UTRECHT UNIVERSITY

MASTER THESIS

Predicting Novasure surgery outcomes based on patient characteristics and treatment activities

Author:
Zoey E. HULZEBOS
6904289

Supervisor:
Dr. ir. X. LU
Second supervisor:
Dr. G.C. VAN DE WEERD
Supervisor Maxima Medisch Centrum:
Jaklien C. LEEMANS, PhD

Thesis Master of Science

Graduate School of Natural Sciences
Business Informatics
Business Process Management and Analytics

September 27, 2022

UTRECHT UNIVERSITY

Abstract

Business Informatics
Graduate School of Natural Sciences

Master of Science

Predicting Novasure surgery outcomes based on patient characteristics and treatment activities

by Zoey E. HULZEBOS
6904289

For 30% of the women, their daily life is influenced by heavy menstrual bleeding, meaning that their energy level, mood, work productivity, social interaction, family life, and sexual functioning alternate due to their menstruation cycle. Endometrial ablation, like the Novasure surgery, can be used as definitive surgical treatment. Endometrial tissue is vaporised during this process, preventing the flow possibility. Such surgery has failed when tissue grows back, flow continues or complaints reoccur. Current medical research has found multiple prognostic factors associated with the failure of the Novasure surgery. While these factors are purely focused on the patient's characteristics, the question arises whether process features as well point at the failure of the Novasure surgery. This research investigates in the use of historical data of Novasure patients to provide evidence-based insights into current treatments and their impacts on the outcome of the Novasure surgery per patient. Using six machine learning algorithms in four experiments with patient characteristics and process features as potential prognostics factors, the most important features are predicted and the most effective algorithms is determined. Adenomyosis, age, BMI, cavity length, and cavity width are the patient characteristics which have the most influence on the outcome of the Novasure surgery. The addition of process features led to the awareness that investigating in care activities and appointments brings new insights in predicting reinterventions. Random forest, extreme gradient boosting and neural network are the algorithms which can be used best for predicting which patients are likely to undergo a reintervention.

Acknowledgements

There are some people who deserve to mentioned because I wish thank them for several reasons. First of all Xixi, for all the hours of discussing, for pushing me a little further every time, yourguidance and motivating words when needed. As in the beginning of this project, I hope sunny weather is coming.

Jaklien, for your enthusiasm and inquisitiveness about data science, enabling to attend multiple genealogist surgeries, seeing the department from a different view, and the honest words I sometimes needed. Suhwan, for your insights, lots of help in programming and the perspective from a male on this project.

Inge, for first getting me in touch with Xixi and the curiosity.

My mum, dad, and Peijo, for the talks, drinks and outreaching hands.

Hilde, Jochem, Joost, for making my time in Utrecht during the pandemic a little less lonely.

Tessa, for the great food, talks, rhythm and ability to bike over the summer.

And, for the sweetest words at the right time, Heleen, Karel, Nicky, Joyce, Ruben, Nienke, Alex and multiple Vera's.

And Nicky again, for making it physically possible to finish this project by lending me her charger.

Contents

List of Figures

List of Tables

List of Abbreviations

AUB	**A**bnormal **U**terine **B**leeding
AUC	**A**rea **U**nder the **C**urve
AUE	**A**bdominal **U**terus hyster**E**ctomy
AVF	**A**nte**v**ersie**f**lexie: Anteverted uterus
BMI	**B**ody **M**ass **I**ndex
CRISP-DM	**C**ross **I**ndustry **S**tandard **P**rocess for **D**ata-**M**ining
DMME	**D**ata **M**ining **M**ethodology for **E**ngineering applications
DNN	**D**eep **N**eural **N**etworks
DT	**D**ecision **T**ree
EHR	**E**lectronic **H**ealth **R**ecord
GB	**G**radient **B**oosting
GBM	**G**radient **B**oosting **M**odel
GLM	**G**eneralised **L**inear **M**odel
HMB	**H**eavy **M**enstrual **B**leeding
IS	**I**nformation **S**ystem
k-NN	**k**-**N**earest **N**eighbour
LH	**L**aparoscopic **H**ysterectomy
LPM	**L**ocal **P**rocess **M**odel
MMC	**M**axima **M**edisch **C**entrum
PC	**P**atient **C**haracteristics
PF	**P**rocess **F**eatures
RF	**R**andom **F**orest
ROC	**R**eceiver **O**perating **C**haracteristic
RVF	**R**etro**v**ersie**f**lexie: retroverted uterus
SMOTE	**S**ynthetic **M**inority **O**versampling **Te**chnique
SVM	**S**upport **V**ector **M**achine
TED	**T**otale **E**dometrium **D**ikte: Total endomtrial thickness
VUE	**V**aginal **U**terus **E**xtirpation
XGBoost	E**x**treme **G**ratient **B**oosting

Chapter 1

Introduction

Over the years, data science technologies have brought new opportunities which enabled breakthroughs (Dhar, 2013), also in the medical domain. By using the medical data of many patients, knowledge can be extracted to improve people's health (Abedjan et al., 2019). This knowledge is addressed from different perspectives, like the descriptive, diagnostic, predictive, and prescriptive perspective. Where the descriptive and diagnostic perspective focus on the declaring and explaining parts of the current process, the predictive and prescriptive perspectives focus more on the future of the process, intending to predict the outcome and describe what is going to happen.

Advances in machine learning have created improving possibilities in patient-level prediction. The ongoing development in machine learning offers the potential to move beyond average treatment and to take personal features into clinical decision-making (Reps et al., 2018). Patient-level decision making can be applied to patients undergoing a surgery with the chance of requiring a reintervention. One of these trajectories to apply patient-level decision making is on patients who choose to undergo a Novasure surgery, with different result outcomes.

1.1 NovaSure

Nowadays, a women's menstruation cycle is still seen as taboo. Despite the fact that it is not a subject spoken about very loud and maybe therefore not many people are aware of, approximately 30% of women are affected in their daily life by heavy menstrual bleeding (HMB) or abnormal uterine bleeding (AUB) (Fraser et al., 2015). HMB is defined either as losing at least 80 millilitres of blood per cycle (Hallberg et al., 1966), or as "excessive menstrual blood loss leading to interference with the physical, emotional, social, and material quality of life of a woman"[1]. In other terms, women are affected in their energy level, mood, work productivity, social interaction, family life, and sexual functioning due to their menstruation cycle (Fraser et al., 2015; Matteson et al., 2009; Lukes et al., 2012; Liu et al., 2007; Fraser, Langham, and Uhl-Hochgraeber, 2009; Coulter, Peto, and Jenkinson, 1994; Warner et al., 2001). The definition of AUB depends on the age of the woman. For women of childbearing age, any change in menstrual period frequency or duration, or amount of flow, as well as bleeding in between cycles is seen as AUB (Livingstone and Fraser, 2002). For postmenopausal women, AUB is defined as vaginal bleeding 12 months or more after the cessation of menses, or when suffering from unpredictable bleeding (Lethaby et al., 2004).

