Predicting Surgical Case Durations for a Thorax Centre

Primary Topic: DM, Secondary Topic: DPV
Course: 2020-1B – Group: 106 – Submission Date: 2021-01-29

Yoran Greg de Weert
s1693859
University of Twente
y.r.deweert@student.utwente.nl

Selina Zwerver
s1690388
University of Twente
s.zwerver@student.utwente.nl

ABSTRACT

Surgery scheduling is a challenging task in hospitals as the scheduling depends on the accuracy of surgery duration predictions. In this study, surgery durations were predicted using different data mining methods. A dataset was provided containing information about 4087 surgical cases, containing 360 unique surgery types of which the ten most occurring were predicted. A feature selection method was used to determine which features are included and five regression models were trained and tested. Generally, the models improved the predictions around 7.5% compared to the baseline model. Only relatively small differences were found between the considered models and pointed out that no model stood out from the other models.

KEYWORDS

Data mining, prediction, surgical case duration, thorax centre, data preparation, data visualisation, regression

1 INTRODUCTION

Scheduling surgeries is a highly difficult management task in hospitals. Underestimating surgery duration leads to high waiting time for patients, while overestimating leads to ineffective use of expensive resources. In constructing the schedule, accurate prediction surgery duration is of high importance. Surgery durations have, however, high inherent variability [1]. To decreases this variability, one could take all information about the surgery that is available during planning into account. Many factors are available, such as surgery type and severity of the case, hence it is important to select the ones that have a significant effect on the variability.

A data set is available containing information about surgical cases at a thorax centre. The research question is therefore:

Which model estimates the surgery duration of the 10 most occurring surgery types most accurate?

With corresponding sub-question *How accurate is this model compared to the initial predicted surgery duration?* The data is used to develop a data-driven predictive tool to accurately predict surgery durations. The method consists of two steps: 1) feature selection and 2) prediction.

In Section 2, some relevant researches and methods related to this topic are highlighted. Section 3 describes the approach used in this paper for predicting the surgery duration. The experiments and corresponding results are described in Section 4. Section 5 discusses the results, limitations and occurred problems of this research and Section 6 delivers conclusions.

2 RELATED WORK AND BACKGROUND

Several studies have been conducted on predicting surgery to improve scheduling. Ibrahim et al. [2] consider three families of models. The first family incorporates the physician input, which takes the surgeon's predictions of the surgery duration into account. The second family contains statistical models such as regression historical average models and the third family combines the physician input models with the statistical models. A model of the third family performed best, with a 30% decrease in mean squared error compared to the benchmark model, which predicted the surgery duration directly from the predicted duration.

Kargar et al. [3] used linear regression (LR), multivariate adaptive regression splines (MARS) and random forest (RF) algorithms to build the prediction models. LR is a common approach and assumes a linear relation between the target variable and the set of predictors. MARS extends linear models with nonlinearities and interaction between variables and RF is a combination of decision trees. The RF model performed best, increasing the R^2 value by 34.4% compared to the hospital prediction.

Ng et al. [4] took heterodacity (varying variance of the error term in a regression model) into account by predicting the surgery duration ($\hat{y}$) and standard deviation simultaneously ($\hat{\sigma}$) in a multilayer perceptron network. They showed that surgery durations are indeed heteroscedastic and obtained a 20% improvement compared to current scheduling techniques.

2.1 Background

The related work shows promising outcomes for all models considered. Therefore, LR, MARS, RF and MLP are chosen to experiment with along with Gradient Boosting Regression. Two models that are considered, require additional explanation: Multilayer Perceptron (MLP) Regressor and Gradient Boosting Regression (GBR). These will briefly be explained in this section.

2.1.1 MLP. This modelling is a form of a feed-forward artificial neural network (i.e. connections between nodes in the network do not form cycles). The MLP consists of at least three layers of nodes (input layer, hidden layer and an output layer) connected by weights and output signals, which depend on a function of the sum of inputs to the node. The MLP is described as fully connected, i.e. each node is connected to all other nodes in the previous or next layer. The network can be trained by varying individual weights of connections which results in an so-called error surface, showing the error of the model compared to the function given a set of weights. The goal is to find a combination of weights that minimises the error, corresponding to the minimum point of the error surface. [5]

2.1.2 GBR. This method also applies to function estimation, where a function, F, has to be estimated. The main idea is to combine weak predictors (as for example decision trees) to create one strong predictor. It boosts the process by adding a new estimator to the function representing the residual of the model value and the function value given a input, x, each iteration, m. This can be mathematically described as:

$$F_m(x) = F_{m-1}(x) + \beta \cdot h(x; a_m) \tag{1}$$

Where β_m is the expansion coefficient and $h(x; a_m)$ the base learner, commonly defined as a simple function of x with parameters $a_m = \{a_1, a_2, ...\}$. [6]

3 APPROACH

A total of five models were trained and tested. This section describes these models and the method that is used to select the features influencing the duration, as input for the models.

