Clinical Prognosis and Risk Prediction of Postoperative Complications in Cancer Patients

Daniel Gonçalves
Instituto Superior Técnico
University of Lisbon, Lisbon, Portugal
dmateusgoncalves@tecnico.ulisboa.pt

Rafael Costa
Instituto Superior Técnico
University of Lisbon, Lisbon, Portugal
rafael.s.costa@tecnico.ulisboa.pt

Rui Henriques
Instituto Superior Técnico
University of Lisbon, Lisbon, Portugal
rmch@tecnico.ulisboa.pt

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.

This thesis aims to study and predict postoperative complications risk for cancer patients, offering two major contributions. First, to develop a risk calculator, specif for the Portuguese population, using machine learning models, with 4 outcomes of interest: i) existence of postoperative complications, ii) severity level of said complications, iii) death probability within 1 year, and iv) number of days spent in the intermediate care unit (ICU). 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.

In order to achieve these objectives, 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 a 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

1 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 [1, 7, 23, 24]. The outcome of such surgeries is still widely unpredictable due to the huge number of factors involved. Postoperative risk assessment tools are available, not only for cancer patients but for surgery in general, with the aim of reducing mortality and morbidity rates [33].

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 [33]. The predictions given by these tools 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 [33].

The main objective of this project is to develop a risk score that is able to predict 4 outcomes: i) existence of postoperative complications, ii) the severity of said complications, iii) the number of days spent in the ICU and the iv) 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. For this purpose, a clinical dataset with more than 800 patients and 100 attributes is available.

This document presents the related work review, the proposed methodology for the project, a summarized results discussion, finishing with some concluding remarks, along with limitations and future work.

2 Related Work

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.

2.1 Traditional Statistical Studies

Being able to predict postoperative complications is of crucial importance to assess treatment viability or chances of

survival. Creating opportunity for the consideration of alternative therapies or procedures, adequate intensive care, or even assisted life ending options.

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 difficult interpretability.

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 POSSUM score was created from a general surgery cohort [12]. 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 CARE score was developed in a cardiac surgery cohort [16]. 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 generalization capacity testing of the results would be advised. 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 of people from different medical cohorts and multiple hospitals, like the ACS NSQIP, which makes use of data collected from 393 American hospitals, totaling almost 1,500,000 patients [4]. 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. 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 [19]. 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. 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 [4], which accounts for demographic data, collected from over 393 hospitals all across America, having a solid and proven national applicability.

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 ([9]) or the Surgical Apgar Score ([19]), 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 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.

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. 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 linear approach is estimated through a distance minimizing approximation method, called Ordinary Least Squares, while logistic regression uses Maximum Likelyhood Estimation, a function that determines the parameters that are most likely to produce the observed data [5].

2.2 Machine Learning Studies

More recently, machine learning has stepped into the field, and the studies using this type of models, specifically for the prediction of postoperative complications, have also been increasing. In a primary analysis these studies bring new prediction models to the table, with high dimensional modeling capabilities, each having its own advantages.

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 in review corresponds to the year 2001, while ML studies correspond to the year 2015. In these fourteen years technology has evolved, and now, more than ever, the available hardware allows for feasible application of very complex methods. Big clinical data is also a growing phenomenon. ML models are making use of genomics, biological, physiological, radiomics, demographic and socio-economic data. Another characteristic differentiating ML and traditional statistic studies is that ML approaches seem less connected

to the professional medical setting, partly because the development of such algorithms is held by artificial intelligence researchers at an experimental level.

k-Nearest Neighbors - The kNN algorithm is one of the most intuitive and simple methods available, due to its distance based approach. In the studies reviewed it is used only once by Wang et al. [32]. 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.

Naive Bayes - Naive Bayes models are also suggested in situations where lightweight and simplistic solutions are enough to respond to the challenge. This method is applied in assuming that all the attributes are conditionally independent. According to Danjuma [14], this method is capable of improved prognostic compared with logistic regression. In Parmar et al. [26], in spite of the fact that it wasn't the best, the results were competitive with that of SVM, NN and RF.

