0.052	0.702 ± 0.072	**0.647 ± 0.075**	**0.241 ± 0.118**
kNN	0.628 ± 0.043	0.686 ± 0.080	0.650 ± 0.065	0.728 ± 0.081
DT	**0.653 ± 0.048**	0.702 ± 0.059	**0.683 ± 0.050**	**0.240 ± 0.066**
RF	**0.630 ± 0.038**	**0.721 ± 0.063**	0.667 ± 0.053	**0.209 ± 0.063**
SVM	0.574 ± 0.059	0.712 ± 0.089	0.661 ± 0.069	0.184 ± 0.084
LR	**0.577 ± 0.052**	0.712 ± 0.103	**0.650 ± 0.066**	**0.174 ± 0.078**
XGB	0.716 ± 0.032	**0.700 ± 0.074**	0.689 ± 0.053	0.282 ± 0.077
MLP	0.574 ± 0.053	0.718 ± 0.095	0.650 ± 0.069	0.173 ± 0.082

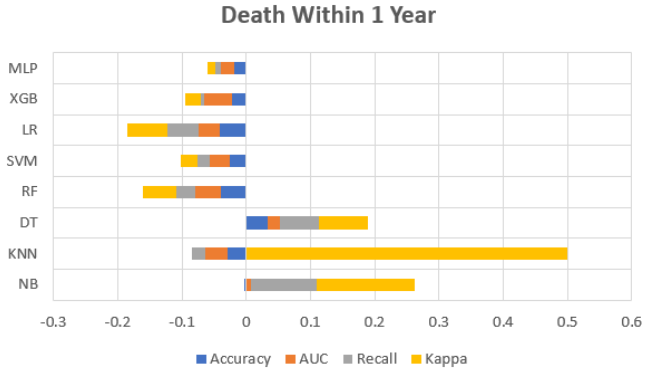


Figure 5.44: Improvement after feature selection (p-value=0.0001) and hyperparameter optimization for 1 year death prediction

Finally, the days in the ICU prediction registers some bad results such as Linear regression, MLP and RF. But also some positive ones, like PLS, XGB and Elastic regression. Table 5.33 shows the results for this step and Fig.5.45 illustrates the improvement since the previous stage for the days in the ICU prediction.

Table 5.33: Results after feature selection (p-value=0.0001) and hyperparameter optimization for the days in the ICU prediction

Algorithm	MAE	RMSE	R^2
Linear	**2.1E+9 ± 6.7E+9**	2.0E+10 ± 6.2E+10	-1.0E+21 ± 3.2E+21
Ridge	1.043 ± 0.143	1.712 ± 0.403	0.058 ± 0.091
Lasso	1.092 ± 0.171	1.785 ± 0.459	-0.009 ± 0.007
SVR	0.942 ± 0.170	1.750 ± 0.474	0.032 ± 0.072
Elastic	**1.012 ± 0.16**	**1.797 ± 0.468**	**-0.023 ± 0.048**
kNN	1.013 ± 0.151	1.722 ± 0.435	0.056 ± 0.051
DT	0.909 ± 0.151	1.784 ± 0.459	-0.013 ± 0.085
RF	**0.946 ± 0.148**	1.77 ± 0.46	0.006 ± 0.052
XGB	0.973 ± 0.167	**1.787 ± 0.471**	**-0.012 ± 0.071**
PLS	**1.056 ± 0.138**	**1.730 ± 0.393**	**0.036 ± 0.100**
MLPR	**1.049 ± 0.140**	1.728 ± 0.397	**0.031 ± 0.149**

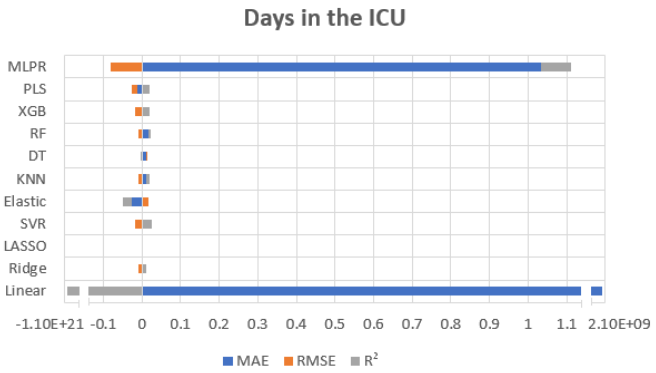


Figure 5.45: Improvement after feature selection (p-value=0.0001) and hyperparameter optimization for the days in the ICU prediction

Once again, the impact (negative or positive) of each change to the models has to be considered in order to choose the models for the final solution. For instance, a model might worsen its results after applying feature selection, but does it compensate to loose a small percentage of accuracy in order to reduce data collection labor at cancer hospitals? This question is explored in section 5.3.1.

5.3 Associative Models - In Depth Analysis

From the beginning of the model development, all models based on Decision Trees showed promising results, such as DT, RF and XGB. This type of algorithms is commonly found in clinical prognostication studies, as illustrated in 3.2. Their popularity is not only due to their relevant performance but also to their ease of interpretation. The difference lies in the way the decision boundary can be visualized, while being able to understand the factors that contributed to that decision. In other types of algorithms, it's possible to obtain a graphical representation of a two dimensional problem, or even a three dimensional one, but as soon as we get to more complex problems we are not able to draw the line. The graphical tree representation offers a unique perspective into how the decisions are made, turning away from the usual black-box notion associated to other algorithms, such as the MLP, for instance.

As an extension to the results obtained from this study, it was possible to explore and improve the traditional visualization associated with tree-based algorithms. Sometimes it is hard to understand where a DT might be struggling to make the right decisions. A simple solution would be to display the error calculated for each node individually. Going further, it is even possible to color code leaf nodes, traducing the error degree associated to the validation process. This specific type of visualization, is an unmatched novelty that can be further extended. Allowing for a quick assessment of the decision process while also showing information about a given performance metric, right at leaf node level.

In this section, a suggestive graphical representation is presented, based on the last stage of development and improvement, where hyperparameter optimization and feature selection, with p-value = 0.0001, was used.

The implementation consists on a reusing the code already available on the Graphviz[2]Python package for plotting various sorts of graphs. Originally, every node uses a random distinctive color to improve the overall interpretability of the tree. As an example, Fig.5.47 shows the tree referring to the prediction of complications' severity is displayed, where every non-leaf node is white and all the leaf nodes can be colored in 10 different ways on a scale from 1 to 10, as illustrated in Fig.5.46. The color of each leaf node is assigned according to the accuracy of each leaf, calculated using the number of correctly guessed instances at the leaf divided by the total amount of instances that landed in that specific leaf.

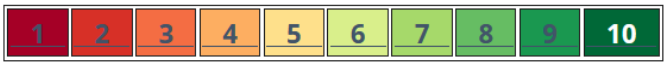


Figure 5.46: Color scheme used for leaf error representation

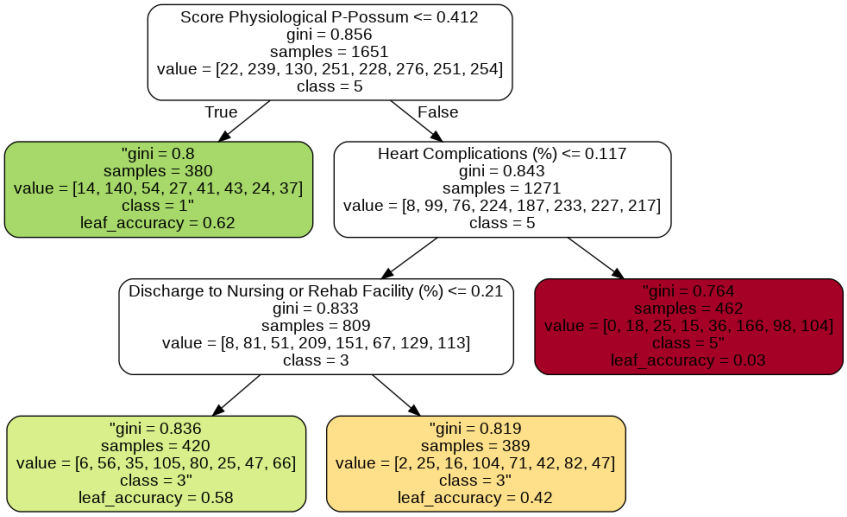


Figure 5.47: Tree graph for the DT from stage 6 used to classify complications' severity

Note that some of these trees might have leafs that classify as the same label. At first it might seem strange but it actually makes sense, according to the tree creation process. No pruning or other constraints were imposed in the hyperparametrization process, because there were not any constraints of that nature. That being said, the algorithm will always try to minimize the Gini Index, within reasonable limits that are imposed, for example, by the minimum number of samples for branching. In analogy, the resultant decision tree behaves like a human that knows 2 ways to the same destiny, but one of them is much safer to get there without ending lost.

[2]https://graphviz.org/

The same principles can be applied to regressor tree-based algorithms, with the difference that the color mapping cannot be made using the accuracy at the leaf node. In continuous settings, the color are attributed depending on the mean absolute error at the leaf node. Calculated using the error of all the predictions resulting from a leaf averaged over the total amount of instances that landed that same node. The following graphs show the same 10 color scheme, over which the error of 4 units is divided. The darker green is attributed to the leafs that have MAE close to 0 and the dark red color is attributed to the leafs that have a MAE closer to 4 units. The units are degrees of the Clavien-Dindo scale, in the tree illustrated by Fig.5.48.

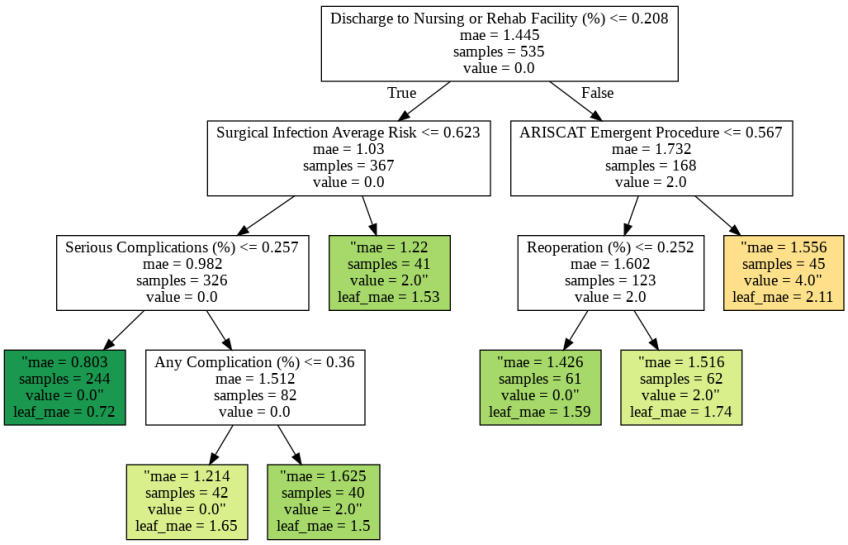


Figure 5.48: Tree graph for the DT from stage 6 used for complications' severity regression

This type of visualization was extracted from the DT algorithm exclusively, but can easily be extended to RF and XGB, with the difference that each model will yield several decision trees, instead of a single one.

5.3.1 Feature Importance in Associative Models

Tree-based models stand out for their intuitive representation, but also for offering information about the importance of each feature in the prediction process. This information might be relevant for doctors in order to reduce the variable collection effort, that can reveal burdensome and too bureaucratic. Right now IPO-Porto is collecting more than 80 pre-operative variables, but not all seem to be of paramount importance for the predictions covered in this project.

Tree-based models offer unique mechanisms to retrieve feature importance ratings, revealing what are the most relevant variables for a certain outcome prediction and giving a great insight into the model knowledge. The models indicate the relative feature importance for each input variable when making

a prediction, as illustrated in Fig.5.49 and Fig.5.50. This type of information can only be matched by certain regression models through variable coefficients, hence the focus on associative models.

All these advantages, contribute for the popularity of this type of models in many areas, including clinical prognostication. A tool that is understandable and transparent contributes to an easier adoption and improved decision confidence.