3.1 Feature selection

Machine learning models need a set of features as input to predict surgery durations and only features that contribute to surgery duration should be taken into account. To estimate this contribution, two types of features are distinguished: numerical and categorical. For both types, the method is described to estimate the effect on the surgery duration.

3.1.1 Dependency on numerical features. The dependency of the surgery duration on a numerical feature can be estimated by determining the correlation between them. The correlation can be represented by Pearson's correlation coefficient for a sample and can be computed using Equation 2: [7]

$$\rho_{X,Y} = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \overline{Y})^2)}}, \tag{2}$$

where $r_{X,Y}$ represents the Pearson's correlation coefficient for numerical features X and Y. The variables $\overline{X}, \overline{Y}$ represent the sample means for X and Y, respectively. Values that are close to 1 indicate a positive correlation, values close to -1 indicate a negative correlation and values close to zero imply a low or no correlation.

3.1.2 Dependency on categorical variables. Correlation cannot be used to determine the contribution of a categorical feature to the surgery duration. Instead, the weighted arithmetic mean of variances is used. The computation of weighted mean of variances for a categorical feature X, M_X, is shown in Equation 3, where N is the amount of categories per feature, $Var(X_j)$ the variance in surgery duration for category $X_i \subset X$, also shown in Equation 4. K_{X_i} is the number of observations for a category X_i and L_X the total number of observations for feature X.

$$M_X = \sum_{j=0}^{N} Var(X_j) \cdot \frac{K_{X_i}}{L_X}, \quad \text{where} \tag{3}$$

$$Var(X_j) = \frac{1}{K_{X_i}} \sum_{j=1}^{K_{X_i}} (X_{i,j} - \overline{X_i})^2 \tag{4}$$

The weighted mean of variances contains the variances of the surgery duration for single categories of a feature. If these variances are relatively low, there is a relatively small amount of deviation of surgery duration the implying that the dependency of the duration on the feature is relatively high. On the other hand, relatively high variances per category result in a higher weighted mean of variances implying a relatively low dependency of the duration on the considered variable. To account for missing data, the weighted mean was multiplied by the percentage of not-`NaN` data in the feature.

3.2 Model

Five models were created: Linear Regression (LR), Multivariate Adaptive Regression Splines (MARS), Random Forest (RF), Multilayer Perceptron Regressor (MLP) and Gradient Boosting Regression (GBR). The models were created using the `sklearn` Python package. Using default parameters for the models, mean absolute errors (MAEs) were determined for various amounts of categorical features taken into account during training, to determine the amount of required features. For the RF, MLP and GBR models, hyperparameter optimisation was performed using grid search. The random state for each model was set to 41 to ensure reproducibility of the results.

To verify the accuracy of the predicted surgery durations for all surgery types, the mean absolute error (MAE), mean absolute percentage error (MAPE) and root mean square error (RMSE) were calculated. The statistics were compared to a baseline prediction, the MAE for the predicted surgery duration and actual surgery duration. To determine which surgery type can be predicted best, the MAE and MAPE between the surgery duration, predicted by the model, and realisation as well as the standard deviation of the MAE were calculated for each model and again compared to the baseline.

4 EXPERIMENTS

The proposed models are applied on a data set containing information about surgical cases. The data set contains 4087 surgical cases performed from January 2013 to January 2016 at TCT. An overview of the features available for each surgery is shown in Appendix A.

4.1 Analysis

In order to use the data in the models, the data was preprocessed as shown in Appendix B.

4.1.1 Statistics. The amount of missing data per feature was determined by counting the occurrence of `NaN`. Features 'Nierfunctie' and 'Linker ventrikel functie' show over 80% of missing data (see Appendix C), indicating that these variables cannot be used for predicting the surgery duration.

Observation shows that some cases have extreme deviations in the predictions of the surgery durations compared to the actual duration. Surgery 632 had a scheduled duration of 50 minutes while the surgery lasted 590 minutes, resulting in a 1080% deviation. In total, 5.7% of the operations showed a deviation greater than 100% and these were therefore excluded from the data set to avoid unrealistic improvements.