Decision Trees - A 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. Danjuma [14] used a DT to predict mortality within 1 year. The results were good, only surpassed by the MLP, a particular type of artificial neural network.

Support Vector Machines - SVMs are another ML model which is frequently used among clinical predictors. SVMs are not as understandable and explicable as other methods like DTs or kNN [3]. Chang et al. [8] 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. [28] 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. [30] 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. [32] 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. Allied with various feature selection methods Parmar et al. [26] tried to predict 3-year mortality on a small dataset of 101 patients, with high dimensionality, containing 404 features. Chang et al. [8] 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. The other network was a fuzzy classifier, a paradigm contrasting with crisp classification. The overall best method was the latter and the overall worst was the normal NN. Danjuma [14] 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.

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, [6]. Zikeba et al. [34] proposed a boosted SVM model to solve inner and between-class imbalanced data problems, by proposing weighted error function with different misclassification costs, for positive and negative examples respectively. The boosting algorithm used is AdaBoost. The results revealed good performance from the ensemble method, and proved the ability to overcome imbalance induced bias.

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. [26] used a RF model among their models set. This model has a competitive performance, but above everything else it proved to be much more stable across tests. Parikh et al. [25] used a Random Forest model (RF) and also Gradient Boosting (GB), both tree based ensemble models. Both models showed good results with a positive predictive value superior to that of traditional statistical values.

2.3 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.

Missing Values - Missing Values are the result of unavailable data at the time of registry and can sometimes be a product of human error. Some predictive models cannot handle missing values, so 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. However, 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 Thottakkara et al. [30] and Van Stiphout et al. [31]. 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. [4] (using a regression method).

Outcome Class Imbalance - Class imbalance is a common problem in medical decision problems [34]. 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. The reduction is the simplest method, but information is precious, and these studies are not making use of very extensive datasets to start with. 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. 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. [34] used various ensemble methods based on SVMs which proved to be efficient at dealing with data imbalance.

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" [2]. 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 Chang et al. [8], Parmar et al. [26] or Parikh et al. [25]. 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 [27], 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. [30].

2.4 Validation

Problems related to poor world wide applicability have been reported in studies [10, 17, 18, 20]. The common conclusions seem to point out that further validation with foreign datasets would be crucial to obtain better reusability. Out of all the studies, 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 [8, 14, 28, 30, 32].

3 Methodology

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.

The dataset has 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.

3.1 The Dataset

A retrospective dataset was provided by courtesy of the Portuguese Institute of Oncology, Porto, Portugal (IPO-Porto). 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 a cancer related surgeries, and is already anonymized. 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.

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. 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.

3.2 Preprocessing

Missing Values - 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. Two types of imputation processes were considered, one using substitution by mean value of the variable and another that is more complex but might offer better results. 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.

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.

Resampling - In clinical data, an imbalance between the positive and the negative class is common, 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. In this project, a mixed strategy is used combining synthetic oversampling with k-Nearest Neighbors informed undersampling, as proposed by He and Garcia [21].

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,our methodology proposes to normalize or standardize data.

Feature Selection - Not all variables might be relevant for a certain prediction, therefore it's common to select a restricted number of variables that will actually be used to build the prediction models. Filter methods offer a p-value representing the probability that a variable is not correlated to an outcome. The Chi-Squared test is used to measure correlation for categorical variables, when the output is also categorical. The 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. Embedded methods are mechanisms intrinsic to the models. Our methods also explore this technique, especially using associative models.

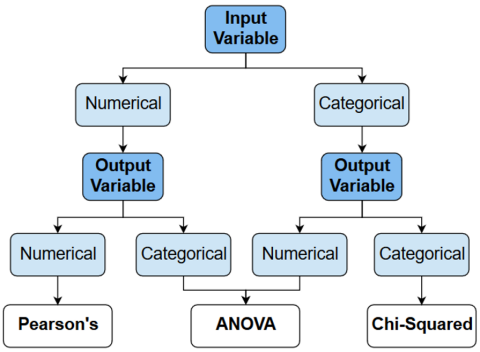

Figure 1. Feature selection process

3.3 Predictive Models

There are plenty of different models and respective variations, and 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: Naive Bayes, K-Nearest Neighbours, Decision Trees, Random Forests, Support Vector Machines, Logistic Regression, Multilayer Perceptron, XGBoost Classifier;
- Regression algorithms: Linear Regression, Ridge Regression, Lasso Regression, SVM Regressor, Elastic Regression, k-Nearest Neighbours Regressor, Decision Tree Regressor, Random Forest Regressor, XGBoost Regressor, Partial Least Squares Regression, Multilayer Perceptron Regressor.