[1] https://www.nice.org.uk/guidance/ng88

While hysterectomy remains the definitive surgical treatment for HMB and AUB, the Novasure surgery can be a uterine sparing alternative. This is a minimally invasive procedure that aims to destroy or remove endometrial tissue and can be used when intracavitary abnormalities are absent (Beelen et al., 2019). Other advantages of endometrial ablation compared to hysterectomy are that the surgery is less invasive, it does not require hysteroscopic visualisation, it can be performed more often under local rather than general anaesthesia, and it has a shorter recovery period (Rodriguez et al., 2019). The Novasure consists of a single-use device and a radio-frequency controller (Bongers, 2007). It includes a conformable bipolar electrode array which is placed on an expandable frame that can create a confluent lesion on the entire interior surface area of the uterine cavity. The device is inserted transcervically into the uterine cavity, and the sheath is retracted to allow the bipolar electrode array to be deployed and conform to the uterine cavity. For 120 seconds at most, an automatically calculated optimal power with a maximum of 180 W is applied to the inside of the uterus. Tissue is vaporised and/or coagulated during this process, hopefully resolving HMB complaints[2]. A thicker endometrial layer results in increasing treatment time of up to 120 seconds at most. The treatment is finished when all of the endometrial tissue is vaporised or when the ablation duration of 120 seconds is reached.

Multiple retrospective population-based trials and randomised control trials have shown that success rates are not 100%. Percentages ranging from 10% to 34% were found as the number of patients requiring further surgery after endometrial ablation, due to persistent or returned complaints after the Novasure (Bongers, 2015; Daniels, 2013; Smith, Malick, and Clark, 2014; Penninx et al., 2011). These surgeries can include another endometrial ablation or hysterectomy.

In the medical domain, research on prognostic factors associated with the failure of the Novasure surgery has been carried out. Bongers (2015) evaluated the factors sterilisation, age, BMI, cavity length, uterine fibroids, smoking, parity, and preoperative dysmenorrhoea. The main conclusion was that the chance of failure decreases as the age of the patient increases and that having dysmenorrhoea contributes negatively to the Novasure surgery outcome. Beelen et al. (2019) provided a systematic literature review on 990 studies carried out from 1998 until February 2019 and concluded is that the following factors are associated with the failure of endometrial ablation: a younger age, prior tubal ligation, and preexisting dysmenorrhea. More research is needed to conclude whether obesity and the presence of large sub-mucous uterine fibroids may be associated with failure.

1.2 Data science for patient-level decision making

Given the relatively high reintervention rate of the Novasure surgery (i.e., 10%-34%), there is a desire to improve the expectations of patients on the effect the Novasure surgery has on their uterus and complaint reduction. Ultimately, the gynaecologist can provide the patient with the right expectations. By comparing the current patient to previous cases, the outcome of the surgery can be predicted. This creates better expectations for both the patient and the gynaecologist department. Therefore, research is done on which data science methods give the best result in predicting the surgery outcome. For this prediction, prognostic factors defined by previous literature which may contribute to the Novasure surgery succeeding or failing are

[2]From NovaSure® Instructions for Use and Controller Operator's Manual: https://www.hologic.com/hologic-products/gynecologic-health/novasure-endometrial-ablation

used. Later in the research, appointment and care activities prior to the surgery are included.

Data science is a broad term for methodology that utilises algorithms for prediction typically with large, complex datasets involving an extensive number of variables (Suchting et al., 2018). Within data science, there are different techniques and technologies which can be used to analyse cases based on similar previous cases. For example, in classification a class label is predicted for the analysis object and in clustering, analysis objects are placed in a group to which the set of objects is more similar than to another group. Decision trees can be used to discover features and extract patterns from data, in order of predictive modelling (Myles et al., 2004). For many years, machine learning is used for a variety of classification tasks and extended with deep learning. By using multiple computational layers that allow an algorithm to learn the appropriate predictive features on the basis of examples, instead of engineering features by hand, predictions become more accurate (Poplin et al., 2018). Progress in machine learning used on datasets has created opportunities in applying patient-level prediction. This offers potential for medical practice to let go of average treatment effects and consider personalised features and risks as part of clinical decision making (Reps et al., 2018).

Not only predictive machine learning, but also process monitoring techniques are interesting to use in healthcare processes. Van der Aalst (2016) defines process mining as the missing link between data-oriented analysis and model-based process analysis techniques. With process mining techniques, process models can be extracted from event data. Three techniques are differentiated. The first, process discovery, aims to automatically construct a model based on observed events. The second is checking conformance. The modelled behaviour and the observed behaviour are compared to each other. Last, the discovered process model can be extended or improved using the event log, using enhancement (Van der Aalst, 2012). Process mining has been used in healthcare before. While the healthcare domain is known for its complexity, another characteristic is that many autonomous, independently developed applications are found (Lenz et al., 2002), instead of an integration between the applications. Mans et al. (2015) uses process mining in order to find out what happens in a healthcare process for a group of patients with the same diagnosis and paths. The focus level is on paths followed by individual patients.

In predictive process monitoring, finished cases in the process are analysed and based on these case outcomes, predictions are made about the future state of the ongoing cases (Teinemaa et al., 2019). Verenich et al. (2019) define predictive (business) process monitoring as "*a family of online process monitoring techniques that seek to predict the future state or properties of ongoing cases of a process based on models extracted from historical (completed) cases recorded in event logs*". With the focus on predictive monitoring of (categorical) case outcomes, in particular, it becomes outcome-oriented predictive process modelling. The techniques usually have an offline stage which exists of four actions: extracting prefixes, clustering or dividing the cases into buckets, encoding features, and training classifiers. Then during the online stage, the outcome of an ongoing case is predicted according to the developed model (Wang et al., 2019). According to Teinemaa et al. (2019) and Verenich et al. (2019), no unified approach is evaluated for predictive (business) process monitoring methods.

To my knowledge, little research has been done on combining predictive machine learning techniques and process monitoring techniques to predict surgery outcomes before, more specific, Novasure surgery. Previous work does include examining Novasure surgery outcomes utilising one machine learning algorithm for prediction. This research (Stevens et al., 2021) compares one machine learning algorithm to

logistic regression and focuses on static patient characteristics from electronic health record (EHR) data, without taking notice of process features. Therefore, we wish to research which features patients have influence on the outcome of the Novasure surgery, using predictive machine learning and thereby include process features. In this thesis, a database including 1029 patients who have undergone a Novasure surgery at the Maxima Medisch Centrum (MMC) is analysed.

1.3 Research questions

The goal of this research is to investigate the use of historical data of Novasure patients to provide evidence-based insights into current treatments and their impacts on the Novasure surgery outcome per patient. That results in the following main research question:

> **MRQ** Given patient characteristics and process features, which machine learning algorithm(s) can predict the outcome of Novasure surgery with highest accuracy?