4.1.2 Selecting the 10 most occurring surgeries. To predict the 10 most occurring surgeries, a threshold was determined in order to remove the remaining surgery types. A graph of the fraction of data and amount of unique surgery types left for a certain threshold, the threshold being the amount of times an operation occurs, is shown in Figure 1. To retain 10 unique surgery types, the threshold must be placed at 61. In this case there is 62% of the total data set left. The 10 most occurring surgeries in order of amount of occurrences are 'CABG', 'AVR', 'CABG + Pacemakerdraad tijdelijk', 'CABG + AVR', 'Wondtoilet', 'MVP', 'AVR + MVP shaving', 'Mediastinoscopie', 'Lobectomie of segmentsectie' and 'Rethoracotomie'.

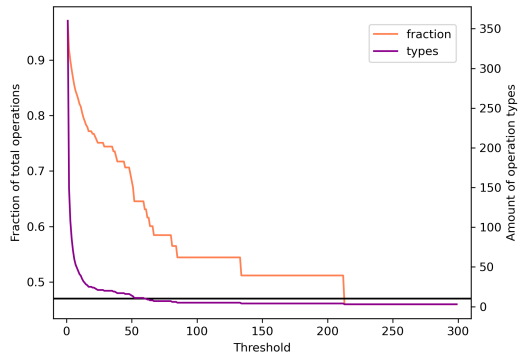

Figure 1: The fraction of data and amount of unique surgery types left for a certain threshold value that indicates the amount of surgery types in the data. The black line indicates where the amount of unique surgery types left equals 10.

4.1.3 Feature dependency. The dependency of features was computed as described in Section 3. The result was a correlation matrix for the numerical features, which is shown in Figure 2. The correlation matrix shows no strong correlation between the numerical features and the surgery duration, indicating that these features are not significantly contributing to the prediction of the duration.

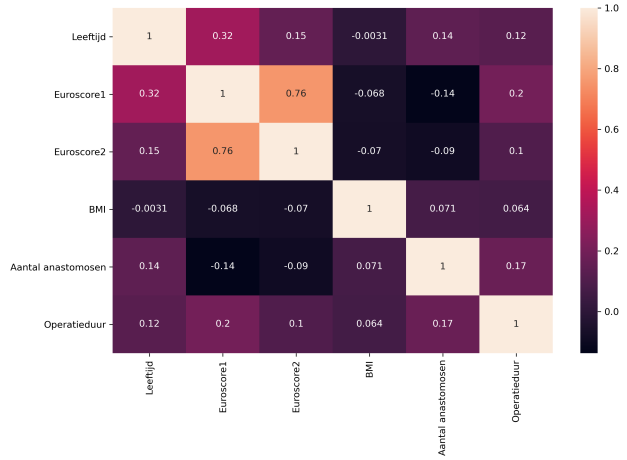

Figure 2: Correlation matrix for the numerical features.

The dependency of the duration on the categorical features can be found in Appendix D and the best scoring variables are shown in Table 1. Features 'Operatietype' can best explain the variance in the target, followed by 'HLM' and 'Benadering'.

Variable	Fraction
Operatietype	0.404334
HLM	0.812712
Benadering	0.817327
⋮	⋮

Table 1: Weighted mean of variances per categorical feature.

4.2 Experimental settings

4.2.1 Feature selection. Features were sorted according to increasing dependency score (see Appendix D). Figure 3 shows that the initial error is high for most models, decreases when taking the second and third feature into account, and then remains approximately constant. The RF models shows an increase in error after taking the second feature into account, after which it gradually decreases. The LR model follows the other models during the first three features, but its error explodes after taking 4 features into account. Due to the amount of missing data at this point, only 8 surgery types can be predicted. It is therefore chosen to use 3 features. These features are 'Operatietype', 'HLM' and 'Benadering'.

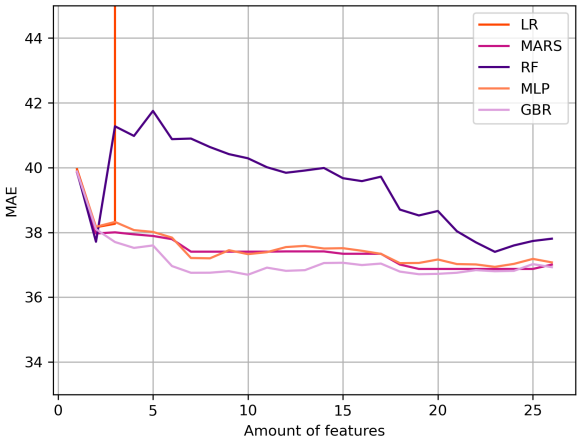

Figure 3: MAE per amount of features for the five models.

4.2.2 Hyperparameter optimisation. As stated in Section 3, hyperparameter optimisation was performed for the RF, MLP and GBR models. The grids are specified in Appendix E. An overview of the optimal parameters used for each model is shown in Appendix F.