3.4 Prediction Outcomes

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.

The Clavien-Dindo Classification [15] is a scale used to standardize in 8 grades the type of therapy needed after a

certain surgery, and is used as the second output of interest: the **severity of complications**. 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.

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. This problem could be treated as a classification problem with the objective of deciding if the patient was going to die within 1 year or not. Some 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 value between 0 and 1. Typically this would be solved by a regression model. In this case, a regression approach would probably not be fit for the task, since the outcome values in the dataset are binary.

The **number of days spent in the ICU** represents important information for medical and also hospital management reasons. Since the dependent variable is continuous (time), this prediction is better solved by a regression model.

3.5 Model Tuning: Hyperparameter Optimization

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. In this project, informed search models are employed. Bayesian optimization [22] associates a probability distribution to the hyperparameters tested, making the search faster than exhaustive approaches. 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);
- Classification models are optimized to maximize their recall (the sensitivity calculated for each target class, and then averaged in a non-weighted formula).

3.6 Evaluation Metrics

Classification Evaluation Metrics - 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. Accuracy can be misleading in situations where the data is imbalanced. In order to overcome the weaknesses of the accuracy metric, others are used to complement it. Like recall/sensitivity, which traduces the positive predictive capacity that the model has for a certain class.

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, the AUC (Area Under the Curve) is more suited, summarizing the results.

Another metric that is used is the Cohen's Kappa [11], 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.

Regression Evaluation Metrics - While using regression models the results are not on a black and white spectrum like classification. 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. 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. 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.

Apart from checking the absolute fitment of the model, the Coefficient of Determination, or R^2, is used to check the relative fitment of a model. This coefficient traduces the percentage variation for the dependent variable explained by the independent variables, being a strong indicator of the goodness-of-fit.

3.7 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.

3.8 Model Comparison

Student's t-test [29] is used for model comparison. 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.

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) [13] is recognized as reliable systems to rank instances according a group of metrics. 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.

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

4 Results & Discussion

This section analyzes some aspects about the results. Since a full display would not be possible, the complete set of results can be consulted in this project's GitHub[1].

The development process was done over 6 steps. The steps details are shown in the schema, Fig. 2. In the case of regression problems, there are only 5 stages, since resampling was not applied.

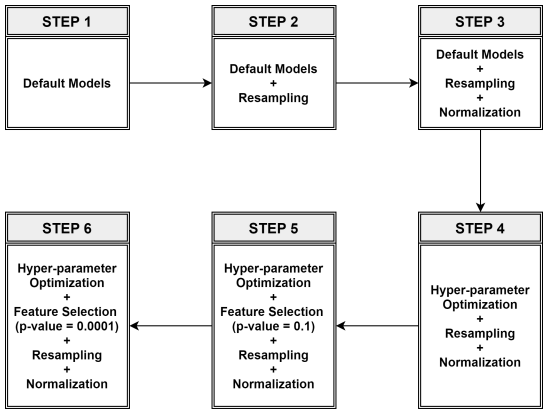

Figure 2. Schema with the main steps to create the models

In this section, 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 these reasons, the models in highlight are merely suggestive, chosen empirically through a Rank Fusion [13] method, as indicated in section 3.8.

[1]https://github.com/danielmg97/master-thesis-iposcore

4.1 Existence of Postoperative Complications

Starting with postoperative complications, 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 best results were achieved on the 4th and 5th stage, where hyperparameters optimization was applied, allied with feature selection, with p-value = 0.1, on the 5th. For all the metrics there are two excelling algorithms, SVM and LR. 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). Table 1 shows the best 5 models, according to the RRF score.