In order to answer the main research question, first the accuracy of the prognostic features from literature are explored. While prognostic factors include patient-specific variables and no process-specific information, these are called patient characteristics. The first sub question addresses these features.

> **RQ1** To what extent do patient characteristics have an influence on the outcome of a Novasure surgery?

With this question, the goal is to find out which patient characteristics at first seem to have the heaviest influence on the outcome of the surgery. Combined with this knowledge, an exploration is done on whether certain process patterns contribute to the surgery success outcome. Process activities are added as input variables to the earlier done machine learning algorithms.

> **RQ2** How are predicting results of the Novasure success outcome influenced by including process features?

Then, to create the best-combined performing model, exploration is done on how different machine learning techniques perform with respect to the other machine learning techniques. Previous studies have shown the potential of using predictive machine learning (Suchting et al., 2018; Stevens et al., 2021) and process monitoring techniques to analyse healthcare processes (Mans et al., 2015; Teinemaa et al., 2019; Verenich et al., 2019; Pijnenborg et al., 2021) on the patient-level decision making. Based on their findings, predictive machine learning techniques are chosen to be compared. This study aims to improve the analysis of these predictive machine learning techniques, by including more techniques than done in previous studies.

> **RQ3** How do the following predictive machine learning techniques perform compared to each other?
>
> - Decision trees
> - Random forests
> - Logistic regression
> - Gradient boosting
> - Neural networks
> - K-nearest neighbour

1.4 Contributions

With this research, we want to deliver an empirical and a methodological contribution. While analysing a dataset in retrospect, my goal is to provide insights on which patient characteristics influence the Novasure surgery. Also, we want to contribute to the knowledge on which machine learning algorithms are suitable to use for patient-level predicting.

1.5 Outline

The structure of the proposal is as follows. Chapter 2 provides background information on predictive modelling techniques. Also, the use of this technique in healthcare is presented in this chapter. Chapter 3 is on the approach of the following research. With the help of cross industry standard process for data mining (CRISP-DM) usage in other domains, the use of CRISP-DM in this research is formulated. The data on which this research is applied is explained in chapter 4. An overview of the dataset is given, the patient flow is described, the data processing is presented, as well as the data analytic strategy. Chapter 5 presents the results from the four experiments, with its discussion in chapter 6. This thesis is concluded in chapter 7.

Chapter 2

Predictive Machine Learning and Process Monitoring techniques in Healthcare

This chapter includes the related work on predictive machine learning and process monitoring techniques. First, general information on these subjects is provided, after which relevant researches on applying both techniques in the medical sector and on patient-level are described.

2.1 Predictive machine learning

Kuhn, Johnson, et al. (2013) define predictive modelling as "*the process of developing a mathematical tool or model that generates an accurate prediction*". Krause, Perer, and Bertini (2016) then state that often, data scientists turn to machine learning. Machine learning is used to solve "big data" problems across a wide range of phenomena. Big data problems are able to include complex data sets involving an extensive number of variables (Suchting et al., 2018). With machine learning, it is possible to create predictive models based on data with ground truth, which is automatically learnt into useful information. The machine learning tools are connected with providing programs with the ability to learn and adopt. The field of machine learning has been branched into several sub fields, standing for different types of learning tasks. Shalev-Shwartz and Ben-David (2014) describe four parameters along which learning paradigms can be classified. For each of the parameters, this research is classified under the explanation, in order to create some perspective.

- **Supervised versus Unsupervised** In supervised learning, the training data contains extra information. The outcome of the the sample is known, called the label. In unsupervised learning, both the training and the test data have no label containing information about the outcome.

  For each of the patients in the dataset, the outcome variable is known: the patient had undergone a reintervention or not. Therefore, this dataset includes supervised data.

- **Active versus Passive Learners** In active learning, the learner interacts with the environment at learning time. During passive learning, the learner only observes information provided by the environment.

  The learning performed during this research is passive, since the data includes the environment as it was between 2008 and 2018 and no further interaction in the form of posing queries or performing experiments is present.

- **Helpfulness of the Teacher** In science, the environment takes the form of the teacher, which can be best thought of as passive. Passive learning scenarios are modelled by postulating training data that is generated by random processes.

  The helpfulness of the teacher is no binary variable, so this research cannot be put in one of the two classes. The learning is more passive than active, since no adversarial teacher is present. Like in most scientific cases, training data is observed.

- **Online versus Batch Learning Protocol** With the use of the batch learning protocol, the learner is set to respond throughout the learning process. During batch learning protocol, the learner is set to engaging the acquired expertise only after having a chance to process large amounts of data.

  The data used in this research is gathered before the research had started. This forms a batch of training data with which experiments can be done before delivering conclusions as output.

Excluding decisions tree, most machine learning techniques are black boxes and chosen for their performance metrics, such as high accuracy scores (Krause, Perer, and Bertini, 2016). In this section seven machine learning techniques are explained with the focus on predicting future outcomes of unfinished cases.

2.1.1 Decision tree

For a long time, decision tree (DT) was the most popular machine learning technique (Myles et al., 2004), mainly caused by their intuitively simple classifier. A decision tree is a recursive split of input data, based on a value belonging to a certain class or the value being higher or lower than a certain threshold (Maxwell, Warner, and Fang, 2018). The pattern of repeated splits is formed by branches representing a path through splits and the leaves forming the ultimate target values. In the case of classification, the leafs represent the classes. The method can be used for both classification and feature selection, and feature reduction purposes (Borak, 1999; Pal and Mather, 2003). Decision trees have several advantages. The trees are able to work with data represented on different measurement scales and visualisation is possible using a set of if-then rules. Also, once the total model is developed, fast classification is possible (Gahegan and West, 2001). Decision tree is added to this research's method mainly to give a first exploration, in the form of visualisation, of how the feature compare to each other. Disadvantages are that it is possible to find a non-optimal solution and there is a chance of overfitting. Overfitting can be prevented by pruning the tree, resulting in a decrease in classified data accuracy, and an increase in dealing with the accuracy of the unknowns.

2.1.2 Random forest

While creating a decision tree can still be successful, an improved version has made its entrance: random forests (RF). RF is a classifier consisting of a collection of decision trees to overcome the weaknesses of decision trees. Each tree is constructed by applying an algorithm A on a training set S, with adding an independent and identically distributed sampled vector θ (Shalev-Shwartz and Ben-David, 2014) and each tree is trained with a randomly generated subset of the training data. By using the majority 'vote' of all trees, the final class for the unknowns is assigned. Using this

global optimum, the non-optimality weakness of the decision tree is solved and significant improvements in classification accuracy are created (Breiman, 2001). Each tree individually gets less accurate due to the reduced training data and the reduced number of variables, but the trees are also less correlated, causing the ensemble to be more reliable (Maxwell, Warner, and Fang, 2018). No individual pruning is needed anymore. One disadvantage is that by creating a forest, the ability to visualise the model is lost. Random forest can take the same input as decision tree and create higher accuracy in classification and regression. Therefore, the technique is added to the method.