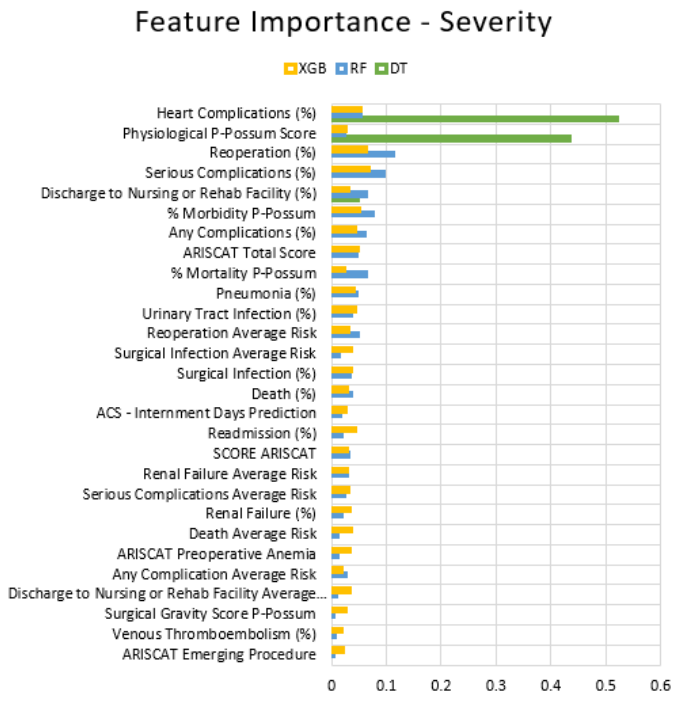


Figure 5.49: Feature importance for the Severity of Complications

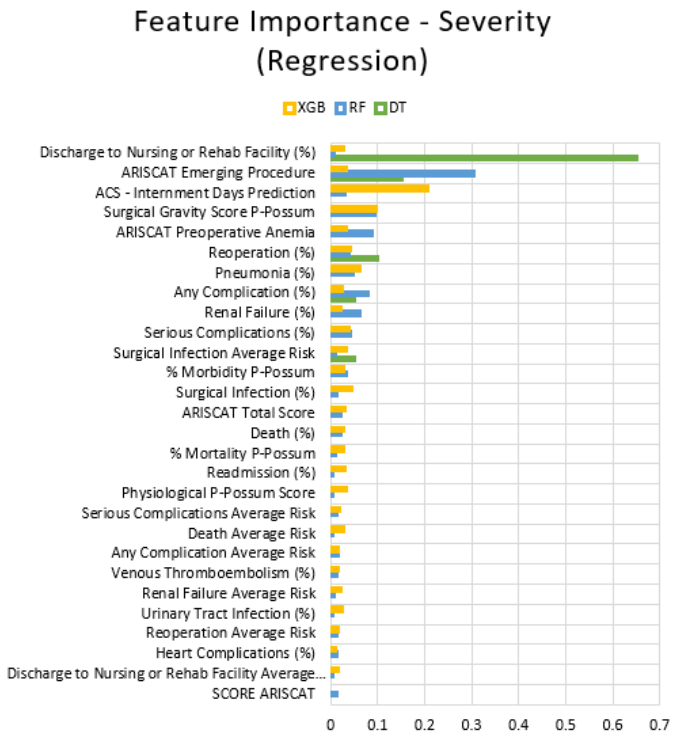


Figure 5.50: Feature importance for the Severity of Complications (regression)

6

Discussion

Contents

6.1 Existence of Postoperative Complications 72
6.2 Severity of Complications 74
6.3 Days Spent in the ICU 78
6.4 Death Probability Within 1 Year 79

Note: As in chapter 5, due to the impossibility of displaying the entire set of results, summarized versions or particular examples of the results can be used. The detailed versions, along with all code, are stored in this thesis' Git Repository: https://github.com/danielmg97/master-thesis-iposcore

After presenting the results in detail, it's important to globally assess the challenges that were initially presented. All the outcomes, in one way or another, were subject to improvements through the various stages of development. The objective of this section is to comment on the reached improvements and the development process. The following sections show graphics that offer a view over the 6 stages of development: 1) default version; 2) resampled; 3) normalized; 4) optimized with no feature selection; 5) optimized with 0.1 *p*-value feature selection; 6) the optimized with 0.0001 feature selection version. In the case of regression problems, there are only 5 stages, since resampling was not applied.

In this chapter, the best models will also be highlighted. The choice process is not trivial here, due to the number of factors influencing the decision, and also the subjectivity associated. For instance, a model can be the best at a certain metric but fail to do so at another. Are all the metrics of equal relevance? What if there are mixed results among 2 different improvement stages? What if 2 algorithms seem to be tied? The answer to all these questions depends on the pursued goals. It depends on the interests of the various future users. Some might desire to have the models with the best recall, but others might sacrifice a small part of the predictive performance for a decreased number of input variables. For these reasons, the models in highlight are merely suggestive, chosen empirically through a Rank Fusion [18] method, as indicated in chapter 4.

6.1 Existence of Postoperative Complications

Starting with postoperative complications, there is a visible positive trend on all metrics, until step 4 or 5. The objective of all the optimizations was the model's sensitivity to both of the output classes (positive and negative), here represented by recall. The same pattern can be seen with AUC and it's more or less the same story with Kappa statistic although it stayed relatively stable from stage 2 to 5.

The best results were achieved on the 4^{th} and 5^{th} stage, where hyperparameters optimization was applied, allied with feature selection, with *p*-value = 0.1, on the 5^{th}. For all the metrics there are two excelling algorithms, SVM and LR. Figures 6.1 to 6.4 show the evolution of the various algorithms, in 4 different perspectives, according to the chosen metrics for performance evaluation.

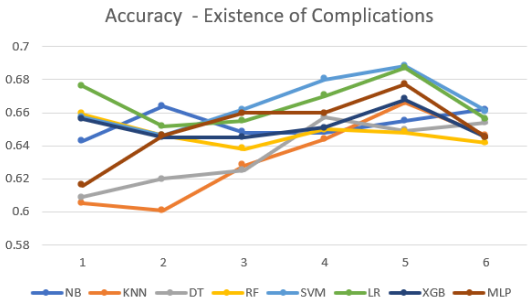


Figure 6.1: Existence of Complications - Global Accuracy

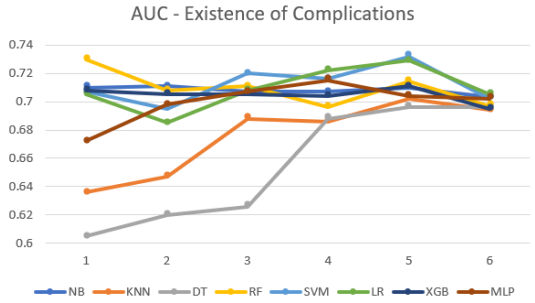


Figure 6.2: Existence of Complications - Global AUC

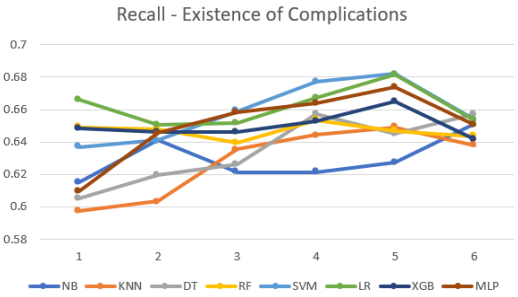


Figure 6.3: Existence of Complications - Global Recall

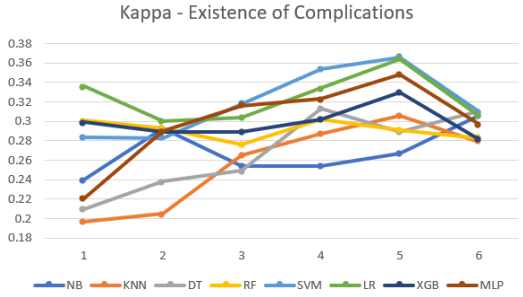


Figure 6.4: Existence of Complications - Global Kappa

This was a problem that seemed relatively straight forward from the beginning with relevant results across all stages of development. The reasons for the success of this prediction are precisely the amount of patients available for each output class (i.e. patients with and without complications). The best model was a SVM from stage 5, ending with the following results: an accuracy of 69%; an average recall score of 68% (averaged for both output classes); an AUC of 0.73; a Cohen's Kappa of 0.37. The Table 6.1 shows the best 5 models according to the Reciprocal Rank Fusion (RRF) results.

Table 6.1: 5 best models for the existence of complications prediction

Algorithm	Stage	Kappa	Recall	AUC	Accuracy	RRF Score
SVM	5	0.366 ± 0.141	0.682 ± 0.070	0.732 ± 0.076	0.688 ± 0.068	0.364
LR	5	0.364 ± 0.178	0.681 ± 0.089	0.729 ± 0.076	0.687 ± 0.068	0.333
SVM	4	0.354 ± 0.123	0.677 ± 0.062	0.716 ± 0.069	0.680 ± 0.061	0.302
LR	4	0.334 ± 0.127	0.667 ± 0.064	0.722 ± 0.068	0.670 ± 0.063	0.277
MLP	5	0.348 ± 0.144	0.674 ± 0.072	0.704 ± 0.083	0.677 ± 0.072	0.260

6.2 Severity of Complications

The complications' severity was the second outcome of interest. For this prediction, two strategies could be applied, classification or regression. The output is a discrete scale, called Clavien-Dindo, ranging from 1 to 8, but it could be modeled continuously if fitting the data this way yielded better results. Both strategies were tested and then compared.

6.2.1 Classification Approach

This challenge in specific revealed to be the hardest outcome to predict out of the 4 initially proposed. Being a multi-class output, with severely underrepresented classes, the results were not surprising. In the 2nd stage of development, resampling techniques were applied to mitigate the imbalance problems. The results remained poor, even when using a mixed strategy like Synthetic Minority Oversampling Technique and Edited Nearest Neighbour (SMOTEENN), due to the reduced number of samples for some of the Clavien-Dindo scale degrees.

Despite the inherent predictive difficulties of this task, the development still shows an improvement in most evaluation metrics after the 2nd development stage, specifically recall score. In fact, out of all the classification tasks in this project, this one shows probably the best recall score improvements. Figures 6.5 to 6.8 illustrate the global improvement of this prediction according to the 4 used metrics.

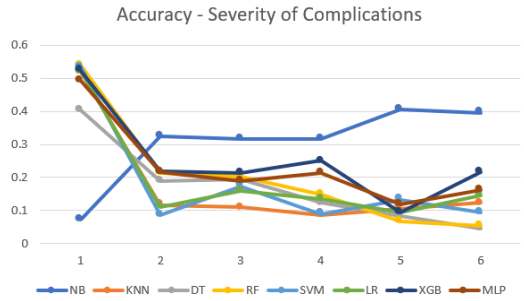


Figure 6.5: Severity of Complications - Global Accuracy

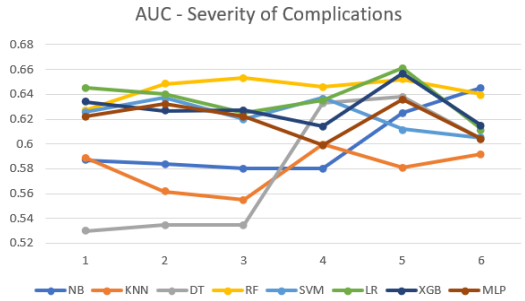


Figure 6.6: Severity of Complications - Global AUC

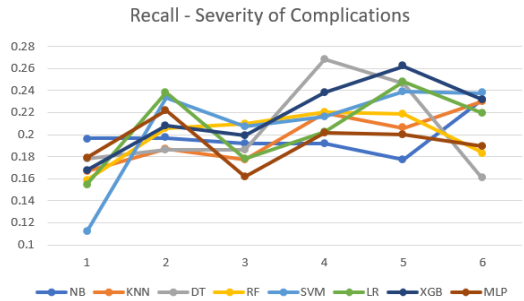


Figure 6.7: Severity of Complications - Global Recall

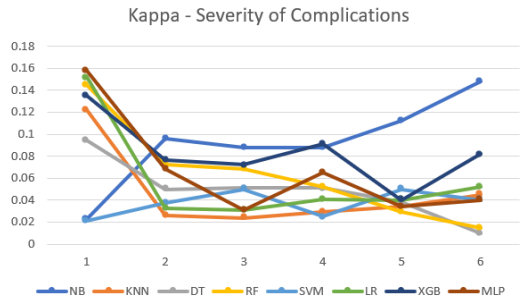


Figure 6.8: Severity of Complications - Global Kappa

The 5 best models for the complications' severity prediction are shown in Table 6.2. Unexpectedly, NB from stage 6 is the best overall model. Firstly, because NB was used as a baseline performance model with little relevant performance expectations resting over one of the more simplistic algorithms in use. And secondly, because this performance peak originated from the 6th development stage, where hyperparameters were optimized while using the most restricted set of input variables, selected with a *p*-value of 0.0001 (appendix B). This NB model, scored an accuracy of about 40%, a recall score of 0.23, an AUC of 0.65 and a kappa statistic of 0.15, which is still relevant performance, considering the values are still above the performance level of a random classifier (chance level of 1/8).