4.2.3 Data split. The data was split into training and testing data. 80% of the data was used for training and 20% for testing. All models were trained and tested on the same sets.

4.3 Results

Experimentation showed that surgeries 'Rethoracotomie' and 'Wondtoilet' did not contain enough data for a proper split after removing the missing data. As a result, only 8 surgeries are predicted. The overall statistics for the models on the test set are shown in Table 2. All models show an improvement compared to the baseline, where RF performs best with an improvement of 8.04% and MARS the worst with 7.08%. Although all models perform better than the baseline, the performances does not differ much between the models as all models show similar values for the MAE, MAPE, RMSE and improvement.

Table 2: Total MSE, MAPE, RMSE and Improvement (Imp) for the models on the test set.

Model	MAE	MAPE	RMSE	Imp (%)
LR	36.2834	21.6062	48.7450	7.3373
MARS	36.3860	21.7019	48.7831	7.0753
RF	36.0100	21.5393	48.4642	8.0357
MLP	36.3995	21.8516	48.8937	7.1505
GBR	36.0277	21.6406	48.4901	7.9904

The statistics of the models are additionally calculated for each individual surgery type. This is shown for RF and MLP in Tables 3 and 4, because these contain the largest individual differences and are based on other methods. Results for the other models are shown in Appendix G. All models show a performance decrease for the surgery types 'AVR', 'AVR + MVP shaving' and 'Mediastinoscopie' compared to the baseline. The most improved surgery is is 'CABG + AVR'.

Tables 3 shows that RF has the lowest MAPE for 'AVR + MVP shaving', 'CABG + Pacemakerdraad tijdelijk' and 'CABG + AVR', hence RF predicts best for these surgery types. The results for 'AVR + MVP shaving' are worse then the baseline. The relative best improvements compared to baseline are for 'CABG + AVR' and the 'Lobectomie of segmentresectie'.

Table 3: MAE, MAPE and Improvement (Imp) per surgery type for the RF model.

Surgery type	MAE	MAPE	Imp (min)
AVR	34.25 ± 29.96	21.69	3.27
AVR + MVP shaving	32.11 ± 25.30	13.48	6.89
CABG	34.28 ± 30.43	25.04	-2.12
CABG + AVR	49.71 ± 44.48	17.18	-19.37
CABG + Pacemakerdraad...	33.34 ± 30.56	16.35	-7.89
Lobectomie of...	60.45 ± 33.60	23.48	-31.05
Mediastinoscopie	48.45 ± 43.86	41.57	2.85
MVP	18.5 ± 2.43	34.69	-5.00

The lowest MAPE values for MLP belong to the surgery types 'AVR + MVP shaving', 'CABG + Pacemakerdraad tijdelijk' and 'CABG + AVR', indicating that MLP can predict these best. This is equal to the results for RF. The main difference with RF is the prediction for the surgery type 'Lobectomie of segmentresectie'. Here, the RF model is better than the baseline and has a MAPE of 23.48%, while MLP performs worse than the baseline with a MAPE of 40.54%. The MAE additionally differs approximately 40 minutes.

Table 4: MAE, MAPE and Improvement (Imp) per surgery type for the MLP model.

Surgery type	MAE	MAPE	Imp (min)
AVR	35.20 ± 30.24	22.31	4.23
AVR + MVP shaving	31.75 ± 22.93	13.21	6.53
CABG	34.25 ± 30.34	24.97	-2.16
CABG + AVR	48.51 ± 43.58	17.43	-20.57
CABG + Pacemakerdraad...	33.33 ± 30.48	16.69	-7.90
Lobectomie of...	102.16 ± 38.89	40.54	10.66
Mediastinoscopie	52.15 ± 44.71	40.46	6.55
MVP	18.50 ± 19.98	41.61	-5.00

The amount of occurrences in the test set of the 8 most occurring surgery types is shown in Table 5. 'CABG' contains the most the occurrences in the test set, while the 'Lobectomie of segmentresectie' and 'MVP' only occur twice. Additionally, the average surgery duration and corresponding standard deviation are shown in the table. The MAE of the models is relatively small compared to the average surgery duration, the MAE falls approximately within the standard deviation.

Table 5: Amount of occurrences (#) of each surgery type in the test set with corresponding mean duration (minutes) and standard deviation (std).