2.1.3 Support vector machines

Support vector machine (SVM) is developed for binary classification problems (Cortes and Vapnik, 1995). Linear predictions are learnt in high dimensional feature spaces, resulting in high sample complexity. To overcome this high sample complexity, optimal boundaries with the support vectors being maximally separate are searched (Myles et al., 2004; Maxwell, Warner, and Fang, 2018). This causes the data points to be separated far from the hyperplane, leading the sample complexity to decrease. The SVM classifier is a binary classifier that can only identify a single boundary between two classes. To overcome this constraint, the classifier is repeatedly applied to multiple combinations of classes, leading to an exponentially increased processing time, but also creating the possibility to find more boundaries. As already mentioned, SVM are originally designed for linear class boundaries. Using the kernel trick, this restriction is bypassed. Under the assumption that linear boundaries do exist in higher-dimension feature spaces, the projection of feature space is set to a higher dimension. There is a difference between hard-SVM and soft-SVM. When it is not possible to completely separate the classes with the largest possible margin, like in a hard-SVM, the decision boundary becomes a soft margin. Parts of the training classes are allowed to be on the wrong side of the decision boundary and are supplied with a cost C, creating a soft-SVM. The higher the cost gets, the more complex the decision boundary gets and the lower the ability to generalise becomes. Due to poor performance in earlier research on patient outcome prediction (Teinemaa et al., 2019) and the main focus on image classification and image retrieval (Chapelle, Haffner, and Vapnik, 1999; Mercier and Lennon, 2003; Hong, Tian, and Huang, 2000; Foody and Mathur, 2004), SVM are not included in the method of this research.

2.1.4 Logistic regression

Logistic regression is a technique used for classification tasks, partly due to its following characteristics. In logistic regression, a family of functions is learnt on the interval $[0, 1]$, with the goal to predict the probability that the label of a case is 0 or 1 (Shalev-Shwartz and Ben-David, 2014). In the field of gynaecology, many prediction models are developed utilising logistic regression (Stevens et al., 2021). One advantage is that logistic regression can take both continuous data and discrete data as input. For this specific research it means that, for example, age and BMI as continuous data can be taken into account, as well as discrete data like whether the patient has had a cesarean section or suffers from endometriosis. Also, logistic regression brings the possibility to identify whether a variable is useful for predicting the outcome by testing whether the variable's effect on predicting the outcome is significantly different from 0. This contributes to finding which features are important to take into

account and which features are less important. However, the method cannot automatically estimate the interconnection between these features, which can result in overestimating the influence of an individual feature. Still, it is useful to include logistic regression in this research, since it has been used many times in gynaecology and therefore can function as an anchor to compare to other methods with.

2.1.5 Gradient boosting

The main idea of gradient boosting is that by adding new models, to the ensemble sequentially, the trained model gets more accurate. With each iteration, a weak, base learner (function) is added to the main gradient boosting model (GBM), and the model gets trained with respect to the error of the whole ensemble so far (Natekin and Knoll, 2013). Most of the time, these base learners are tree models, which means that their input data can be both categorical and continuous and GBMs are suitable for classification. The main goal is to provide a more accurate estimate of the response variable. It rests on the principle that the algorithm constructs new base learners which are maximally correlated with the negative gradient of the loss function associated with the whole ensemble. This loss function can be arbitrary, but it can also be a classic square loss, for example. This would give a better intuition. The researcher has great freedom in choosing the loss function, which brings high flexibility, and thereby makes the technique a highly customisable to fit any data-driven task. Successes are considerably high in not only practical applications but also in machine learning and data mining challenges. Gradient boosting came out as the best technique in sort-like researches (Teinemaa et al., 2019; Pijnenborg et al., 2021) and is therefore added.

2.1.6 Neural networks

Inspired by the most complex organ in the body, neural networks are designed to carry out high complex computations. In simplified models, the human brain is presented as a large number of computing devices, called neurons, forming a complex communication network. Here, a neural network is conceptualised as a mathematical analogue of animal brain axons and interactions through synapses (Maxwell, Warner, and Fang, 2018). Simply said, a directed graph is presented, where nodes represent the neurons, and edges form the links between the neurons. Each neuron processes a weighted sum of outputs from the neuron directing to it. The network is organised into layers. A set of nodes can be decomposed into a union of disjoint subsets so that every edge connects one node to another node forming layers. The neural network is trained by randomly guessing values for the weights in the input of the neurons in the different layers iteratively while observing the effect of the outcomes. Each adjustment that improves the classification is kept and reinforced. Each adjustment that worsens the classification is discarded. The technique is improved by increasing the number of neurons in the hidden layer and adding more hidden layers, which together cause an increase in the potential for describing complex decision boundaries. Although, it did not perform the best compared to other methods (Suchting et al., 2018), neural networks are included in this research's method for its potential and ability to apply one or more non-linear layers.

2.1.7 K-Nearest Neighbour

The k-nearest neighbour (k-nn) classifier is different than the previously mentioned machine learning techniques. The technique is a local version of a univariate location estimator (Altman, 1992), based on the assumption that features are used to describe the domain points relevant to their label, in such a way that close-by points are alike (Shalev-Shwartz and Ben-David, 2014). Instead of producing a trained model, the training set is memorised and the label of each sample is predicted based on its direct neighbours from the original training data. The direct neighbours form the most common class of k training samples in the near feature space, all equally spread on either side of the point of estimation. The lower k is chosen, the more complex the decision boundary gets. Using a class with a high number of k training samples improves generalisation. Due to the fact that there is no model, the resources require to become greater when the amount of training samples increases (Maxwell, Warner, and Fang, 2018). K-nn is added to the method for its simplicity in finding a suitable class for each data point.

2.1.8 Inclusion of algorithms

For each of the machine learning techniques has been considered whether or not to include them in the research. Decision trees are added to this research's method mainly to give a first exploration, in the form of visualisation, of how the features compare to each other. Random forest can take the same input as decision tree and create higher accuracy in classification and regression. Support vector machines delivered poor performance in earlier research on patient prediction outcome (Teinemaa et al., 2019) and their main focus is on image classification and image retrieval (Chapelle, Haffner, and Vapnik, 1999). Due to its extensive use in the field of genealogy in developing prediction models (Stevens et al., 2021), logistic regression is added to the research method as baseline. In researches comparable to this one, gradient boosting came out as the best technique(Teinemaa et al., 2019; Pijnenborg et al., 2021), which is why the technique is added. Neural network is included for its potential and ability to apply one or more non-linear layers and k-nearest neighbour is added for its simplicity in finding a suitable class for each data point.