Table 6.2: 5 best models for complication's severity prediction

Algorithm	Stage	Kappa	Recall	AUC	Accuracy	RRF Score
NB	6	0.148 $\pm$ 0.052	0.232 $\pm$ 0.077	0.645 $\pm$ 0.030	0.395 $\pm$ 0.079	0.294
XGB	4	0.091 $\pm$ 0.039	0.238 $\pm$ 0.071	0.614 $\pm$ 0.062	0.250 $\pm$ 0.037	0.253
NB	5	0.112 $\pm$ 0.072	0.177 $\pm$ 0.047	0.625 $\pm$ 0.051	0.407 $\pm$ 0.075	0.25
LR	5	0.040 $\pm$ 0.030	0.248 $\pm$ 0.081	0.661 $\pm$ 0.066	0.095 $\pm$ 0.031	0.248
XGB	5	0.040 $\pm$ 0.032	0.262 $\pm$ 0.060	0.657 $\pm$ 0.055	0.095 $\pm$ 0.030	0.244

6.2.2 Regression Approach

After testing the discrete approach, a continuous strategy was employed. There is a slight decrease of the prediction error overall but the goodness of fit metric, R^2 shows that the models are only slightly better fitted than a model making predictions based on the average output value. The best models are able to predict the output with an error inferior to 1.2 units, in a severity scale of 1 to 8. Figures 6.9 to 6.11 show the results for the various development stages. The sudden spikes are explained due to Linear Regression having high error values and low R^2, that would upset the resolution if included.

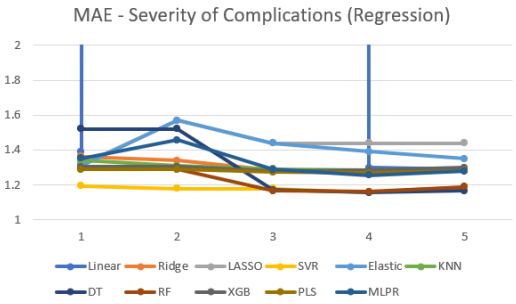


Figure 6.9: Severity of Complications - Global MAE

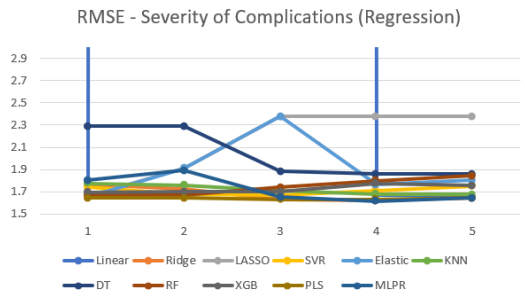


Figure 6.10: Severity of Complications - Global RMSE

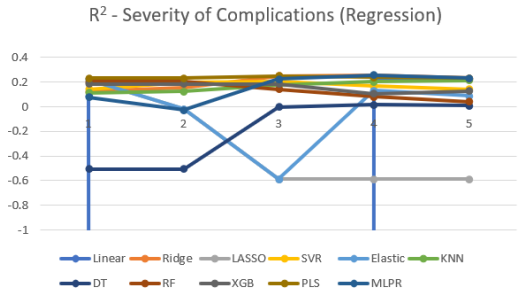


Figure 6.11: Severity of Complications - Global R^2

In order to be able to make comparisons later, the predictions made were rounded in order to obtain scores for accuracy, recall and kappa statistic. Figures 6.12 to 6.14 show the improvements of the regression models during the 5 development stages, which seem to peak at the 3rd stage.

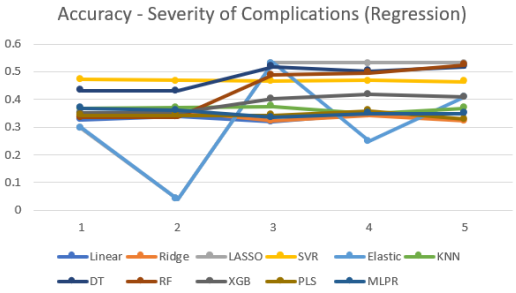


Figure 6.12: Severity of Complications - Global Accuracy

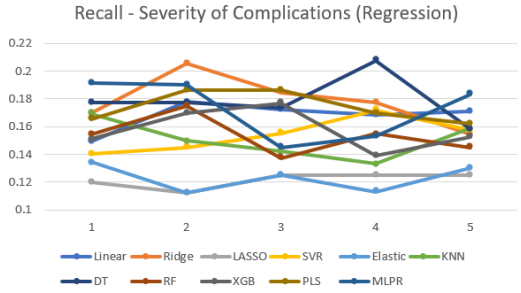


Figure 6.13: Severity of Complications - Global Recall

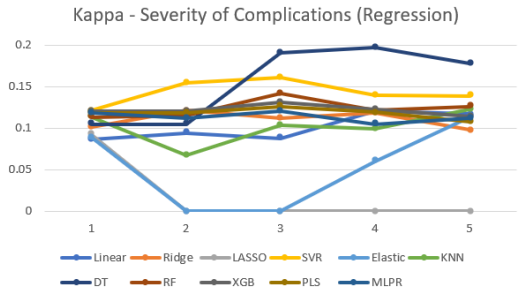


Figure 6.14: Severity of Complications - Global Kappa

The 5 best regression model setups are shown in the Table 6.3. For this ranking, only the MAE, RMSE and R^2 were considered, excluding the metrics used to compare this approach with the discrete one. The best model is the MLP, a fact that might point to a higher complexity problem. Its performance is tightly followed by Ridge regression and PLS, both algorithms armed with mechanisms that try to simplify the problem. Ridge regression applies a penalty to the independent variables which might not be as relevant to the outcome prediction. PLS is a regression method suited for situations where there is

multicollinearity among input variables, or when you have more variables than observations. All aspects that help illustrate the complexity of this challenge.

Table 6.3: 5 best models for complication's severity (regression) prediction

Algorithm	Stage	MAE	RMSE	R^2	RRF Score
MLP	4	1.255 ± 0.135	1.617 ± 0.187	0.258 ± 0.130	0.232
Ridge	4	1.268 ± 0.125	1.623 ± 0.192	0.253 ± 0.124	0.214
PLS	3	1.274 ± 0.143	1.629 ± 0.218	0.248 ± 0.144	0.192
Ridge	3	1.283 ± 0.138	1.631 ± 0.211	0.248 ± 0.123	0.185
PLS	4	1.268 ± 0.130	1.633 ± 0.198	0.244 ± 0.125	0.179

6.2.3 Approach Comparison

In order for the comparison to be possible, the results from the regression model were rounded to the closest integer value. This way, apart from the normal regression evaluation metrics, it was possible to extract the accuracy, recall score and kappa statistic from the model. The last three discrete metrics can be compared to the ones obtained from the classification approach, allowing for a direct predictive performance comparison.

The ranking was established using the results (accuracy, recall and kappa statistic) from all the algorithms in both classification and regression approach. With an accuracy superior to that of other algorithms, while maintaining good results for both recall and kappa statistic.

Table 6.4 shows the best 5 algorithms in order to more accurately assess the best solution. The results seem to point to regression as the best strategy to solve this problem, since only 1 out of the top 5 models are classifiers. At the top we have a DT with an accuracy of 50%, a recall of 0.21 and a kappa statistic of around 0.20.

Table 6.4: 5 best models for severity prediction (approach comparison)

Algorithm	Approach	Stage	Kappa	Recall	Accuracy	RRF Score
DT	REGR	4	0.197 ± 0.050	0.213 ± 0.052	0.502 ± 0.035	0.185
DT	REGR	3	0.191 ± 0.065	0.181 ± 0.024	0.517 ± 0.042	0.168
DT	REGR	5	0.178 ± 0.049	0.165 ± 0.023	0.518 ± 0.059	0.161
NB	CLASS	6	0.148 ± 0.052	0.232 ± 0.077	0.395 ± 0.079	0.154
SVM	REGR	3	0.161 ± 0.077	0.159 ± 0.061	0.466 ± 0.049	0.136

6.3 Days Spent in the ICU

The prediction of days spent in the ICU is a difficult task given the typical short stays of 1 or 2 days. Within the small improvements made, the algorithms decreased their error to a MAE of approximately 1 day. The result is that models will be trying to fit about 350 points with the output 1.0 days and 250 points for 2.0 days. The remaining 200 records will be split between patients that spend 3.0 or 4.0 days, and also patients that spend less than 1 day.

Overall, it is difficult to have a real perception of model performance due to the imbalanced setting, which is confirmed by low R^2 values, meaning that the models perform similarly to a model based on average values. Figures 6.15 to 6.17 show the global view of the metrics used for the days in the ICU regression.

The relevance of this prediction might also be questionable, due to the extremely low variability of the output. Any model with the capacity to predict if a certain patient will spend more or less than the average value, of exactly 2.0 days, should suffice as a management support tool for the ICU. Provided that this is the case, and should health professionals be interested in getting an insight into the time the patient will be spending specifically at the ICU, the obtained models might reveal useful. Reminding that this is a prediction that is not common, and for which IPO-Porto does not have any predictive score presently.

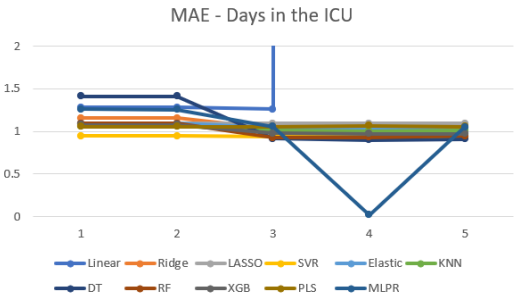


Figure 6.15: Days in the ICU - Global MAE

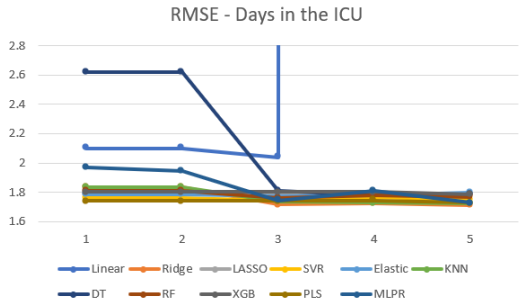


Figure 6.16: Days in the ICU - Global RMSE

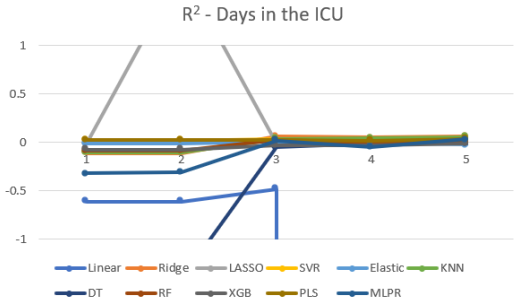


Figure 6.17: Days in the ICU - Global R^2

Once more, the 5 best models are presented in the table 6.5. For the prediction of the days a patient will be spending in the ICU Ridge Regression and kNN were the major contenders. The success of Ridge Regression over other regression models might be a sign that not all independent variables are as important to the outcome prediction, since this is a model that applies penalties in order to reduce the impact of certain variables. The kNN algorithm is also in the top 5, which is slightly surprising due to its simplistic nature.

Table 6.5: 5 best models for days spent in the ICU prediction

Algorithm	Stage	MAE	RMSE	R^2	RRF Score
Ridge	3	1.036 ± 0.158	1.716 ± 0.412	0.060 ± 0.080	0.211
Ridge	5	1.043 ± 0.143	1.712 ± 0.403	0.058 ± 0.091	0.208
kNN	5	1.013 ± 0.151	1.722 ± 0.435	0.056 ± 0.051	0.187
Ridge	4	1.043 ± 0.150	1.722 ± 0.404	0.047 ± 0.110	0.183
kNN	4	1.002 ± 0.143	1.730 ± 0.410	0.046 ± 0.070	0.169

6.4 Death Probability Within 1 Year

This outcome was predicted using a classification approach since the available data was simply a binary variable stating whether the patient had died or not within a 1 year period after surgery. The development efforts soon revealed the severe imbalance of 1:8, towards the negative result for 1 year death. However, this imbalance was not critical since there were still close to 100 patients representing the minority class. Allied to this number of factor, the quality of the data available, contributed greatly for the prediction of death. In fact, the vast majority of the variables selected as the most relevant set for this outcome were results of scores already in use at IPO-Porto. This fact is not a validation of those scores alone, but rather a confirmation that they do a good job standardizing input data and giving rough indications for the patients prognostic.

The development process shows a somewhat stagnant trend across all the evaluation metrics, except for recall, where some improvement were introduced. The models met their peak performance in the 4^{th} and 5^{th} stages as expected. Showing that the restriction of information from step 5 to 6 impacts performance, a reduction of close to 50% of the input data (from 33 to 16 input variables). Figures 6.18 to 6.21 show the global view of the metrics used across the 6 stages of development.