Surgery type	#	Mean + std
AVR	75	205.13 ± 44.46
AVR + MVP shaving	14	236.64 ± 34.66
CABG	223	230.70 ± 48.62
CABG + AVR	36	283.17 ± 64.30
CABG + pacemaker tijdelijk	69	263.10 ± 46.70
Lobectomie of segmentresectie	2	247.50 ± 38.90
Mediastinoscopie	20	224.05 ± 68.93
MVP	2	60.50 ± 26.16

5 DISCUSSION

5.1 Results

5.1.1 Significance of results. Table 2 shows that the general results of the models are similar, the difference between the best and worst performing model compared to the baseline being less than 1%. The chosen models use different approaches to predict the surgery duration, and optimisation was performed where possible, hence the results suggest that there is not enough information in the data set to obtain better results. This is, however, dependent on the feature-selection process that was used. It is possible that feature selection was sub-optimal and better results could be obtained using a different method. This is confirmed when regarding the standard deviation, which is relatively large, indicating that more features are required for accurate prediction. The improvement additionally is small compared to the literature presented in Section 2.

In the individual results it can be seen that the greatest and smallest errors can be found in the 'Lobectomie of segmentresectie' and 'MVP' surgery types, respectively. Taking Table 5 into account, these results are likely unreliable. Both surgery types only contain two observations in the test set. If these observations are far from, or close to, the predicted value, the error value can respectively be very large or small. Therefore, these surgery types can not be taken into account when deciding model accuracy. The other surgeries show little difference in error and improvement between the models.

5.2 Methodology

5.2.1 Feature selection. Figure 3 shows the MAE of default models when taking a certain amount of features into account. The chosen amount of features was three, primarily based on the amount of surgery types that could be predicted. It is seen that the error decreases when taking more than 3 features into account. This suggests that an increase in the amount of features during training would result in a better prediction of the surgery duration. The case is however that the amount of data that is left when using more features decreases. As a result, the test set becomes small and the error can decrease. Another possibility to increase the accuracy of the results would be to include the numerical features. Currently, the categorical features were chosen as the model input because the numerical features showed little to no correlation with the surgery duration. The methods of assessment (correlation and weighted variation) differ from each other hence it is not possible to compare the results directly. It might be that some of the numerical features are more useful than some of the categorical features that are currently taken into account.

5.2.2 Multiple correspondence analysis. Instead of the current assessment of importance of the features, multiple correspondence analysis (MCA) could be used. MCA is an extension of principal component analysis (PCA), meant for categorical features [8]. MCA is, however, only useful for features that take binary values (1 or 0). This is useful for the features that only contain 'yes' or 'no', but not for, for example, surgery type. As there are 360 surgery types in the data set, each of these would result in a separate column as explained in one hot encoding. This would make the amount of categorical features much larger than the initial set and likely useless.

5.2.3 Nan data. The loss in data during feature selection can be prevented by filling the empty (`NaN`) slots. There are multiple algorithms available to perform this action. Examples include bottom-up filling, top-down filling, filling with average or most frequent and Multivariate Imputation by Chained Equation (MICE). In the first two cases, the missing data point would be filled by either the previous or next data point. If the column would be `0–NaN–1` it would either be filled as 0–0–1 (top-down) or 0–1–1 (bottom-up). The second solution, filling with average or most frequent simply fills the gap with the average or most frequent value in the column. The problem with these two solutions is that the other data in the set is not taken into account. As a result, the gap could be filled with a value that is either not possible or makes no sense. A solution would be MICE, which creates multiple imputations to account for statistical uncertainty and performs a chained approach to handle variables of varying types, basing the missing values on the complete set [9]. MICE can, however, not deal with string types hence the categorical variables must be label encoded. For the surgery types, the data set would therefore contain integers but when performing MICE, the missing data is filled with floats. Therefore, a surgery type '2.4' can exist in the data, which is useless because it does not correspond to a true surgery type.

5.2.4 LR error. As can be seen in Figure 3, the MAE for the LR model explodes after taking three features into account. This behaviour can be due to various causes. A first explanation could be that, after using three features, the data set becomes highly nonlinear. The MARS model, taking nonlinearities into account to a certain severity, is however unaffected, suggesting that nonlinearity is not the cause of this problem. The second explanation is that the one hot encoded data is multicollinear. Multicollinearity means that values in certain columns can be predicted by combinations of other columns [10]. This is a consequence of one hot encoding the data. Label encoding would not resolve this problem, since the input for a linear regression model is absurd in such a case. Since the MARS and MLP models no not show this large increase in error from three features onward, this multicollinearity seems only problematic for the LR model.

5.2.5 K-cross validation. In the current training and testing of the models, only one split is used. To properly train and validate the models, k-cross validation could be used. In such a case, the data would be split into k groups. Then for each group, the group would be taken as the test set and the remaining groups as training sets. The model would be fit on the training groups and tested on the test group. The evaluation scores would be retained and the trained model would be discarded. As a final step, the scores of all models (same type of model, trained and tested on various sets) would be combined into final evaluation scores. This method could provide a better estimate of the accuracy of the models. [11]

6 CONCLUSIONS

Five models were trained to predict the surgery duration. Except for surgery types 'AVR' and 'AVR + MVP shaving', the models showed an improvement in prediction of approximately. The differences in results between the five models were relatively small, indicating that the performances of the models are similar, taking into account the used approach and data set. No best performing model could be pointed out from the five tested models. All models showed an overall improvement of approximately 7.5%.