Table 2.1 compares the techniques on several points to each other. The table is based on the characteristics of different models discussed and found in the individual model sections. All techniques are suitable for classification, which was a criterion for finding techniques in the first place. Class prediction is the desired outcome of this research. Second, due to the variety in scales of the features, the technique should be able to handle both discrete and continuous input data at the same time. For the techniques not being able to handle both at the same time, categorical data had to be made binary or numerical. Finding the contribution which each feature delivers to the predicted outcome is wishful, but not obligated. The last characteristic treats the speed, performance, memory usage, and overall time taken for model training, based on Tatsat, Puri, and Lookabaugh (2020).

2.2 Use of predictive modelling techniques and process mining in healthcare

Concerning predictive machine learning on patient-level and outcome-oriented process monitoring, the researches worth mentioning are described in this section. The

TABLE 2.1: Comparison of machine learning techniques

Machine learning technique	Classification	Discrete and continuous input data	Training time	Used in method
Decision tree	✓	✓	✓	✓
Random forest	✓	✓		✓
Support vector machine	✓			
Logistic regression	✓	✓	✓	✓
Gradient boosting	✓	✓		✓
Neural network	✓			✓
K-nearest neighbour	✓		✓	✓

researches are divided based on their main method. First, researches with the focus on machine learning related outcome prediction are discussed, and then process mining related outcome prediction. This section ends with a discussion on how the papers differ from this research and what makes the researches worth mentioning.

2.2.1 Machine learning related outcome prediction on patient-level

Suchting et al. (2018) used four different machine learning techniques to predict whether or not patients would perform aggressive event against staff or other patients. To their knowledge, there is limited previous work examining patient aggression in mental health facilities utilising machine learning algorithms for prediction based on available EHR data. They provide a retrospective study utilising 29,841 EHRs from a psychiatric centre. Based on 328 predictors, including patient's full demographic profiles, vitals and comprehensive psycho social assessments, a prediction model was made to predict the one outcome measure: an aggressive event or not. They split the data in 80% training set and 20% test set and applied GLM, RF, GBM and Deep Neural Networks (DNN). Validation was done using 5-cross fold validation and the model performance is measured by the highest area under the receiver operating characteristic curve (AUC). GLM was the best performing technique with an AUC of 0.7794 on training data and an AUC of 0.7801 on the test data. These results surpass three previous machine learning efforts (Wu, Roy, and Stewart, 2010; Gowin et al., 2015; Passos et al., 2016) in behavioural sciences.

Currently, the study closest to this research is done by Stevens et al. (2021). Their goal was to develop a prediction model to predict surgical reintervention within two years after endometrial ablation. The retrospective cohort study analysed EHRs of 446 patients of the Catharina Hospital in Eindhoven and the Elkerliek Hospital in Helmond. The authors compared their earlier developed logistic regression model to RF, which they first trained on the patient characteristics age, duration of menstruation, dysmenorrhea, parity and previous cesarean section. Where their final logistic regression model achieved an AUC of 0.71 after correcting by the shrinkage factor, their RF model first achieved an AUC of 0.63 and 0.65 after optimising. This leads them to the conclusion that machine learning models do not perform better

than predicting with logistic regression models, but this difference is not significant due to overlapping confidence intervals.

2.2.2 Process mining related outcome prediction

Koorn et al. (2019) also took aggressive events under patients into account. They used a process mining approach to analyse patterns of aggressive behaviour using records from 1115 patients in Dutch residential care facilities over three years. They created an event log in which each case had a unique case identifier, an activity description and a timestamp and applied process mining using Disco. They found a clear distinction in cases using exclusively a certain type of aggression and cases containing a mix of aggressive behaviour types. Later, Koorn et al. (2020) continued this research in proposing and formalising a technique to discover action-response-effect patterns. They tried to provide processes to organisations that are appropriately represented and effectively filtered to show meaningful relations. In this paper a causal mining algorithm is presented which can be used as a discovery technique. It includes statistical tests to uncover potential dependency relations between responses and their effects on the cases. The technique can be used to support decision making process. Defining this in the medical domain, it can help with decisions considering multiple treatment options. After that, Koorn et al. (2022) proposed a novel and generic process mining approach with the goal of producing insights into statistical relations in patient pathways. From an event log, they created a state action log on which statistical tests were used to discover significant relations. After that, a graph was created with the help of a heuristic selection miner in order to visualise uncovered statistical relations from the analysis step. While validating their novel approach, they came to the conclusions that domain knowledge is required to define actions and states and that their approach cannot confirm that relations are causal relations.

Teinemaa et al. (2019) had as goal to train a model which could accurately and efficiently predict outcomes given a prefix only. They did that by comparing the performance of different outcome-oriented monitoring methods for business processes. The models were trained on a given event log of completed business process execution cases with a final outcome class. Then, the prefix traces in the historical log were divided into several buckets and different classifiers were trained for each bucket. These classifiers were single bucketing, k-nearest neighbour, state, clustering, prefix length and domain knowledge. At run-time, the most suitable bucket for the ongoing case was determined. Depending on the bucket, decision tree, random forest, gradient boosting model and support vector machine were utilised, after which the sequence of classifiers was encoded. Mentionable results are that XGBoosting performed best wist the highest AUC in 15 of the 24 cases and the highest f1 score in 11 cases. Support vector machine did in general not meet the same level of accuracy as the other used methods.

Also in the medical field, Pijnenborg et al. (2021) investigated the application of process mining techniques on palliative care pathways to obtain an evidence-based understanding of which palliative treatments are commonly carried out and how they are associated with the patients' survival time. Using an event log of completed treatments and the patients' survival time, they trained a model that can predict the life expectancy of patients currently under treatment. In the offline phase they extracted prefixes and created buckets, after which sequence encoding techniques were applied, like in Teinemaa et al. (2019). The prediction models used were random forest, gradient boosting model and decision tree. Then, they used a local process

model (LPM) miner (ProM 6.9 framework)[1] to extract common practices of the palliative process and used the earlier found metrics to rank the obtained models. In 14 of the 23 cases, extreme gradient boosting performed best, with once reaching a f1 score of 0.87 and an accuracy of 0.8.

2.2.3 Medical specific domain

One main difference between the in this section described researches and the research which is going to be carried out, is how data science techniques are used. For example Suchting et al. (2018), who predicted whether or not patients would undergo an aggressive event based on the patients' available EHR data including predictors like their demographic profile, vitals and comprehensive psycho social assessments. The results included outcome events. In contrast to this research, they do not use event data which could have revealed triggers or patterns through time.

The main focus of Koorn et al. (2019), Koorn et al. (2020), and Koorn et al. (2022) and Pijnenborg et al. (2021) analysed and found patterns, based on event data. One factor left unpractised is the use of personal patient features. The influence of these features is not researched in finding patterns. Where Suchting et al. (2018) purely focused on patient features, and where Koorn et al. (2019), Koorn et al. (2020), and Koorn et al. (2022) and Pijnenborg et al. (2021) only used process features, we are combining these two to create more comprehensive models.