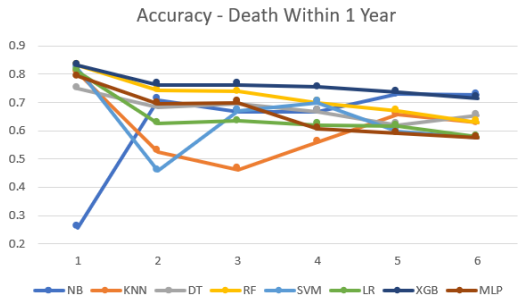


Figure 6.18: Death Within 1 Year - Global Accuracy

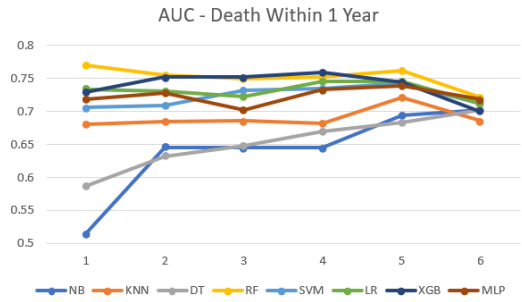


Figure 6.19: Death Within 1 Year - Global AUC

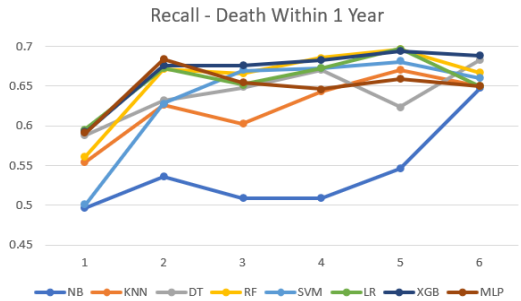


Figure 6.20: Death Within 1 Year - Global Recall

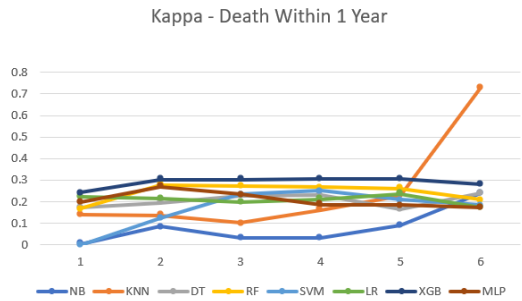


Figure 6.21: Death Within 1 Year - Global Kappa

This outcome shows particularly good results when predicted by tree-based models. More specifically, the most accurate results were achieved by XGB and RF. The best model was the XGB algorithm from 4th stage, with an accuracy of 75%, an average recall score of 0.68, an AUC of 0.76 and a Cohen's Kappa of 0.31. Table 6.6 shows the 5 best models according to the rank established by the Reciprocal Rank Fusion method.

Table 6.6: 5 best models for 1 year death prediction

Algorithm	Stage	Kappa	Recall	AUC	Accuracy	RRF Score
XGB	4	0.305 $\pm$ 0.088	0.683 $\pm$ 0.059	0.759 $\pm$ 0.085	0.754 $\pm$ 0.036	0.320
XGB	5	0.305 $\pm$ 0.069	0.694 $\pm$ 0.075	0.744 $\pm$ 0.047	0.738 $\pm$ 0.054	0.314
RF	5	0.261 $\pm$ 0.095	0.696 $\pm$ 0.068	0.762 $\pm$ 0.081	0.669 $\pm$ 0.058	0.296
RF	4	0.267 $\pm$ 0.082	0.686 $\pm$ 0.064	0.752 $\pm$ 0.085	0.700 $\pm$ 0.041	0.278
LR	5	0.237 $\pm$ 0.081	0.697 $\pm$ 0.065	0.746 $\pm$ 0.008	0.618 $\pm$ 0.107	0.252

7

Conclusion

Contents

7.1 Conclusions 82

7.2 Limitations and Future Work 82

7.3 Scientific Communication 83

As the closing remarks, this chapter is left with the conclusions drawn from this study, the limitations and some work that might still be developed, giving rise to future studies.

7.1 Conclusions

This study provides relevant insights about the world of clinical prognostication, with greater incidence over surgical complications in the oncology domain. From the extensive literature review work, to the specific study to which this thesis is dedicated to. The existent tools are by no means aligned with high throughput and big data systems that are being adopted nowadays, hence the relevance of machine learning studies for clinical prognostication. Models that are able to learn and adapt are equipped with the scalability necessary for foreign adaptation, not only in international context but also different surgical areas.

In this work several supervised learning algorithms were developed and compared, which allowed the prediction of four main outcomes, with the goal of increasing the accuracy of previous risk score tools used at IPO-Porto. The 4 outcomes of interest for this study were: the existence of postoperative complications, the severity of the complications, number of days the patient will spend in the ICU, and the probability of death within 1 year. All predicted with relevant results by the models presented here. Offering the possibility to Portuguese cancer hospitals, more specifically IPO-Porto, to have specialized tools, better suited to their needs and practices. These models introduce the capability of learning from previous data, recycling the good standardization and more or less accurate prediction work already made by older prognostication tools and risk scores. Model interpretability is also covered, by offering new visualization options to tree-based ML models, in order to support medical decision processes. Additionally, information about relevant variables for the outcomes prediction is provided, contributing to more efficient data acquisition processes.

7.2 Limitations and Future Work

This study was developed aiming to predict only 4 outcomes out of many present in the same dataset, such as the total of days a patient will spend in the hospital, or the amount of work a patient will require from nurses. Being a study in the surgical oncology area, it also could be relevant to predict the same, or a different, set of outcomes, but using more specific surgical profiles. For instance, the dataset offers information about the area of the body which is affected by the cancer. Such studies could be important because different types of cancer result in different postoperative risks and complications.

In order to help the study being more inline with hospital interests, it would also be good to have information about the collection effort for each of the input variables. This way, the studies could be

directed towards the use of low effort collection variables. Easing the burden of data collection, that could contribute to the creation of more meaningful and complete datasets in the future.

One of the limitations of this work, is the fact that there is not enough metadata on the dataset, covering acquisition, insertion and other aspects. This aspect makes it extremely difficult to decide without external help what values should be imputed, and what should not. For that reason, some of the variables in this study might have been incorrectly imputed, making the learning process more difficult.

In the future, IPO-Porto will also be releasing new datasets and extensions to already existing ones, which could impact the knowledge fed to the models improving them, especially in outcomes with severe imbalance problems.

The ”final” models resulting from this study offer relevant predictive performance. With this in mind, the hypothesis of creating ensemble methods using the algorithms developed is still in the open, and therefore great potential could be available if explored.

Lastly, an external validation process could not be conducted at the time this project was developed, since it requires the availability of an independent unseen dataset. This step should be crucial to verify the true generalization capabilities of the ML models.

7.3 Scientific Communication

This project's literature review process was developed around 2 main approaches to surgical prognostication, relating traditional statistics and ML models. About this topic and more, the review article “Predicting postoperative complications in cancer patients: a survey bridging classical and machine learning contributions to post-surgical risk analysis” was submitted for publication in the journal Data Mining and Knowledge Discovery[1], by Springer Nature, and is presently under review.

[1] https://www.springer.com/journal/10618

Bibliography

[1] Amin Andalib, Agnihotram V Ramana-Kumar, Gillian Bartlett, Eduardo L Franco, and Lorenzo E Ferri. Influence of postoperative infectious complications on long-term survival of lung cancer patients: a population-based cohort study. *Journal of thoracic oncology*, 8(5):554–561, 2013.

[2] Ahsan M Arozullah, Shukri F Khuri, William G Henderson, and Jennifer Daley. Development and validation of a multifactorial risk index for predicting postoperative pneumonia after major noncardiac surgery. *Annals of internal medicine*, 135(10): 847–857, 2001.

[3] Richard E Bellman. *Adaptive control processes: a guided tour*. Princeton university press, 1961.

[4] Adrien Bibal and Benoît Frénay. Interpretability of machine learning models and representations: an introduction. In *ESANN*, 2016.

[5] Karl Y Bilimoria, Yaoming Liu, Jennifer L Paruch, Lynn Zhou, Thomas E Kmiecik, Clifford Y Ko, and Mark E Cohen. Development and evaluation of the universal acs nsqip surgical risk calculator: a decision aid and informed consent tool for patients and surgeons. *Journal of the American College of Surgeons*, 217(5):833–842, 2013.

[6] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.

[7] Paula Branco, Luís Torgo, and Rita P Ribeiro. Smogn: a pre-processing approach for imbalanced regression. In *First International Workshop on Learning with Imbalanced Domains: Theory and Applications*, pages 36–50. PMLR, 2017.

[8] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[9] A Breugom, E Bastiaannet, CB van den Broek, JWT Dekker, LG van der Geest, C Puylaert, W-H Steup, CJ van de Velde, G-J Liefers, and JE Portielje. Colon cancer patients with postoperative complications have higher risk of recurrences. *Journal of geriatric oncology*, 4:S42, 2013.

[10] Jaume Canet, Lluís Gallart, Carmen Gomar, Guillem Paluzie, Jordi Valles, Jordi Castillo, Sergi Sabate, Valentín Mazo, Zahara Briones, and Joaquín Sanchis. Prediction of postoperative pulmonary complications in a population-based surgical cohort. *Anesthesiology: The Journal of the American Society of Anesthesiologists*, 113(6):1338–1350, 2010.

[11] Siow-Wee Chang, Sameem Abdul-Kareem, Amir Feisal Merican, and Rosnah Binti Zain. Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. *BMC bioinformatics*, 14(1):170, 2013.

[12] Mary E Charlson, Peter Pompei, Kathy L Ales, and C Ronald MacKenzie. A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of chronic diseases*, 40(5):373–383, 1987.

[13] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.

[14] Chee Tang Chin, T Chua, and S LIM. Risk assessment models in acute coronary syndromes and their applicability in singapore. *Ann Acad Med Singapore*, 39(3):216–220, 2010.

[15] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[16] Jonathan A Cook and Gary S Collins. The rise of big clinical databases. *British Journal of Surgery*, 102(2):e93–e101, 2015.

[17] GP Copeland, D Jones, and MPOSSUM Walters. Possum: a scoring system for surgical audit. *British Journal of Surgery*, 78(3):355–360, 1991.

[18] Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, page 758–759, New York, NY, USA, 2009. Association for Computing Machinery. ISBN 9781605584836. doi: 10.1145/1571941.1572114. URL https://doi.org/10.1145/1571941.1572114.

[19] Kwetishe Joro Danjuma. Performance evaluation of machine learning algorithms in post-operative life expectancy in the lung cancer patients. *arXiv preprint arXiv:1504.04646*, 2015.

[20] Carol E. DeSantis, Jiemin Ma, Mia M. Gaudet, Lisa A. Newman, Kimberly D. Miller, Ann Goding Sauer, Ahmedin Jemal, and Rebecca L. Siegel. Breast cancer statistics, 2019. *CA: A Cancer Journal for Clinicians*, 69(6):438–451, 2019. doi: 10.3322/caac.21583. URL https://acsjournals.onlinelibrary.wiley.com/doi/abs/10.3322/caac.21583.

[21] Daniel Dindo, Nicolas Demartines, and Pierre-Alain Clavien. Classification of surgical complications: a new proposal with evaluation in a cohort of 6336 patients and results of a survey. *Annals of surgery*, 240(2):205, 2004.

[22] A Donati, M Ruzzi, E Adrario, P Pelaia, F Coluzzi, V Gabbanelli, and P Pietropaoli. A new and feasible model for predicting operative risk. *British journal of anaesthesia*, 93(3):393–399, 2004.

[23] Jean-Yves Dupuis, Feng Wang, Howard Nathan, Miu Lam, Scott Grimes, and Michael Bourke. The cardiac anesthesia risk evaluation scorea clinically useful predictor of mortality and morbidity after cardiac surgery. *Anesthesiology: The Journal of the American Society of Anesthesiologists*, 94(2):194–204, 2001.

[24] Francesc Formiga, Joan Masip, David Chivite, and Xavier Corbella. Applicability of the heart failure readmission risk score: A first european study. *International journal of cardiology*, 236:304–309, 2017.

[25] Silvia Bueno Garofallo, Daniel Pinheiro Machado, Clarissa Garcia Rodrigues, Odemir Bordim Jr, Renato AK Kalil, and Vera Lúcia Portal. Applicability of two international risk scores in cardiac surgery in a reference center in brazil. *Arquivos brasileiros de cardiologia*, 102(6):539–548, 2014.

[26] Atul A Gawande, Mary R Kwaan, Scott E Regenbogen, Stuart A Lipsitz, and Michael J Zinner. An apgar score for surgery. *Journal of the American College of Surgeons*, 204(2):201–208, 2007.