6.1 Future work

As a future direction, results could be improved by increasing the amount of surgeries as model input. This can be achieved by adding more records of surgery cases or by using the data more effectively, using methods to predict the values for missing data.

In this study, surgery type is used as input variable to predict the duration. It could, however, be worthwhile to develop and optimise independent models for each surgery type. This has as advantage that the surgery is not an input feature and a future direction is to see if this results in higher prediction accuracy.

As only one feature selection method is considered in this study, an interesting direction is to study different feature selection methods. This does not only change the prediction of the surgery duration, it may as well change the behaviour and performance of the considered models.

REFERENCES

[1] Enis Kayis et al. "Improving prediction of surgery duration using operational and temporal factors". In: *AMIA ... Annual Symposium proceedings. AMIA Symposium* (2012), pp. 456–462. ISSN: 1942-597X. URL: https://europepmc.org/articles/PMC3540440.

[2] Rouba Ibrahim and Song-Hee Kim. "Predicting Surgery Duration: Physician Input, Statistical Models, and Combined Models". In: (2018).

[3] Zahra Shahabi Kargar et al. "Predicting Procedure Duration to Improve Scheduling of Elective Surgery". In: (2014). DOI: 10.1007/978-3-319-13560-1_86.

[4] Nathan H Ng et al. "Predicting Surgery Duration with Neural Heteroscedastic Regression". In: (2017).

[5] Matt W Gardner and SR Dorling. "Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences". In: *Atmospheric environment* 32.14-15 (1998), pp. 2627–2636.

[6] Jerome H Friedman. "Stochastic gradient boosting". In: *Computational statistics & data analysis* 38.4 (2002), pp. 367–378.

[7] Jacob Benesty et al. "Pearson correlation coefficient". In: *Noise reduction in speech processing*. Springer, 2009, pp. 1–4.

[8] Michael Greenacre and Jorg Blasius. *Multiple Correspondence Analysis and Related Methods*. Chapman and Hall/CRC, 2006. ISBN: 9781584886280.

[9] Melissa J. Azur et al. "Multiple imputation by chained equations: what is it and how does it work?" In: *Int J Methods Psychiatr Res* 20 (1 Mar. 2011), pp. 40–49. DOI: 10.1002/mpr.329.

[10] Moeedlodhi. *How Multicollinearity Is a Problem in Linear Regression*. Accessed on: 23 January 2021. Aug. 2020. URL: https://medium.com/swlh/how-multicollinearity-is-a-problem-in-linear-regression-dbb76e25cd80.

[11] G. James et al. *An Introduction to Statistical Learning with Applications in R*. Springer-Verlag New York, 2013. ISBN: 978-1-4614-7138-7. DOI: 10.1007/978-1-4614-7138-7.

A DATASET OVERVIEW

Feature	Category (definition)
Operatietype	AVR
	AVR + MVP
	CABG + AVR
	⋮
Chirurg	1,00-15,00
	Ander specialisme
Anesthesioloog	3,00-19,00
	Onbekend
Benadering	Volledige sternotomie
	Antero lateraal links
	Antero lateraal rechts
	⋮
OK	HCK1
	HCK3
	OK 1
	⋮
Casustype	Electief
	Spoed < 24 uur
	Acuut < 30 minuten
	⋮
Dagdeel	Ochtend/Middag/Avond/Nacht
Leeftijd	Continuous
Geslacht	M/F
AF	J/N
Chronische longziekte	J/N
Extracardinale vaatpathie	J/N
Eerdere hartchirurgie	J/N
Actieve endocarditis	J/N
Kritische preoperatieve status	J/N
Myocard infact < 90 dagen	J/N
Aorta chirurgie	J/N
Pulmonale hypertensie	Normaal/Matig/Ernstig
Linker ventrikel functie	Goed/Matig/Slecht/Heel slecht
Euroscore 1	Continuous
Euroscore 2	Continuous
Nierfunctie	Normaal/Matig/Slecht/Dialyse
Slechte mobiliteit	J/N
BMI	Continuous
DM	J/N
Hypercholesterolemie	J/N
Hypertensie	J/N
Perifeer vaatlijden	J/N
CSS	0, 1, 2, 3, 4
NYHA	1, 2, 3, 4
Aantal anastomosen	Continuous
HLM	J/N

Table 6: Overview of the features in the dataset.