When it comes to comparing machine learning techniques on patient-level outcome prediction, only four machine learning techniques are compared at most in earlier research (Suchting et al., 2018; Teinemaa et al., 2019; Pijnenborg et al., 2021).

The research done by Stevens et al. (2021), the one most alike to this research, differs on multiple characteristics. The differences which are likely to lead to a different conclusion are:

- The follow-up period of Stevens et al. (2021) is ended on the day of hysterectomy, in case of death or on April 15, 2015, whereas we keep the follow-up period on three years after the Novasure surgery.
- Stevens et al. (2021) compare one machine learning technique, RF, to their logistic regression model. In this research multiple machine learning techniques are compared to each other, and combined with process mining.
- The amount of patients (446 vs. 1029) and the amount of variables (5 vs. 18) differ significantly. Machine learning is known to work better with more data and variables (Myles et al., 2004; Suchting et al., 2018; Maxwell, Warner, and Fang, 2018).
- While Stevens et al. (2021) only look at patient characteristics, this research also included appointments and care activities in the two years prior to the Novasure surgery. The total amount of features taken into account increases to 276 features with these appointments and care activities.

Also, they concluded that their results were not significant due to overlapping confidence intervals. By comparing six, instead of two or four, data science techniques, including process features and by using a relatively wide dataset, the expectation is to create better performing prediction models with, for example, a higher AUC.

[1] https://www.promtools.org/doku.php

Chapter 3

Research method

The methodology used in this study is based on CRISP-DM. CRISP-DM is originally developed to help translate business problems into data mining tasks, with the aim of making these large data mining projects less costly, more reliable, more repeatable, more manageable and faster (Wirth and Hipp, 2000). This research is not a data science project with the focus on making the business more efficient in terms of time and money. Therefore, two adjustments on the CRISP-DM lifecycle are adopted. In this chapter, first, a look is taken at how CRISP-DM is used in other domains. Then, the adjustments done to create a fitting CRISP-DM lifecycle for this research are discussed in section 3.2. This chapter concludes with the CRISP-DM lifecycle as it is used in this research.

3.1 CRISP-DM in other domains

In this section, we briefly discuss the CRISP-DM methodology and its application in other domains, to justify the use of CRISP-DM for this research project. CRISP-DM can be used to extract useful knowledge from data by systematically following a process with reasonably well-defined stages (Provost and Fawcett, 2013). The six stages of the life cycle include business understanding, data understanding, data preparation, modelling, evaluation and deployment. Although CRISP-DM is presented as a life cycle, the sequence of the stages is not rigid and moving back and forth between the phases is always required (Chapman et al., 2000). CRISP-DM is originally designed to address business problems, but other fields have adopted the method and applied extensions and adjustments make the method fit their specific field of research:

- Venter, de Waal, and Willers (2007) argue that forensic analysis can benefit from research knowledge in discovery that data mining and adjusted the life cycle so that crimes can be "re-enacted" by analysing electronic evidence left behind by subject's actions.
- Niaksu (2015) created the CRISP-MED-DM, which addresses specific challenges of data mining in the medical domain, by introducing generic and specialised tasks to the original CRISP-DM cycle to resolve five well-known challenges in medical data mining.
- Huber et al. (2019) extended CRISP-DM with Data Mining Methodology for Engineering applications (DMME) to provide communication and planning foundation for data production within the production domain.
- Cazacu and Titan (2021) used CRISP-DM as a method to standardise analysing large volumes of unstructured data and thereby generate analytical insights for well-being and social science topics.

These examples show that a method which is originally created with the aim of making large data mining projects less costly, more reliable and more efficient by helping translate business problems into data mining task can be used in other fields. Again, this research has the aim of improving a medical problem using data mining tasks, and not making a process more efficient and less costly. Some changes need to be adjusted to the cycle to make it fit this master thesis project.

3.2 CRISP-DM for data science research

This research has the aim of improving a medical problem using data mining tasks, so that efficiency increases through correcter predictions, which slightly differs from the main goal of the original CRISP-DM. Therefore, the cycle should be slightly adjusted. Two main changes, which include the business understanding and evaluation phases, are discussed.

3.2.1 Domain Understanding

According to Martínez-Plumed et al. (2019), data science spans both the industry and academia, because in both domains value is extracted from data using scientific methods like machine learning. The emphasis is on solving the domain-specific problem in a data-driven way. In this case, the medical related problem is addressed using data science techniques, like machine learning and process mining. They also mention academic discovery being usually question-driven, rather than data-driven or goal-driven as a difference between the industry and academia. In this research, to get to the answer of the research questions, the focus is on a provided database. Still, one adjustment is done to make the CRISP-DM applicable to this scientific research includes the first phase of the life cycle. The adjustment addresses the aim of understanding of the process of carrying out a research, instead of business understanding. The first phase is transferred into a more fitting understanding: *Domain understanding*. Venter, de Waal, and Willers (2007) use the word case, because each evidence mining project is associated with a specific case. In this research project it is important to understand the specific domain, the gynaecology domain in which Novasure surgeries are carried out. Medically, the Novasure surgery and the process around it need to be well-understood.

3.2.2 Evaluation

Within process mining methodologies, there are several points where the data analyst, who is carrying out the project, and the domain specialist, who is in need of insights, have a moment of contact. First, at the start of the project. The domain specialist and the data analyst discuss the problem, the goals and the desired outcome. Then, during the data extraction and pre-processing phase, the data analyst might have questions about the data. At the end of the project, the data analyst and the domain specialist evaluate the project and its results, in order to make findings valid. It is important to create results with actual organisational value.

According to Koorn et al. (2021), current process mining methodologies fall short in providing evaluation on the findings of a research and this shortage is also present in data science (Martinez, Viles, and Olaizola, 2021). Untill now, qualitative evaluation methods used are case studies (Thomas, 2006; Kaufmann et al., 2017), interviews

with experts (Baijens, Helms, and Iren, 2020), focus groups and undefined discussions (Koorn et al., 2021), but this concludes to no structured evaluation method which can be turned into actionable insights and recommendations is provided.