[27] Louise GH Goh, Satvinder S Dhaliwal, Timothy A Welborn, Peter L Thompson, Bruce R Maycock, Deborah A Kerr, Andy H Lee, Dean Bertolatti, Karin M Clark, Rakhshanda Naheed, et al. Cardiovascular disease risk score prediction models for women and its applicability to asians. *International journal of women's health*, 6:259, 2014.

[28] Prateek K Gupta, Himani Gupta, Abhishek Sundaram, Manu Kaushik, Xiang Fang, Weldon J Miller, Dennis J Esterbrooks, Claire B Hunter, Iraklis I Pipinos, Jason M Johanning, et al. Development and validation of a risk calculator for prediction of cardiac risk after surgery. *Circulation*, 124(4):381–387, 2011.

[29] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

[30] Benjamin T Hazen, Christopher A Boone, Jeremy D Ezell, and L Allison Jones-Farmer. Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics*, 154:72–80, 2014.

[31] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[32] David W Hosmer Jr, Stanley Lemeshow, and Rodney X Sturdivant. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

[33] Mohammad Hossin and MN Sulaiman. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2):1, 2015.

[34] Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. Automated machine learning-methods, systems, challenges, 2019.

[35] Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

[36] Warren Kaplan. Background paper 6.5 cancer and cancer therapeutics. *World Health Organization (ed) Priority medicines for Europe and the world: update*, pages 6–5, 2013.

[37] Muhammad Umer Khan, Jong Pill Choi, Hyunjung Shin, and Minkoo Kim. Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5148–5151. IEEE, 2008.

[38] Dong Wook Kim, Sanghoon Lee, Sunmo Kwon, Woong Nam, In-Ho Cha, and Hyung Jun Kim. Deep learning-based survival prediction of oral cancer patients. *Scientific reports*, 9(1):1–10, 2019.

[39] William A Knaus, Elizabeth A Draper, Douglas P Wagner, and Jack E Zimmerman. Apache ii: a severity of disease classification system. *Critical care medicine*, 13(10):818–829, 1985.

[40] Wai Lun Law, Hok Kwok Choi, Yee Man Lee, and Judy WC Ho. The impact of postoperative complications on long-term outcomes following curative resection for colorectal cancer. *Annals of surgical oncology*, 14(9):2559–2566, 2007.

[41] Edward R Marcantonio, Lee Goldman, Carol M Mangione, Lynn E Ludwig, Brenda Muraca, Christine M Haslauer, Magruder C Donaldson, Anthony D Whittemore, David J Sugarbaker, Robert Poss, et al. A clinical prediction rule for delirium after elective noncardiac surgery. *Jama*, 271(2):134–139, 1994.

[42] Michal Nowakowski, Magdalena Pisarska, Mateusz Rubinkiewicz, Grzegorz Torbicz, Natalia Gajewska, Magdalena Mizera, Piotr Major, Pawel Potocki, Dorota Radkowiak, and Michal Pedziwiatr. Postoperative complications are associated with worse survival after laparoscopic surgery for non-metastatic colorectal cancer–interim analysis of 3-year overall survival. *Videosurgery and Other Miniinvasive Techniques*, 13(3):326, 2018.

[43] Ravi B Parikh, Christopher Manz, Corey Chivers, Susan Harkness Regli, Jennifer Braun, Michael E Draugelis, Lynn M Schuchter, Lawrence N Shulman, Amol S Navathe, Mitesh S Patel, et al. Machine learning approaches to predict 6-month mortality among patients with cancer. *JAMA network open*, 2(10):e1915997–e1915997, 2019.

[44] Chintan Parmar, Patrick Grossmann, Derek Rietveld, Michelle M Rietbergen, Philippe Lambin, and Hugo JWL Aerts. Radiomic machine-learning classifiers for prognostic biomarkers of head and neck cancer. *Frontiers in oncology*, 5:272, 2015.

[45] Karl Pearson. X. on the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 50(302):157–175, 1900.

[46] Karl Pearson. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[47] Karl Pearson. *Drapers' Company Research Memoirs: Biometric series*, volume 1. Cambridge University Press, 1904.

[48] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[49] DR Prytherch, MS Whiteley, B Higgins, PC Weaver, WG Prout, and SJ Powell. Possum and portsmouth possum for predicting mortality. *British Journal of Surgery*, 85(9):1217–1220, 1998.

[50] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[51] F Roques, SAM Nashef, P Michel, E Gauducheau, C De Vincentiis, E Baudet, J Cortina, M David, A Faichney, F Gavrielle, et al. Risk factors and outcome in european cardiac surgery: analysis of the euroscore multinational database of 19030 patients. *European Journal of Cardio-thoracic Surgery*, 15(6):816–823, 1999.

[52] Meyer Saklad. Grading of patients for surgical procedures. *Anesthesiology: The Journal of the American Society of Anesthesiologists*, 2(3):281–284, 1941.

[53] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal. Cancer statistics, 2019. *CA: A Cancer Journal for Clinicians*, 69 (1):7–34, 2019. doi: 10.3322/caac.21551. URL https://onlinelibrary.wiley.com/doi/abs/10.3322/caac.21551.

[54] Cristina Soguero-Ruiz, Kristian Hindberg, Inmaculada Mora-Jiménez, José Luis Rojo-Álvarez, Stein Olav Skrøvseth, Fred Godtliebsen, Kim Mortensen, Arthur Revhaug, Rolv-Ole Lindsetmo, Knut Magne Augestad, et al. Predicting colorectal surgical complications using heterogeneous clinical data and kernel methods. *Journal of biomedical informatics*, 61:87–96,

2016.

[55] Student. The probable error of a mean. *Biometrika*, pages 1–25, 1908.

[56] R Sutton, S Bann, M Brooks, and S Sarin. The surgical risk scale as an improved tool for risk-adjusted analysis in comparative surgical audit. *British Journal of Surgery*, 89(6):763–768, 2002.

[57] Luigi Tavazzi. Big data: is clinical practice changing? *European Heart Journal Supplements*, 21(Supplement-B):B98–B102, 03 2019. ISSN 1520-765X.

[58] Paul Thottakkara, Tezcan Ozrazgat-Baslanti, Bradley B Hupf, Parisa Rashidi, Panos Pardalos, Petar Momcilovic, and Azra Bihorac. Application of machine learning techniques to high-dimensional clinical data to forecast postoperative complications. *PloS one*, 11(5), 2016.

[59] Sachin Vaid, Ted Bell, Rod Grim, and Vanita Ahuja. Predicting risk of death in general surgery patients on the basis of preoperative variables using american college of surgeons national surgical quality improvement program data. *The Permanente journal*, 16(4):10, 2012.

[60] RGPM Van Stiphout, EO Postma, V Valentini, and P Lambin. The contribution of machine learning to predicting cancer outcome. *Artificial Intelligence*, 350:400, 2010.

[61] Guanjin Wang, Kin-Man Lam, Zhaohong Deng, and Kup-Sze Choi. Prediction of mortality after radical cystectomy for bladder cancer by machine learning techniques. *Computers in biology and medicine*, 63:124–132, 2015.

[62] MS Whiteley, DR Prytherch, B Higgins, PC Weaver, and WG Prout. An evaluation of the possum surgical scoring system. *British journal of surgery*, 83(6):812–815, 1996.

[63] Duminda N Wijeysundera. Predicting outcomes: Is there utility in risk scores? *Canadian Journal of Anesthesia/Journal canadien d'anesthésie*, 63(2):148–158, 2016.

[64] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.

[65] Maciej Zikeba, Jakub M Tomczak, Marek Lubicz, and Jerzy Swikatek. Boosted svm for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients. *Applied soft computing*, 14: 99–108, 2014.

A

Appendix - Selected Features

There are 3 sets of features used for each model optimization stage. One using the full set of available features, other using a set selected with a *p*-value = 0.1, and an even more restricted set of variables using a *p*-value = 0.0001. The selected sets are shown below, in ranking order:

A.1 Full Set of Variables

1. 'Days in ICU'
2. 'Postoperative Complication'
3. 'Clavien-Dindo Classification'
4. '1 Year Death'
5. 'Days at IPOP'
6. 'Death - Time After Surgery (1 Year Limit)'
7. 'Age'
8. 'Score Physiological P-Possum'
9. 'Score Surgical Gravity P-Possum'
10. '% Morbidity P-Possum'
11. '% Mortality P-Possum'
12. 'ACS Height'
13. 'ACS Weight'
14. 'Serious Complications (%)'
15. 'Serious Complications Average Risk'
16. 'Any Complication (%)'
17. 'Any Complication Average Risk'
18. 'Pneumonia (%)'
19. 'Pneumonia Average Risk'
20. 'Heart Complications (%)'
21. 'Heart Complications Average Risk'
22. 'Surgical Infection (%)'
23. 'Surgical Infection Average Risk'
24. 'UTI (%)'
25. 'UTI Average Risk'
26. 'Venous Thromboembolism (%)'
27. 'Venous Thromboembolism Average Risk'
28. 'Renal Failure (%)'
29. 'Renal Failure Average Risk'
30. 'Readmission (%)'
31. 'Readmission Average Risk'
32. 'Reoperation (%)'
33. 'Reoperation Average Risk'
34. 'Death (%)'
35. 'Death Average Risk'
36. 'Discharge to Nursing or Rehab Facility (%)'
37. 'Discharge to Nursing or Rehab Facility Average Risk'
38. 'ACS - Internment Days Prediction'

39. 'ARISCAT TOTAL SCORE'
40. 'SCORE ARISCAT'
41. 'Specialty'
42. 'LOCATION'
43. 'Preoperative Diagnostic'
44. 'Preoperative Comorbilities'
45. 'ACS_Procedure'
46. 'Surgery Type'
47. 'Specialty_COD'
48. 'ASA'
49. 'PP Age'
50. 'PP Cardiac'
51. 'PP Respiratory'
52. 'PP EKG'
53. 'PP Systolic Blood Pressure'
54. 'PP Pulse'
55. 'PP Hemoglobin'
56. 'PP Leukocytes'
57. 'PP Urea'
58. 'PP Sodium'
59. 'PP Potassium'
60. 'PP Glasglow Scale'
61. 'PP Surgery Type'
62. 'PP N. of Procedure'
63. 'PP Blood Loss'
64. 'PP Peritoneal Contamination'
65. 'PP Malignancy State'
66. 'PP CEPOD-Surgery Classification'
67. 'ACS Age'
68. 'ACS Functional State'
69. 'ACS ASA'
70. 'ACS Systemic Sepsis'
71. 'ACS Diabetes'
72. 'ACS Dyspnoea'
73. 'ARISCAT Age'
74. 'ARISCAT SpO2 '
75. 'ARISCAT Surgical Incision'
76. 'ARISCAT Surgery Duration'
77. 'Gender'
78. '1st Surgery at IPO'
79. 'Preoperative Chemo'
80. 'ACS Gender'
81. 'ACS Emergency'
82. 'ACS Steroids'
83. 'ACS Ascites'
84. 'ACS Ventilator Dependent'
85. 'ACS Disseminated cancer'
86. 'ACS Hypertension'
87. 'ACS CHF'
88. 'ACS Smoker'
89. 'ACS COPD'
90. 'ACS Dialysis'
91. 'ACS Acute Renal Failure'
92. 'ARISCAT Last Month Respiratory Infection'
93. 'ARISCAT Preoperative Anemia'
94. 'ARISCAT Emergent Procedure'
95. 'Date of Surgery'

A.2 First Feature Selection Stage (p-value = 0.1)

A.2.0.1 Feature ranking for the output "Days in the ICU":