Outcome	Category (definition)
Geplande operatieduur	Continuous
Operatieduur	Continuous
Ziekenhuis ligduur	Continuous
IC ligduur	Continuous

Table 7: Overview of the targets in the dataset.

B DATA PREPROCESSING

Comma separation in numeric values was replaced by point separation in order to reliably use the data in processing. Unknown ('Onbekend') data, and 'Andere specialisme' was replaced by `NaN`.

In order to use the categorical features in the LR, MARS and MLP models, the categories must be encoded. Categories can be encoded using one hot encoding or label encoding. One hot encoding creates a binary column for each category, resulting in various binary arrays for the combinations of features while label encoding assigns a numerical value to categories. An example of both encoding type is shown in Table 8. Label encoding allows the models to capture a ranked relation between the variables, e.g. in the example we have that Apple > Banana > Pear. This behaviour must be avoided, hence one hot encoding is used.

Table 8: Encoding of category 'Fruit' (a) using label encoding (b) and one hot encoding (c).

Fruit	**Fruit**	**Apple**	**Pear**	**Banana**
Apple	0	1	0	0
Pear	2	0	1	0
Banana	1	0	0	1
a)	**b)**		**c)**	

C MISSING DATA PER VARIABLE

The percentage of missing data per feature is shown in Figure 4.

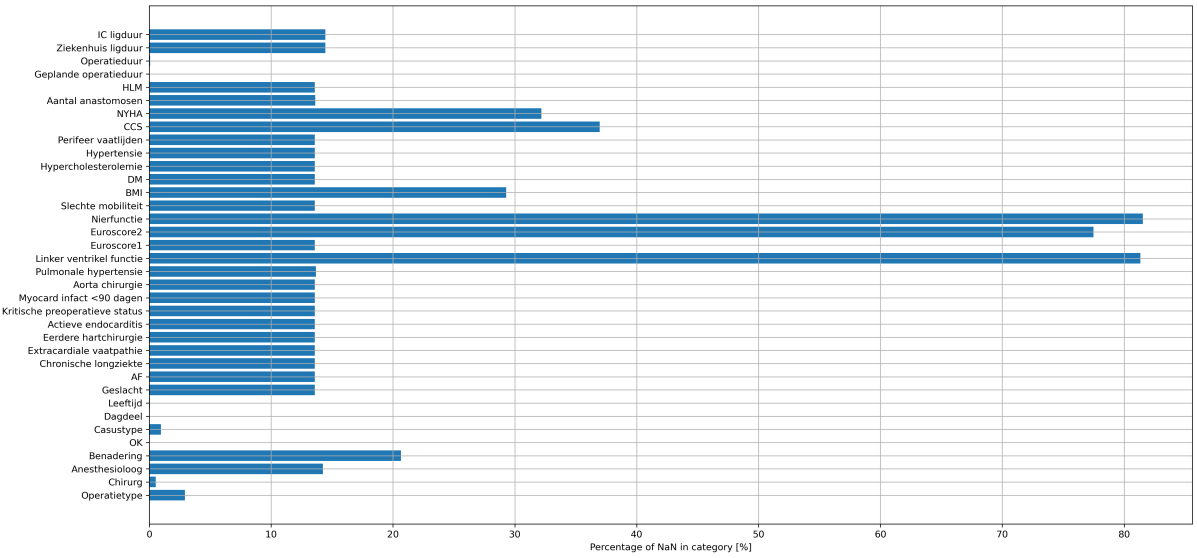

Figure 4: Percentage of missing data per feature.

D FRACTION OF VARIANCE PER CATEGORICAL VARIABLE

Table 9: Weighted variance of each categorical feature.

Feature	Variance
Operatietype	0.404334
HLM	0.812712
Benadering	0.817327
Aorta chirurgie	0.836164
AF	0.876498
Hypertensie	0.8779
Eerdere hartchirurgie	0.878513
Actieve endocarditis	0.879336
Hypercholesterolemie	0.883028
Extracardiale vaatpathie	0.884076
Chronische longziekte	0.884145
DM	0.884389
Myocard infact <90 dagen	0.885059
Perifeer vaatlijden	0.885122
Geslacht	0.88514
Pulmonale hypertensie	0.885155
Kritische preoperatieve status	0.8856
Slechte mobiliteit	0.88602
Anesthesioloog	0.894166
Dagdeel	0.949306
Casustype	0.956833
OK	0.979303
Chirurg	0.999037
NYHA	1.21675
CCS	1.3583
Linker ventrikel functie	5.1656
Nierfunctie	5.24122

E GRID PARAMETERS

E.1 Random forest

Table 10: Grid parameters for the random forest regressor. In case of ranges, the first number indicated the minimum, the second the maximum and the third the step size.