In quantitative research, the focus is on verifying the effectiveness of the proposed techniques, using established metrics. Koorn et al. (2021) observed 80 papers about projects with organisational partners being included in the evaluation at the end of the projects in-depth. Based on these observations, six validation strategies are proposed from the qualitative research perspective. Lewis (2015) recommends that one should always follow two validation strategies at the minimum to reach a sufficient accepting level of validation. To bridge the gap between the industry-focused CRISP-DM and academic research requirements, four of these evaluation strategies proposed by Koorn et al. (2021) are present in the additional evaluation phase of this research project and are carried as follows:

- **Engagement and understanding of the field** Engagement and understanding are achieved by visiting the hospital and observing different gynaecology surgeries. Also, meetings with domain experts are held regularly to see if no misinformation slips in the data and to align the interpretation of results.
- **Triangulation** Adequate triangulation results in completeness and consistency of results. There are two ways to achieve this: using multiple data sources and applying a mixed method approach to the data. In this research multiple methods are applied to the data, see section 2.1. Valuable information is retrieved by the results provided and then cross-validate that the outcomes provided are adequate.
- **Peer review and external audit** During the research presentations are given to domain experts from both the gynaecology field and the machine learning field. Machine learning expert peer review the methods which are used in this research and give feedback on the whether the data is handled in a compliant manner. Experts from the medical domain can reflect on the credibility of the results obtained and on whether they are likely to be true. Afterwards, medical experts are also asked to predict whether a patient is likely to need a reintervention, based on patient features, process features and human knowledge, to validate whether machine predicting is more accurate than human predicting.
- **Clarify biases** Biases slip in each research, especially when process mining is combined with another domain and influences the quality of the outcomes. Where the process analyst has no expert knowledge on the domain, which can result in difficulties when in determining the causes of unexpected analysis results (Van Eck et al., 2015), the domain experts miss expertise in interpreting process mining results. A researcher should always be transparent about the results towards the expert and discuss the reliability with the expert. Therefore, the possible biases from a qualitative perspective are reflected and described at the end of the research project.

3.3 Research lifecycle

With the previously described adjustments, the lifecycle results in the following. The visualisation can be found in figure 3.1. For each of the six stages, a brief description is given and the implementation of that stage in this research is explained.

FIGURE 3.1: CRISP-DM with adjusted stages

3.3.1 Domain understanding

In order to gain more domain knowledge on the field of HMB and the medical interventions applied, several steps are taken. I was present at different gynaecology surgeries being carried out, including the Novasure surgery, and I attended information evenings of the gynaecology department. Then, together with the hospital and with the use of provided background literature, the research problem is defined. Literature research is carried out to obtain knowledge on the research gap.

3.3.2 Data understanding

Understanding the data includes studying the dataset provided. To gain full understanding, unclarities are discussed with a professional. While going back and forth between the data understanding and the data preparation phases, a table describing the data and its values (table 4.1) is created, with the main goal of summarising the dataset. Here, a clear distinction between predictive features and process-aware features is discovered.

3.3.3 Data preparation

During the data preparation phase, the initial raw data is transformed into a dataset that can be used for modelling. The dataset, which is provided by the hospital, undergoes multiple iterations in which the data is cleaned, missing values are extracted from the rest of the dataset, and outliers are removed. Then, from the most important columns a new, smaller dataset is created to work with more easily. A detailed explanation of this can be found in section 4.4.1 and section 4.4.2.

3.3.4 Modelling

In this phase, six modelling techniques are selected and applied, and their parameters are calibrated to optimal values. This is done in two sub phases, since there are two different processes to focus on. During the first sub phase, the focus is on patient characteristics. During the second sub phase, also appointments and care activities per patients are involved. For both phases, the following data science algorithms are used; decision trees, random forests, logistic regression, gradient boosting, neural networks and k-nn. A more detailed description is provided in sub section 4.6. During the second sub phase the activity features are also taken into account and the modelling focus is shifted to process mining. The modelling phase is going to be carried out in three different, time-dependent iterations.

3.3.5 Evaluation

The evaluation objective is to validate and compare the applied machine learning techniques used in the modelling phase. Like described in 3.2.2, this is achieved by visiting the hospital to have meetings with experts, cross validation, peer review, and with a description of all possible biases. Also, validation is done by experts in the form of a look-over to see whether no remarkable results have occurred.

3.3.6 Deployment

Deployment is out of the scope of this project, mainly due to time constraints. In the original CRISP-DM, the last phase is deployment, meaning that the knowledge gained is organised and presented so that the customer can use it. Often, this includes applying “live" models within the organisation’s decision-making process. The focus of this research is on answering the research questions. Building a live model or dashboard which experts could use to make patient-level decisions, would take too much time and is beyond my abilities.

Chapter 4

Data description

This chapter describes the data which is used in this case to find out whether the combination of predictive modelling and process mining provides more valuable results. First, a brief description of the data and its origin is given, after which the flow the patients can undergo is described. Then, extra information on the groups of variables is provided. The chapter finishes with a description of how the data is processed, the analytic strategy applied, and concludes with a presentation of the features used.

4.1 Data overview

For this retrospective research, a dataset provided by the gynaecology and obstetrics department of the MMC is provided. It covers information about all patients who have undergone a Novasure surgery between January 1^{st}, 2008 and December 31^{st}, 2018 at the MMC in Eindhoven and Veldhoven, except for when they had undergone a previous Novasure surgery before this period. This data is extended with all care activities belonging to the patient from two years before the Novasure surgery, stored in different tabs. The total number of patients in the data set is 1039 patients, of which 10 are outliers, resulting in 1029 useful patients. At the moment of writing, the MMC is one of the 94 healthcare institutions providing a Novasure in the Netherlands[1]. The patients received usual care and did not have to follow any additional procedures. No additional permission is requested for the use of this medical data because this research is part of a retrospective nWMO research with a general purpose in which:

- the effort required to request consent is disproportionate to the purpose it serves (>1000 women);
- at the moment the pressure on healthcare is so great that asking for permission is considered disproportionate (Ministry of Health, Welfare and Sport);
- patients have been informed that data can be used for research (via the patient folder and website)
- there was the possibility to lodge an objection if the patient did not want to share its data (objection procedure or opt-out procedure); and
- we the preconditions are met: purpose limitation, data minimisation (no more than necessary for the question) and careful handling of the datasets (coded and within the MMC and UU environment).

[1] https://www.hevigbloedverlies.nl/zoek-een-ziekenhuis

4.2 Flow description

To get a clearer view of the process, figure 4.1 is created. It describes the path that patients can undergo. Patients with different complaints (*cycle disorder*, *benign adnexal abnormality*, or *uterine fibroids*) are referred to the gynaecologist. Since the inclusion criterion of the dataset is that the patient has undergone a Novasure surgery's, all 1029 patients visit the next phase, the Novasure surgery. After the Novasure surgery has taken place, there are two possibilities; (1) the patient is satisfied during the three up-following years and no reintervention takes place, or (2) the patient still has complaints and undergoes a reintervention of some sort in the three up-following years. Section 4.3 goes deeper into the patient's preoperative and perioperative features and explains the different categories of reintervention.