1. 'Reoperation (%)'
2. 'Serious Complications (%)'
3. 'ACS - Internment Days Prediction'
4. 'Any Complication (%)'
5. 'ARISCAT TOTAL SCORE'
6. 'Serious Complications Average Risk'
7. 'Pneumonia (%)'
8. 'Any Complication Average Risk'
9. 'Heart Complications Average Risk'
10. 'Venous Thromboembolism (%)'
11. 'Reoperation Average Risk'
12. 'PP N. of Procedure'
13. 'ACS Ventilator Dependent'
14. 'SCORE ARISCAT'
15. '% Mortality P-Possum'
16. 'Pneumonia Average Risk'
17. 'ARISCAT Surgery Duration'
18. 'ACS Systemic Sepsis'
19. 'Score Surgical Gravity P-Possum'
20. 'Discharge to Nursing or Rehab Facility (%)'
21. 'Surgical Infection Average Risk'
22. 'Heart Complications (%)'
23. 'Death (%)'
24. 'PP Leukocytes'
25. 'Renal Failure (%)'
26. 'PP Peritoneal Contamination'
27. 'Discharge to Nursing or Rehab Facility Average Risk'
28. 'Venous Thromboembolism Average Risk'
29. 'Surgical Infection (%)'
30. 'PP Blood Loss'
31. '% Morbidity P-Possum'
32. 'Score Physiological P-Possum'
33. 'ARISCAT Emergent Procedure'
34. 'PP Urea'
35. 'PP CEPOD-Surgery Classification'
36. 'Readmission (%)'
37. 'Readmission Average Risk'
38. 'ARISCAT Preoperative Anemia'
39. 'Specialty_COD'
40. 'ACS Ascites'
41. 'Renal Failure Average Risk'
42. 'PP Hemoglobin'
43. 'ACS Dyspnoea'
44. 'ACS Emergency'
45. 'Surgery Type'
46. 'ACS Smoker'
47. 'ARISCAT Surgical Incision'
48. 'Death Average Risk'
49. 'Gender'
50. 'PP Respiratory'
51. 'ASA'
52. 'ACS Gender'
53. 'PP Pulse'
54. 'ACS ASA'
55. 'ACS COPD'
56. 'PP Systolic Blood Pressure'
57. 'PP Sodium'
58. 'ACS Functional State'
59. 'PP Potassium'
60. 'ACS Weight'

A.2.0.2 Feature ranking for the output "Existence of Postoperative Complications":

1. 'Serious Complications (%)'
2. 'Any Complication (%)'
3. 'ACS - Internment Days Prediction'
4. 'Pneumonia (%)'
5. 'Reoperation (%)'
6. 'Discharge to Nursing or Rehab Facility (%)'
7. '% Morbidity P-Possum'
8. 'Death (%)'
9. 'Venous Thromboembolism (%)'
10. 'Renal Failure (%)'
11. '% Mortality P-Possum'

12. 'Heart Complications (%)'
13. 'Readmission (%)'
14. 'Score Surgical Gravity P-Possum'
15. 'ARISCAT Emergent Procedure'
16. 'Score Physiological P-Possum'
17. 'Any Complication Average Risk'
18. 'Serious Complications Average Risk'
19. 'ARISCAT Preoperative Anemia'
20. 'Surgical Infection (%)'
21. 'UTI (%)'
22. 'Surgical Infection Average Risk'
23. 'Reoperation Average Risk'
24. 'Discharge to Nursing or Rehab Facility Average Risk'
25. 'Death Average Risk'
26. 'ARISCAT TOTAL SCORE'
27. 'Readmission Average Risk'
28. 'SCORE ARISCAT'
29. 'ARISCAT SpO2 '
30. 'Renal Failure Average Risk'
31. 'Venous Thromboembolism Average Risk'
32. 'UTI Average Risk'
33. 'Pneumonia Average Risk'
34. 'ARISCAT Last Month Respiratory Infection'
35. 'Heart Complications Average Risk'
36. 'PP Peritoneal Contamination'
37. 'ACS Functional State'
38. 'PP Respiratory'
39. 'PP Hemoglobin'
40. 'ACS Systemic Sepsis'
41. 'Age'

A.2.0.3 Feature ranking for the output "Complications Severity":

1. 'ARISCAT Emergent Procedure'
2. 'ARISCAT Preoperative Anemia'
3. 'Death (%)'
4. 'Pneumonia (%)'
5. 'Serious Complications (%)'
6. 'ACS - Internment Days Prediction'
7. 'Any Complication (%)'
8. 'ARISCAT SpO2 '
9. 'Discharge to Nursing or Rehab Facility (%)'
10. '% Mortality P-Possum'
11. 'ARISCAT Last Month Respiratory Infection'
12. 'Renal Failure (%)'
13. 'Heart Complications (%)'
14. 'Reoperation (%)'
15. '% Morbidity P-Possum'
16. 'Venous Thromboembolism (%)'
17. 'Score Surgical Gravity P-Possum'
18. 'ACS Functional State'
19. 'Score Physiological P-Possum'
20. 'ARISCAT Surgical Incision'
21. 'PP Peritoneal Contamination'
22. 'Readmission (%)'
23. 'ACS Systemic Sepsis'
24. 'Reoperation Average Risk'
25. 'UTI (%)'
26. 'ARISCAT TOTAL SCORE'
27. 'Death Average Risk'
28. 'Serious Complications Average Risk'
29. 'Any Complication Average Risk'
30. 'Discharge to Nursing or Rehab Facility Average Risk'
31. 'Surgical Infection (%)'
32. 'Surgical Infection Average Risk'
33. 'SCORE ARISCAT'
34. 'Renal Failure Average Risk'
35. 'Readmission Average Risk'
36. 'ACS Height'
37. 'Heart Complications Average Risk'
38. 'Pneumonia Average Risk'
39. 'Venous Thromboembolism Average Risk'
40. 'ACS Weight'
41. 'UTI Average Risk'

A.2.0.4 Feature ranking for the output "1 Year Death":

1. 'Pneumonia (%)'
2. 'Serious Complications (%)'
3. 'Any Complication (%)'
4. 'ACS - Internment Days Prediction'
5. 'Readmission (%)'
6. 'Discharge to Nursing or Rehab Facility (%)'
7. '% Morbidity P-Possum'
8. 'Reoperation (%)'
9. 'Death (%)'
10. '% Mortality P-Possum'
11. 'Venous Thromboembolism (%)'
12. 'Score Physiological P-Possum'
13. 'Score Surgical Gravity P-Possum'
14. 'Heart Complications (%)'
15. 'Renal Failure (%)'
16. 'UTI (%)'
17. 'ACS Weight'
18. 'ARISCAT Preoperative Anemia'
19. 'Reoperation Average Risk'
20. 'Age'
21. 'Discharge to Nursing or Rehab Facility Average Risk'
22. 'Any Complication Average Risk'
23. 'Serious Complications Average Risk'
24. 'ACS Functional State'
25. 'Surgical Infection (%)'
26. 'ARISCAT Emergent Procedure'
27. 'Death Average Risk'
28. 'ARISCAT Surgical Incision'
29. 'PP Hemoglobin'
30. 'Surgical Infection Average Risk'
31. 'Pneumonia Average Risk'
32. 'Readmission Average Risk'
33. 'ARISCAT SpO2 '

A.3 Second Feature Selection Stage (p-value = 0.0001)

A.3.0.1 Feature ranking for the output "Days in the ICU":

1. 'Reoperation (%)'
2. 'Serious Complications (%)'
3. 'ACS - Internment Days Prediction'
4. 'Any Complication (%)'
5. 'ARISCAT TOTAL SCORE'
6. 'Serious Complications Average Risk'
7. 'Pneumonia (%)'
8. 'Any Complication Average Risk'
9. 'Heart Complications Average Risk'
10. 'Venous Thromboembolism (%)'
11. 'Reoperation Average Risk'
12. 'PP N. of Procedure'

13. 'ACS Ventilator Dependent'
14. 'SCORE ARISCAT'
15. '% Mortality P-Possum'
16. 'Pneumonia Average Risk'
17. 'ARISCAT Surgery Duration'
18. 'ACS Systemic Sepsis'
19. 'Score Surgical Gravity P-Possum'
20. 'Discharge to Nursing or Rehab Facility (%)'
21. 'Surgical Infection Average Risk'
22. 'Heart Complications (%)'
23. 'Death (%)'
24. 'PP Leukocytes'
25. 'Renal Failure (%)'
26. 'PP Peritoneal Contamination'
27. 'Discharge to Nursing or Rehab Facility Average Risk'

A.3.0.2 Feature ranking for the output "Existence of Postoperative Complications":

1. 'Serious Complications (%)'
2. 'Any Complication (%)'
3. 'ACS - Internment Days Prediction'
4. 'Pneumonia (%)'
5. 'Reoperation (%)'
6. 'Discharge to Nursing or Rehab Facility (%)'
7. '% Morbidity P-Possum'
8. 'Death (%)'
9. 'Venous Thromboembolism (%)'
10. 'Renal Failure (%)'
11. '% Mortality P-Possum'
12. 'Heart Complications (%)'
13. 'Readmission (%)'
14. 'Score Surgical Gravity P-Possum'
15. 'ARISCAT Emergent Procedure'
16. 'Score Physiological P-Possum'
17. 'Any Complication Average Risk'
18. 'Serious Complications Average Risk'
19. 'ARISCAT Preoperative Anemia'
20. 'Surgical Infection (%)'
21. 'UTI (%)'
22. 'Surgical Infection Average Risk'
23. 'Reoperation Average Risk'
24. 'Discharge to Nursing or Rehab Facility Average Risk'
25. 'Death Average Risk'
26. 'ARISCAT TOTAL SCORE'
27. 'Readmission Average Risk'

A.3.0.3 Feature ranking for the output "Complications' Severity":

1. 'ARISCAT Emergent Procedure'
2. 'ARISCAT Preoperative Anemia'
3. 'Death (%)'
4. 'Pneumonia (%)'
5. 'Serious Complications (%)'
6. 'ACS - Internment Days Prediction'
7. 'Any Complication (%)'
8. 'Discharge to Nursing or Rehab Facility (%)'
9. '% Mortality P-Possum'
10. 'Renal Failure (%)'
11. 'Heart Complications (%)'
12. 'Reoperation (%)'
13. '% Morbidity P-Possum'
14. 'Venous Thromboembolism (%)'
15. 'Score Surgical Gravity P-Possum'
16. 'Score Physiological P-Possum'
17. 'Readmission (%)'
18. 'Reoperation Average Risk'
19. 'UTI (%)'
20. 'ARISCAT TOTAL SCORE'
21. 'Death Average Risk'
22. 'Serious Complications Average Risk'
23. 'Any Complication Average Risk'
24. 'Discharge to Nursing or Rehab Facility Average Risk'
25. 'Surgical Infection (%)'
26. 'Surgical Infection Average Risk'
27. 'SCORE ARISCAT'
28. 'Renal Failure Average Risk'

A.3.0.4 Feature ranking for the output "1 Year Death":

1. 'Pneumonia (%)'
2. 'Serious Complications (%)'
3. 'Any Complication (%)'
4. 'ACS - Internment Days Prediction'
5. 'Readmission (%)'
6. 'Discharge to Nursing or Rehab Facility (%)'
7. '% Morbidity P-Possum'
8. 'Reoperation (%)'
9. 'Death (%)'
10. '% Mortality P-Possum'
11. 'Venous Thromboembolism (%)'
12. 'Score Physiological P-Possum'
13. 'Score Surgical Gravity P-Possum'
14. 'Heart Complications (%)'
15. 'Renal Failure (%)'
16. 'UTI (%)'

B

Appendix - Models' Hyperparameters

The models used in this document were tested using 4 different sets of hyperparameters: the first was the default set (blank); the second was the result of hyperparameter optimization without feature selection; the third was the result of hyperparameter optimization after feature selection with a p-value - 0.1; the fourth is the result of hyperparameter optimization after feature selection with a p-value = 0.0001. The 3 optimized sets are shown in this appendix.