Parameter	Options
Number of trees	10, 200, 10
Minimum number of samples required to split an internal node	2, 10, 1
Minimum number of samples required to be at a leaf node	1, 10, 1
Number of features to consider	Auto, square root logarithmic
Maximum depth of the tree	None and 1, 100, 1
Random state	41
Bootstrap	True or false
Function to measure split quality	MSE or MAE

E.2 Multilayer perceptron

Table 11: Grid parameters for the multilayer perceptron regressor. In case of ranges, the first number indicated the minimum, the second the maximum and the third the step size.

Parameter	Options
Hidden layer size	1, 200, 1
Activation function	Identity, sigmoid, tan, rectified linear unit
Solver	Quasi-Newton, stochastic gradient descent, adam
Learning rate schedule	Constant, gradual decrease, adaptive
Random state	41

E.3 Gradient boosting

Table 12: Grid parameters for the Gradient Boosting regressor. In case of ranges, the first number indicated the minimum, the second the maximum and the third the step size.

Parameter	Options
Loss function	Least squares regression, least absolute deviation, combination of the previous or quantile regression
Number of boosting stages	10, 200, 10
Function to measure split quality	Friedman MSE, MSE, MAE
Minimum number of samples required to split an internal node	2, 10, 1
Minimum number of samples required to be at a leaf node	1, 10, 1
Maximum depth of the individual regression estimators	None and 10, 100, 1
Number of features to consider	Auto, square root, logarithmic, none
Random state	41

F MODEL PARAMETERS

F.1 Random Forest

Table 13: Optimised parameter values for the RF model.

Parameter	Value
Number of trees	178
Minimum number of samples required to split an internal node	2
Minimum number of samples required to be at a leaf node	2
Number of features to consider	Square root
Maximum depth of the tree	None
Random state	41
Bootstrap	False
Function to measure split quality	MAE

Table 14: Optimised parameter values for the MLP model.

Parameter	Options
Hidden layer size	49
Activation function	Tan
Solver	Quasi-Newton
Learning rate schedule	Constant
Random state	41

F.2 Gradient Boosting Regression

Table 15: Optimised parameter values for the GBR model.

Parameter	Value
Loss function	Combination of least squares regression and least absolute deviation
Number of boosting stages	52
Function to measure split quality	MAE
Minimum number of samples required to split an internal node	2
Minimum number of samples required to be at a leaf node	4
Maximum depth of the individual regression estimators	None
Number of features to consider	Square root
Random state	41

G RESULTS

Table 16: MAE, MAPE, Improvement (Imp) and standard deviation (Std) per surgery type for the LR model.

Surgery type	MAE	MAPE	Imp (%)
AVR	35.95 ± 30.41	22.52	4.97
AVR + MVP shaving	31.79 ± 24.79	13.43	6.57
CABG	33.97 ± 30.02	24.63	-2.43
CABG + AVR	48.67 ± 44.17	17.32	-20.42
CABG + Pacemakerdraad...	33.46 ± 30.49	16.86	-7.77
Lobectomie of...	109.00 ± 24.04	43.34	17.50
Mediastinoscopie	48.45 ± 44.12	40.90	2.85
MVP	18.50 ± 2.12	34.57	-5.00

Table 17: MAE, MAPE and Improvement (Imp) per surgery type for the MARS model.

Surgery type	MAE	MAPE	Imp (%)
AVR	36.64 ± 30.85	22.79	5.66
AVR + MVP shaving	31.15 ± 24.97	13.43	5.93
CABG	34.02 ± 30.19	24.87	-2.39
CABG + AVR	49.75 ± 44.47	17.35	-19.33
CABG + Pacemakerdraad...	33.72 ± 30.68	17.44	-7.51
Lobectomie of...	76.59 ± 1.42	30.08	-14.91
Mediastinoscopie	48.46 ± 43.52	40.58	2.86
MVP	18.50 ± 15.56	27.60	-5.00

Table 18: MAE, MAPE and Improvement (Imp) per surgery type for the GBR model.

Surgery type	MAE	MAPE	Imp (%)
AVR	34.40 ± 30.06	21.87	3.43
AVR + MVP shaving	31.81 ± 24.24	13.37	6.59
CABG	34.33 ± 30.57	25.15	-2.07
CABG + AVR	48.90 ± 44.38	17.17	-20.18
CABG + Pacemakerdraad...	33.32 ± 30.48	16.66	-7.91
Lobectomie of...	70.38 ± 33.14	27.54	21.12
Mediastinoscopie	48.45 ± 43.37	41.70	2.85
MVP	18.50 ± 3.06	32.53	-5.00