Table 1. Best 5 models - existence of complications prediction

Model	Kappa	Recall	AUC	Accuracy	RRF
SVM-5	0.37±0.14	0.68±0.07	0.73±0.08	0.69±0.07	0.36
LR-5	0.36±0.18	0.68±0.09	0.73±0.08	0.69±0.09	0.33
SVM-4	0.35±0.12	0.68±0.06	0.72±0.07	0.68±0.06	0.30
LR-4	0.33±0.13	0.67±0.06	0.72±0.07	0.67±0.06	0.28
MLP-5	0.35±0.14	0.67±0.07	0.70±0.08	0.68±0.07	0.26

4.2 Severity of Complications

Classification Approach - 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. This challenge, in specific, revealed to be the hardest outcome to predict out of the 4 initially proposed. Even after applying the SMOTEENN resampling technique to mitigate the imbalance problems, the results remained poor due to the reduced number of samples for some of the Clavien-Dindo scale degrees. The 5 best models for the complications' severity prediction are shown in Table 2. 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 2. Best 5 models - complication's severity (classification)

Model	Kappa	Recall	AUC	Accuracy	RRF
NB-6	0.15±0.05	0.23±0.08	0.65±0.03	0.40±0.08	0.29
XGB-4	0.09±0.04	0.24±0.06	0.61±0.07	0.25±0.04	0.25
NB-5	0.11±0.07	0.18±0.05	0.63±0.05	0.41±0.08	0.25
LR-5	0.04±0.03	0.25±0.08	0.66±0.07	0.10±0.07	0.25
XGB-5	0.04±0.03	0.26±0.06	0.66±0.06	0.10±0.03	0.24

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. 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.

The 5 best regression model setups are shown in the Table 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 support the higher complexity problem theory.

Table 3. Best 5 models - complication's severity (regression)

Model	MAE	RMSE	R^2	RRF
MLP-4	1.26±0.14	1.62±0.19	0.26±0.13	0.23
Ridge-4	1.27±0.13	1.62±0.19	0.25±0.12	0.21
PLS-3	1.27±0.14	1.63±0.22	0.25±0.14	0.19
Ridge-3	1.28±0.14	1.63±0.21	0.25±0.12	0.19
PLS-4	1.27±0.13	1.63±0.20	0.24±0.13	0.18

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. Table 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.

Table 4. Best 5 models - severity prediction

Model	Kappa	Recall	Accuracy	RRF
DT-Regr-4	0.20±0.07	0.21±0.02	0.50±0.04	0.19
DT-Regr-3	0.19±0.07	0.18±0.04	0.52±0.04	0.17
DT-Regr-5	0.18±0.05	0.17±0.05	0.52±0.04	0.16
NB-Class-6	0.15±0.05	0.23±0.08	0.40±0.08	0.15
SVM-Regr-3	0.16±0.08	0.16±0.06	0.47±0.06	0.14

4.3 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.

Once more, the 5 best models are presented in the table 5. 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.

Table 5. Best 5 models - days in the ICU prediction

Model	MAE	RMSE	R^2	RRF
Ridge-3	1.04±0.16	1.72±0.41	0.06±0.08	0.21
Ridge-5	1.04±0.14	1.71±0.40	0.06±0.09	0.21
kNN-5	1.01±0.15	1.72±0.43	0.06±0.05	0.19
Ridge-4	1.04±0.15	1.72±0.40	0.05±0.11	0.18
kNN-4	1.00±0.14	1.73±0.43	0.05±0.07	0.17

4.4 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 models met their peak performance in the 4th and 5th 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). This outcome shows particularly good results when predicted by tree-based models, as shown in Table 6.

Table 6. Best 5 models - 1 year death prediction

Model	Kappa	Recall	AUC	Accuracy	RRF
XGB-4	0.31±0.09	0.68±0.06	0.76±0.09	0.75±0.04	0.32
XGB-5	0.31±0.07	0.69±0.08	0.74±0.05	0.74±0.05	0.31
RF-5	0.26±0.0.1	0.70±0.07	0.76±0.08	0.67±0.06	0.30
RF-4	0.27±0.08	0.69±0.06	0.75±0.09	0.70±0.04	0.28
LR-5	0.24±0.08	0.70±0.0.7	0.75±0.08	0.62±0.11	0.25

4.5 Associative Model Study

Graphical Representation - 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. The test set error is calculated for each node individually and displayed. Additionally, leaf nodes are colored, 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, improving interpretability and confidence. A suggestive graphical representation is presented in Fig. 3, based on a Decision Tree used to predict the existence of complications.