B.1 Optimized Set - No Feature Selection

B.1.1 Classification - Existence of Complications

- k-Nearest Neighbors:

```
{'n_neighbors': 10, 'weights': 'uniform'}
```

- Decision Trees:

```
{'ccp_alpha': 0.00033514029600854867, 'criterion': 'gini', 'max_depth': 84, 'min_samples_leaf': 16, '
    min_samples_split': 10, 'min_weight_fraction_leaf': 0.006997454163191403, 'splitter': 'random'}
```

- Random Forests:

```
{'ccp_alpha': 0.0011099469890931818, 'criterion': 'gini', 'min_samples_leaf': 14, 'min_samples_split':
    14, 'min_weight_fraction_leaf': 0.0003697013624808938, 'n_estimators': 50}
```

- Support Vector Machines:

{'C': 1.8685145743875649, 'decision_function_shape': 'ovo', 'degree': 3, 'gamma': 'scale', 'kernel': 'rbf', 'shrinking': False, 'tol': 0.045855638056383506}

- Logistic Regression:

{'C': 0.08035299230850745, 'fit_intercept': False, 'penalty': 'l2', 'solver': 'saga'}

- Extreme Gradient Boosting:

{'booster': 'gbtree', 'gamma': 0.020378724221450446, 'learning_rate': 0.06999558433910678, 'max_depth': 18, 'n_estimators': 270, 'reg_alpha': 0.00010039334340736686}

- Multilayer Perceptron:

{'activation': 'identity', 'alpha': 0.00023922065113711865, 'batch_size': 250, 'early_stopping': False, 'hidden_layer_sizes': (25), 'learning_rate': 'adaptive', 'learning_rate_init': 0.0016662140938665715, 'solver': 'sgd'}

B.1.2 Classification - Complications' Severity

- k-Nearest Neighbours:

{'n_neighbors': 48, 'weights': 'distance'}

- Decision Trees:

{'ccp_alpha': 0.0001321333448409692, 'criterion': 'gini', 'max_depth': 75, 'min_samples_leaf': 14, 'min_samples_split': 16, 'min_weight_fraction_leaf': 0.019094661328607775, 'splitter': 'best'}

- Random Forests:

{'ccp_alpha': 0.0009103207327372043, 'criterion': 'entropy', 'min_samples_leaf': 16, 'min_samples_split': 6, 'min_weight_fraction_leaf': 0.014311533908442989, 'n_estimators': 150}

- Support Vector Machines:

{'C': 0.9189234151553689, 'decision_function_shape': 'ovo', 'degree': 3, 'gamma': 'auto', 'kernel': 'rbf', 'shrinking': False, 'tol': 0.003970023949811552}

- Logistic Regression:

{'C': 0.2246304891861346, 'fit_intercept': False, 'penalty': 'l2', 'solver': 'lbfgs'}

- Extreme Gradient Boosting:

{'booster': 'dart', 'gamma': 0.012925475325721735, 'learning_rate': 0.005812754377284124, 'max_depth': 3, 'n_estimators': 420, 'reg_alpha': 0.0001406038579548865}

- Multilayer Perceptron:

{'activation': 'tanh', 'alpha': 0.0163020172962808, 'batch_size': 300, 'early_stopping': False, 'hidden_layer_sizes': (50,100,50,), 'learning_rate': 'invscaling', 'learning_rate_init': 0.009857518651835451, 'solver': 'adam'}

B.1.3 Classification - 1 Year Death

- k-Nearest Neighbours:

{'n_neighbors': 4, 'weights': 'uniform'}

- Decision Trees:

{'ccp_alpha': 0.0008164441131025791, 'criterion': 'entropy', 'max_depth': 51, 'min_samples_leaf': 2, 'min_samples_split': 18, 'min_weight_fraction_leaf': 0.4390857558643149, 'splitter': 'best'}

- Random Forests:

{'ccp_alpha': 0.0015350846450790524, 'criterion': 'entropy', 'min_samples_leaf': 6, 'min_samples_split': 16, 'min_weight_fraction_leaf': 0.0025524643060734096, 'n_estimators': 200}

- Support Vector Machines:

{'C': 1.6136204472012943, 'decision_function_shape': 'ovo', 'degree': 5, 'gamma': 'scale', 'kernel': 'rbf', 'shrinking': False, 'tol': 0.0016385662109461184}

- Logistic Regression:

{'C': 0.13462010975587382, 'fit_intercept': True, 'penalty': 'l2', 'solver': 'lbfgs'}

- Extreme Gradient Boosting:

{'booster': 'dart', 'gamma': 0.007534385901076864, 'learning_rate': 0.5240628318691967, 'max_depth': 3, 'n_estimators': 120, 'reg_alpha': 0.03506282747303021}

- Multilayer Perceptron:

{'activation': 'identity', 'alpha': 0.12284268593974768, 'batch_size': 100, 'early_stopping': False, 'hidden_layer_sizes': (50,), 'learning_rate': 'adaptive', 'learning_rate_init': 0.044626743912542836, 'solver': 'adam'}

B.1.4 Regression - Complications' Severity

- Linear Regression:

{'fit_intercept': True, 'normalize': False}

- Ridge Regression:

{'alpha': 1.433616782546514, 'fit_intercept': True, 'normalize': False, 'solver': 'auto'}

- LASSO Regression:

{'alpha': 6.097527112471915, 'fit_intercept': False, 'normalize': False, 'precompute': False}

- Support Vector Regression:

{'C': 1.2947486555069185, 'degree': 5, 'gamma': 'scale', 'kernel': 'rbf', 'shrinking': False}

- Elastic Regression:

{'alpha': 1.2916130552465075, 'fit_intercept': False, 'l1_ratio': 0.001801666698174747, 'normalize': False}

- K-Nearest Neighbours Regressor:

{'algorithm': 'kd_tree', 'n_neighbors': 22, 'weights': 'uniform'}

- Decision Tree Regressor:

{'ccp_alpha': 0.004342561405542502, 'criterion': 'mae', 'max_depth': 65, 'min_samples_leaf': 14, 'min_samples_split': 14, 'min_weight_fraction_leaf': 0.005684146290952337, 'splitter': 'random'}

- Random Forest Regressor:

{'ccp_alpha': 0.017580241052617063, 'criterion': 'mae', 'min_samples_leaf': 4, 'min_samples_split': 10, 'min_weight_fraction_leaf': 0.004809827759508562, 'n_estimators': 80}

- XGBoost Regressor:

{'booster': 'gbtree', 'gamma': 0.001615896871147276, 'learning_rate': 0.004144048200160541, 'max_depth': 9, 'n_estimators': 320, 'reg_alpha': 0.05908488686823017}

- Partial Least Squares Regression:

{'n_components': 1}

- Multilayer Perceptron Regressor:

{'activation': 'logistic', 'alpha': 0.00010781136106364605, 'batch_size': 250, 'early_stopping': True, 'hidden_layer_sizes': (50,), 'learning_rate': 'adaptive', 'learning_rate_init': 0.0029368234170485115, 'solver': 'adam'}

B.1.5 Regression - Days in the ICU

- Linear Regression:

{'fit_intercept': False, 'normalize': True}

- Ridge Regression:

{'alpha': 1.0084999130619332, 'fit_intercept': True, 'normalize': True, 'solver': 'auto'}

- LASSO Regression:

{'alpha': 2.30527979234979, 'fit_intercept': True, 'normalize': False, 'precompute': True}

- Support Vector Regression:

{'C': 1.2015431074644953, 'degree': 4, 'gamma': 'scale', 'kernel': 'rbf', 'shrinking': False}

- Elastic Regression:

{'alpha': 1.0036833450105551, 'fit_intercept': False, 'l1_ratio': 0.020572520729076584, 'normalize': True}

- K-Nearest Neighbours Regressor:

{'algorithm': 'ball_tree', 'n_neighbors': 46, 'weights': 'uniform'}

- Decision Tree Regressor:

{'ccp_alpha': 0.0013643422221902815, 'criterion': 'mae', 'max_depth': 95, 'min_samples_leaf': 14, 'min_samples_split': 16, 'min_weight_fraction_leaf': 0.0033686540008876905, 'splitter': 'random'}

- Random Forest Regressor:

{'ccp_alpha': 0.005275566397171974, 'criterion': 'mae', 'min_samples_leaf': 8, 'min_samples_split': 20, 'min_weight_fraction_leaf': 0.005085724373320969, 'n_estimators': 120}

- XGBoost Regressor:

{'booster': 'gbtree', 'gamma': 0.0002491578788307851, 'learning_rate': 0.002155922123347837, 'max_depth': 3, 'n_estimators': 500, 'reg_alpha': 0.0006542724632320029}

- Partial Least Squares Regression:

{'n_components': 2}

- Multilayer Perceptron Regressor:

{'activation': 'logistic', 'alpha': 0.0011216587184417774, 'batch_size': 50, 'early_stopping': True, 'hidden_layer_sizes': (100,100,), 'learning_rate': 'invscaling', 'learning_rate_init': 0.0011787100570382375, 'solver': 'adam'}

B.2 Optimized Set - p-value = 0.0001 Feature Selection

B.2.1 Classification - Existence of Complications

- k-Nearest Neighbours:

{'n_neighbors': 34, 'weights': 'distance'}

- Decision Trees:

{'ccp_alpha': 0.00429295112201027, 'criterion': 'gini', 'max_depth': 93, 'min_samples_leaf': 6, 'min_samples_split': 14, 'min_weight_fraction_leaf': 0.13708520673251146, 'splitter': 'best'}

- Random Forests:

{'ccp_alpha': 0.0384258250479654, 'criterion': 'gini', 'min_samples_leaf': 16, 'min_samples_split': 10, 'min_weight_fraction_leaf': 0.017596034317898027, 'n_estimators': 20}

- Support Vector Machines:

{'C': 1.9996513568560503, 'decision_function_shape': 'ovo', 'degree': 5, 'gamma': 'scale', 'kernel': 'linear', 'shrinking': False, 'tol': 0.009148823602882007}

- Logistic Regression:

{'C': 0.4534006074557504, 'fit_intercept': True, 'dual': False, 'penalty': 'l2', 'solver': 'liblinear'}

- Extreme Gradient Boosting:

{'booster': 'dart', 'gamma': 0.0003173319796021907, 'learning_rate': 0.005636323967519565, 'max_depth': 6, 'n_estimators': 350, 'reg_alpha': 0.00015182046374735654}

- Multilayer Perceptron:

{'activation': 'identity', 'alpha': 0.0031199417306985162, 'batch_size': 250, 'early_stopping': True, 'hidden_layer_sizes': (25), 'learning_rate': 'adaptive', 'learning_rate_init': 0.07281583397017921, 'solver': 'sgd'}

B.2.2 Classification - Complications' Severity

- k-Nearest Neighbours:

```
{'n_neighbors': 12, 'weights': 'uniform'}
```

- Decision Trees:

```
{'ccp_alpha': 0.0012499567107979162, 'criterion': 'gini', 'max_depth': 81, 'min_samples_leaf': 8, '
    min_samples_split': 20, 'min_weight_fraction_leaf': 0.22882264730753812, 'splitter': 'random'}
```

- Random Forests:

```
{'ccp_alpha': 0.01639097706223533, 'criterion': 'gini', 'min_samples_leaf': 20, 'min_samples_split': 10,
     'min_weight_fraction_leaf': 0.0001675222584292594, 'n_estimators': 200}
```

- Support Vector Machines:

```
{'C': 0.40377293036689427, 'decision_function_shape': 'ovr', 'degree': 4, 'gamma': 'scale', 'kernel': '
    rbf', 'shrinking': True, 'tol': 0.07709880964539434}
```

- Logistic Regression:

```
{'C': 1.9197684226946312, 'fit_intercept': True, 'penalty': 'none', 'solver': 'sag'}
```

- Extreme Gradient Boosting:

```
{'booster': 'gbtree', 'gamma': 1.2300321508261962, 'learning_rate': 0.042396341810979996, 'max_depth':
    17, 'n_estimators': 170, 'reg_alpha': 0.044095288835510976}
```

- Multilayer Perceptron:

```
{'activation': 'logistic', 'alpha': 0.0520090314379177, 'batch_size': 200, 'early_stopping': True, '
    hidden_layer_sizes': (25), 'learning_rate': 'invscaling', 'learning_rate_init': 0.00127923087363695
    , 'solver': 'lbfgs'}
```

B.2.3 Classification - 1 Year Death

- k-Nearest Neighbours:

```
{'n_neighbors': 4, 'weights': 'uniform'}
```

- Decision Trees:

```
{'ccp_alpha': 0.0002621591045514422, 'criterion': 'gini', 'max_depth': 72, 'min_samples_leaf': 12, '
    min_samples_split': 12, 'min_weight_fraction_leaf': 0.03225070598332315, 'splitter': 'best'}
```

- Random Forests:

```
{'ccp_alpha': 0.0006476706731739106, 'criterion': 'gini', 'min_samples_leaf': 10, 'min_samples_split':
    2, 'min_weight_fraction_leaf': 0.0003392453212472795, 'n_estimators': 120}
```

- Support Vector Machines:

```
{'C': 1.9524429654131306, 'decision_function_shape': 'ovr', 'degree': 5, 'gamma': 'auto', 'kernel': '
    linear', 'shrinking': False, 'tol': 0.08488271936872323}
```

- Logistic Regression:

```
{'C': 1.5149546680431834, 'fit_intercept': True, 'penalty': 'none', 'solver': 'sag'}
```

- Extreme Gradient Boosting:

```
{'booster': 'gbtree', 'gamma': 0.01217811556240899, 'learning_rate': 0.7944278910400787, 'max_depth':
    36, 'n_estimators': 210, 'reg_alpha': 0.004646922554067186}
```

- Multilayer Perceptron:

```
{'activation': 'tanh', 'alpha': 0.0019528798409905903, 'batch_size': 100, 'early_stopping': False, '
    hidden_layer_sizes': (25), 'learning_rate': 'constant', 'learning_rate_init': 0.0014018047686829552
    , 'solver': 'adam'}
```

B.2.4 Regression - Complications' Severity

- Linear Regression:

```
{'fit_intercept': True, 'normalize': True}
```

- Ridge Regression:

```
{'alpha': 1.4356765645820724, 'fit_intercept': True, 'normalize': False, 'solver': 'auto'}
```

- LASSO Regression:

```
{'alpha': 9.663052603303248, 'fit_intercept': False, 'normalize': True, 'precompute': True}
```

- Support Vector Regression:

```
{'C': 1.3029077878653237, 'degree': 5, 'gamma': 'scale', 'kernel': 'rbf', 'shrinking': True}
```

- Elastic Regression:

```
{'alpha': 1.1132276444070772, 'fit_intercept': False, 'l1_ratio': 0.009312491615277252, 'normalize':
    False}
```

- K-Nearest Neighbours Regressor:

```
{'algorithm': 'brute', 'n_neighbors': 22, 'weights': 'uniform'}
```

- Decision Tree Regressor:

```
{'ccp_alpha': 0.0010966158782928297, 'criterion': 'mae', 'max_depth': 70, 'min_samples_leaf': 12, '
    min_samples_split': 8, 'min_weight_fraction_leaf': 0.06502645425165987, 'splitter': 'random'}
```

- Random Forest Regressor:

```
{'ccp_alpha': 0.023415474996602674, 'criterion': 'mae', 'min_samples_leaf': 6, 'min_samples_split': 2, '
    min_weight_fraction_leaf': 0.004088545306169188, 'n_estimators': 100}
```

- XGBoost Regressor:

```
{'booster': 'dart', 'gamma': 0.003642914237204647, 'learning_rate': 0.0035603832947454345, 'max_depth':
    3, 'n_estimators': 460, 'reg_alpha': 0.014395166907462276}
```

- Partial Least Squares Regression:

```
{'n_components': 3}
```

- Multilayer Perceptron Regressor:

```
{'activation': 'logistic', 'alpha': 0.016879916221349926, 'batch_size': 300, 'early_stopping': False, '
    hidden_layer_sizes': (50,), 'learning_rate': 'adaptive', 'learning_rate_init': 0.01749690261049504,
     'solver': 'adam'}
```

B.2.5 Regression - Days in the ICU

- Linear Regression:

```
{'fit_intercept': False, 'normalize': True}
```

- Ridge Regression:

```
{'alpha': 1.0161020029593626, 'fit_intercept': True, 'normalize': True, 'solver': 'auto'}
```

- LASSO Regression:

```
{'alpha': 3.65586591657791, 'fit_intercept': True, 'normalize': True, 'precompute': False}
```

- Support Vector Regression:

```
{'C': 1.201966659062525, 'degree': 5, 'gamma': 'scale', 'kernel': 'rbf', 'shrinking': True}
```

- Elastic Regression:

```
{'alpha': 1.0244126085161624, 'fit_intercept': False, 'l1_ratio': 0.0009062756077093115, 'normalize':
    True}
```

- K-Nearest Neighbours Regressor:

```
{'algorithm': 'brute', 'n_neighbors': 46, 'weights': 'uniform'}
```

- Decision Tree Regressor:

```
{'ccp_alpha': 0.021270317984444176, 'criterion': 'mae', 'max_depth': 10, 'min_samples_leaf': 12, '
    min_samples_split': 6, 'min_weight_fraction_leaf': 0.0023657322805192217, 'splitter': 'best'}
```

- Random Forest Regressor:

{'ccp_alpha': 0.0037892053379269632, 'criterion': 'mae', 'min_samples_leaf': 8, 'min_samples_split': 6, 'min_weight_fraction_leaf': 0.005681808128580106, 'n_estimators': 180}

- XGBoost Regressor:

{'booster': 'gbtree', 'gamma': 0.00179866468766607, 'learning_rate': 0.0033309758058766986, 'max_depth': 3, 'n_estimators': 460, 'reg_alpha': 0.0033595214600199196}

- Partial Least Squares Regression:

{'n_components': 1}

- Multilayer Perceptron Regressor:

{'activation': 'tanh', 'alpha': 0.003850255057655146, 'batch_size': 150, 'early_stopping': False, 'hidden_layer_sizes': (50,100,50,), 'learning_rate': 'invscaling', 'learning_rate_init': 0.011371928356166733, 'solver': 'sgd'}

B.3 Optimized Set - p-value = 0.1 Feature Selection

B.3.1 Classification - Existence of Complications

- k-Nearest Neighbours:

{'n_neighbors': 18, 'weights': 'uniform'}

- Decision Trees:

{'ccp_alpha': 0.003158836284337476, 'criterion': 'entropy', 'max_depth': 12, 'min_samples_leaf': 18, 'min_samples_split': 8, 'min_weight_fraction_leaf': 0.04150970034587553, 'splitter': 1}

- Random Forests:

{'ccp_alpha': 0.007376371845815248, 'criterion': 'entropy', 'min_samples_leaf': 12, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.011213022432220807, 'n_estimators': 20}

- Support Vector Machines:

{'C': 1.4315349110765008, 'decision_function_shape': 'ovo', 'degree': 5, 'gamma': 'auto', 'kernel': 'linear', 'shrinking': True, 'tol': 0.041651534375602026}

- Logistic Regression:

{'C': 1.904889278629025, 'fit_intercept': True, 'l1_ratio': 0.28207145242820697, 'penalty': 'elasticnet', 'solver': 'saga'}

- Extreme Gradient Boosting:

{'booster': 'dart', 'gamma': 0.028248409091746977, 'learning_rate': 0.8201332490632605, 'max_depth': 36, 'n_estimators': 110, 'reg_alpha': 0.0003410258189114528}

- Multilayer Perceptron:

{'activation': 'logistic', 'alpha': 0.00012855766180589005, 'batch_size': 50, 'early_stopping': True, 'hidden_layer_sizes': (50,), 'learning_rate': 'invscaling', 'learning_rate_init': 0.09624815038246906, 'solver': 'lbfgs'}

B.3.2 Classification - Complications' Severity

- k-Nearest Neighbours:

{'n_neighbors': 12, 'weights': 'uniform'}

- Decision Trees:

{'ccp_alpha': 0.00021848070587403436, 'criterion': 0, 'max_depth': 60, 'min_samples_leaf': 6, 'min_samples_split': 20, 'min_weight_fraction_leaf': 0.1360284580094963, 'splitter': 0}

- Random Forests:

{'ccp_alpha': 0.021741436475361905, 'criterion': 0, 'min_samples_leaf': 2, 'min_samples_split': 2, 'min_weight_fraction_leaf': 0.0004011837899573551, 'n_estimators': 30}

- Support Vector Machines:

{'C': 1.137043709865312, 'decision_function_shape': 'ovr', 'degree': 5, 'gamma': 'scale', 'kernel': 'linear', 'shrinking': True, 'tol': 0.05224048951134231}

- Logistic Regression:

{'C': 0.487199883552072, 'fit_intercept': True, 'penalty': 'l1', 'solver': 'liblinear'}

{'booster': 'dart', 'gamma': 0.7078791378965007, 'learning_rate': 0.006003103216192764, 'max_depth': 1, 'n_estimators': 130, 'reg_alpha': 0.055628048317247485}

- Extreme Gradient Boosting:

{'activation': 'tanh', 'alpha': 0.00010257336852803629, 'batch_size': 250, 'early_stopping': False, 'hidden_layer_sizes': (25), 'learning_rate': 'adaptive', 'learning_rate_init': 0.0010656109130631173, 'solver': 'adam'}

B.3.3 Classification - 1 Year Death

- k-Nearest Neighbours:

{'n_neighbors': 10, 'weights': 'uniform'}

- Decision Trees:

{'ccp_alpha': 0.0005021528951313316, 'criterion': 'entropy', 'max_depth': 36, 'min_samples_leaf': 10, 'min_samples_split': 12, 'min_weight_fraction_leaf': 0.014198089748714372, 'splitter': 1}

- Random Forests:

{'ccp_alpha': 0.00016431154701314744, 'criterion': 'entropy', 'min_samples_leaf': 12, 'min_samples_split': 20, 'min_weight_fraction_leaf': 0.0028766141876542515, 'n_estimators': 70}

- Support Vector Machines:

{'C': 1.9657376172241314, 'decision_function_shape': 'ovr', 'degree': 4, 'gamma': 'auto', 'kernel': 'rbf', 'shrinking': False, 'tol': 0.06761584751747673}

- Logistic Regression:

{'C': 1.6525410121670268, 'fit_intercept': False, 'penalty': 'l1', 'solver': 'liblinear'}

- Extreme Gradient Boosting:

{'booster': 2, 'gamma': 0.0022120526934845323, 'learning_rate': 0.5753762481785569, 'max_depth': 27, 'n_estimators': 150, 'reg_alpha': 0.002825934300702562}

- Multilayer Perceptron:

{'activation': 'logistic', 'alpha': 0.01584678952495818, 'batch_size': 150, 'early_stopping': False, 'hidden_layer_sizes': (25), 'learning_rate': 'adaptive', 'learning_rate_init': 0.020905841532323482, 'solver': 'sgd'}

B.3.4 Regression - Complications' Severity

- Linear Regression:

{'fit_intercept': False, 'normalize': False}

- Ridge Regression:

{'alpha': 9.999477475455103, 'fit_intercept': False, 'normalize': False, 'solver': 'auto'}

- LASSO Regression:

{'alpha': 2.807369031058963, 'fit_intercept': False, 'normalize': False, 'precompute': True}

- Support Vector Regression:

{'C': 0.7803017216533735, 'degree': 4, 'gamma': 'scale', 'kernel': 'rbf', 'shrinking': True}

- Elastic Regression:

{'alpha': 1.0867523402170132, 'fit_intercept': False, 'l1_ratio': 0.0002741011631456902, 'normalize': True}

- K-Nearest Neighbours Regressor:

{'algorithm': 'kd_tree', 'n_neighbors': 28, 'weights': 'distance'}

- Decision Tree Regressor:

{'ccp_alpha': 0.0005739913845507395, 'criterion': 2, 'max_depth': 75, 'min_samples_leaf': 14, '
    min_samples_split': 12, 'min_weight_fraction_leaf': 0.0017470814564770842, 'splitter': 1}

- Random Forest Regressor:

{'ccp_alpha': 0.020094511321263102, 'criterion': 'mae', 'min_samples_leaf': 8, 'min_samples_split': 18,
    'min_weight_fraction_leaf': 0.0002541535243351869, 'n_estimators': 20}

- XGBoost Regressor:

{'booster': 2, 'gamma': 0.0010137138524742317, 'learning_rate': 0.005574834742044873, 'max_depth': 21, '
    n_estimators': 240, 'reg_alpha': 0.000271615786949664}

- Partial Least Squares Regression:

{'n_components': 3}

- Multilayer Perceptron Regressor:

{'activation': 'tanh', 'alpha': 0.02137869340621498, 'batch_size': 200, 'early_stopping': True, '
    hidden_layer_sizes': (50,), 'learning_rate': 'invscaling', 'learning_rate_init': 0
    .004678722534514956, 'solver': 'adam'}

B.3.5 Regression - Days in the ICU

- Linear Regression:

{'fit_intercept': False, 'normalize': False}

- Ridge Regression:

{'alpha': 2.026731377532221, 'fit_intercept': True, 'normalize': True, 'solver': 'auto'}

- LASSO Regression:

{'alpha': 4.575918499164157, 'fit_intercept': True, 'normalize': True, 'precompute': False}

- Support Vector Regression:

{'C': 0.606522040866752, 'degree': 3, 'gamma': 'scale', 'kernel': 'rbf', 'shrinking': True}

- Elastic Regression:

{'alpha': 2.331178395333172, 'fit_intercept': False, 'l1_ratio': 0.0010250858372409643, 'normalize':
    False}

- K-Nearest Neighbours Regressor:

{'algorithm': 'ball_tree', 'n_neighbors': 38, 'weights': 'uniform'}

- Decision Tree Regressor:

{'ccp_alpha': 0.0024348321503631673, 'criterion': 2, 'max_depth': 10, 'min_samples_leaf': 12, '
    min_samples_split': 18, 'min_weight_fraction_leaf': 0.0002395319616883768, 'splitter': 1}

- Random Forest Regressor:

{'ccp_alpha': 0.006959011651425836, 'criterion': 'mae', 'min_samples_leaf': 20, 'min_samples_split': 18,
     'min_weight_fraction_leaf': 0.02156897547425383, 'n_estimators': 200}

- XGBoost Regressor:

{'booster': 2, 'gamma': 0.1447112553018635, 'learning_rate': 0.0024665053205051852, 'max_depth': 3, '
    n_estimators': 420, 'reg_alpha': 0.019900067616538546}

- Partial Least Squares Regression:

{'n_components': 1}

- Multilayer Perceptron Regressor:

{'activation': 'relu', 'alpha': 0.12379681206829639, 'batch_size': 200, 'early_stopping': True, '
    hidden_layer_sizes': (50,50,50,), 'learning_rate': 'adaptive', 'learning_rate_init': 0
    .0062462078547118535, 'solver': 'sgd'}