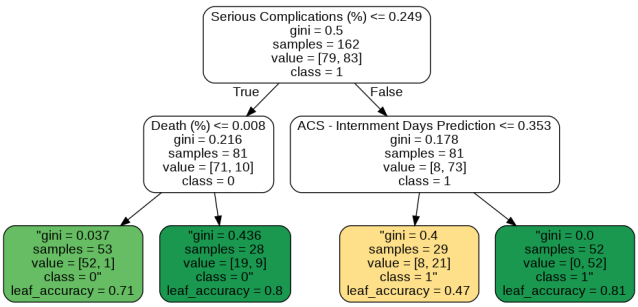

Figure 3. Example Decision Tree - Complication Prediction

5 Conclusion

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 by the doctors 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. 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.

6 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.

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 acquisition, 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.

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.

Acknowledgments

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). We thank IPO-Porto (Dr. Lúcio L. Santos) for providing the dataset.

References

[1] Amin Andalib, Agnihotram V Ramana-Kumar, Gillian Bartlett, Eduardo L Franco, and Lorenzo E Ferri. 2013. Influence of postoperative infectious complications on long-term survival of lung cancer patients:

a population-based cohort study. *Journal of thoracic oncology* 8, 5 (2013), 554–561.

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

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

[4] Karl Y Bilimoria, Yaoming Liu, Jennifer L Paruch, Lynn Zhou, Thomas E Kmiecik, Clifford Y Ko, and Mark E Cohen. 2013. 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 (2013), 833–842.

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

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

[7] 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. 2013. Colon cancer patients with postoperative complications have higher risk of recurrences. *Journal of geriatric oncology* 4 (2013), S42.

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

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

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

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

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

[13] Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. 2009. 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* (Boston, MA, USA) *(SIGIR '09)*. Association for Computing Machinery, New York, NY, USA, 758–759. https://doi.org/10.1145/1571941.1572114

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

[15] Daniel Dindo, Nicolas Demartines, and Pierre-Alain Clavien. 2004. 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 (2004), 205.

[16] Jean-Yves Dupuis, Feng Wang, Howard Nathan, Miu Lam, Scott Grimes, and Michael Bourke. 2001. 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 (2001), 194–204.

[17] Francesc Formiga, Joan Masip, David Chivite, and Xavier Corbella. 2017. Applicability of the heart failure Readmission Risk score: A first European study. *International journal of cardiology* 236 (2017), 304–309.

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

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

[20] 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. 2014. Cardiovascular disease risk score prediction models for women and its applicability to Asians. *International journal of women's health* 6 (2014), 259.

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

[22] Frank Hutter, Lars Kotthoff, and Joaquin Vanschoren. 2019. Automated Machine Learning-Methods, Systems, Challenges.

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

[24] Michal Nowakowski, Magdalena Pisarska, Mateusz Rubinkiewicz, Grzegorz Torbicz, Natalia Gajewska, Magdalena Mizera, Piotr Major, Pawel Potocki, Dorota Radkowiak, and Michal Pedziwiatr. 2018. 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 (2018), 326.

[25] 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. 2019. Machine Learning Approaches to Predict 6-Month Mortality Among Patients With Cancer. *JAMA network open* 2, 10 (2019), e1915997–e1915997.

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

[27] Karl Pearson. 1901. 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 (1901), 559–572.

[28] 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. 2016. Predicting colorectal surgical complications using heterogeneous clinical data and kernel methods. *Journal of biomedical informatics* 61 (2016), 87–96.

[29] Student. 1908. The probable error of a mean. *Biometrika* (1908), 1–25.

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

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

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

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

[34] Maciej Zikeba, Jakub M Tomczak, Marek Lubicz, and Jerzy Swikatek. 2014. 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 (2014), 99–108.