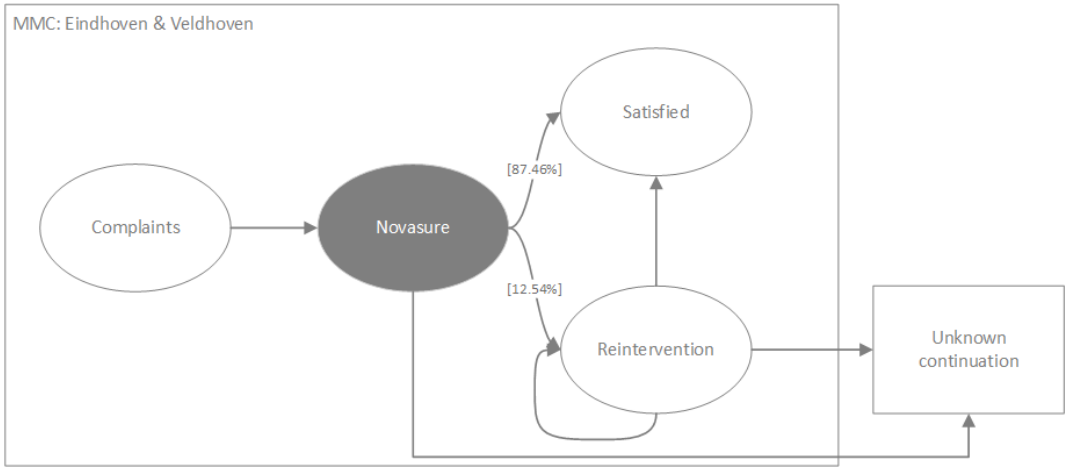


FIGURE 4.1: Graphical representation of patient flow

4.3 Variables

The dataset is divided into different variables; predictor variables and outcome variable, where the predictor variables are further categorised. Predictor variables exist of pre-existent patient characteristics, activities in the process towards the Novasure surgery, and occurrences during the surgery. The outcome variable includes whether a patient has undergone a reintervention or not.

4.3.1 Predictor variables

Predictor variables are the features occurring before and during the Novasure surgery, which are highly associated with having an influence on the outcome of the surgery (Beelen et al., 2019). They are divided into three categories:

- **Patient features**: include the features which apply solely to the patient. They are asked and researched prior to the surgery, with the aim of finding out whether the patient meets all requirements to undergo the Novasure surgery and apply all to the patient's body.
- **Process features**: include the care activities and surgeries at the same hospital which took place between two years before the Novasure surgery, and the waiting time between the surgery request and Novasure surgery in case it takes place in the operation theatre.

- **Perioperative features**: include the variables which occur during the Novasure surgery. They include the form of anaesthesia applied to the patient and the power and endurance the Novasure machine delivers during the surgery.

4.3.2 Outcome variable

The outcome variable is whether the patient undergoes a reintervention within three years after the Novasure surgery. When the patient is satisfied, no reintervention takes place. If the patient keeps suffering from HMB or complaints start reappearing, different interventions belong to the solution options. An overview of the data before specifying certain categories and all possible reintervention types can be found in Appendix A. Table 4.1 presents a summary of both the predictor variables and outcome variables, including the values that the variables can take and the amount of missing values in case there were.

4.4 Data processing

The dataset provided by the gynaecology department is not ready to put in a model. According to Bose, Mans, and Van der Aalst (2013), there are four categories of problems that contribute to data quality: missing data, incorrect data, imprecise data and irrelevant data. We do several iterations to make an input-ready dataset and discuss them in the following sections.

4.4.1 Patient selection

Before the dataset was provided, the domain expert had taken an extensive look at the dataset. In her opinion, it includes ten outliers. The reasons why these patients are outliers differ per patient and can include technical difficulties, the presence of a septum, perforation while expanding Novasure, passing out, not passing the cavity assessment, and a registration error. The patients seem to have undergone a reintervention, based on the dataset, but in reality, their first Novasure operation was not carried out. When retrieving their EHR another time, the string format did not match the original dataset anymore. These datapoints are imprecise and this makes the patients inappropriate for this research. Therefore, the datapoints are removed from the dataset which includes 1029 patients to analyse.

4.4.2 Feature selection and subtraction

The database started with 356 different features, one being more important than the other. Based on literature study (Bongers, 2007; Beelen et al., 2019) and on expert opinion, the most valuable and probable predictive features are chosen. For each column in the excel dataset, the meaning of the variable name is validated by the expert and the possible range of outcomes of the variable is investigated. After this investigation, the consideration is made whether the features could have an influence on the total process, the Novasure surgery, or the outcome. This is done in consultation with an expert and based on literature studies.

The rest of the data is seen as irrelevant data and is not taken into account while modelling the data. Due to the set-up of EHRs, and the way information about the patient is gathered, the following features have to be calculated or extracted from other columns in the provided dataset:

TABLE 4.1: Summary of dataset, including their feature, range, or possible values and contingent notes.

Feature	**Range/Values**	**Notes**
Patient features - predictor variables		*Amount of missing values*
Age	$[26, 60]$	-
Complaint	*Cycle disorder*, *benign adnexal abnormality*, *uterine fibriods*, or *unknown*	18
BMI	$[15.57, 50.20]$	442
Parity	$[0, 1, 2, 3$ *and more*$]$ or *unknown*	238
Cesarean section	$0 \vee 1$	-
Uterus position	*AVF*, *Nothing found*, *RVF* or *Stretch*	-
Endometrial thickness (mm)	$[1, 26]$	571
Cavity length (mm)	$[22, 65]$	453
Cavity width (mm)	$[25, 55]$	315
Dysmenorrhea	$0 \vee 1$ or *unknown*	678
Endometriosis	$0 \vee 1$	-
Adenomyosis	$0 \vee 1$ or *unknown*	863
Uterine fibroids	$0 \vee 1$	-
Sterilisation	$0 \vee 1$	-
Process features - predictor variables		*Notes*
Appointments 2 years pre Novasure	$[0, 20]$	*77 additional appointments*
Care activities 2 years pre Novasure	$[0, 25]$	*181 additional care activities*
Waiting time (days)	[0,265]	313 missing values
Perioperative features - predictor variables		*Amount of missing values*
Anaesthesia	*Local anaesthetic*, *Sedation* or *General anaesthetic*	-
Ablation duration (sec)	[6,120]	580
Ablation power (watt)	[1,180]	237
Reintervention information - outcome variable		*Amount of missing values*
Reintervention	$0 \vee 1$	-

- **Ablation duration:** The ablation duration is extracted from the free text fields with the help of a written function. It searches through the text fields for parts of strings which indicate that the number around it provides information on the ablation duration. Chances are that, due to inconsistent ways of notation, not all noted duration values are found. This makes the value less reliable than when it would would have been gathered in an obligated numerical field.
- **Ablation power:** The ablation power is extracted from the free text fields with the help of a written function. It searches through the text fields for parts of strings which indicate that the number around it provides information on the ablation power. Chances are that, due to inconsistent ways of notation, not all noted power values are found. This makes the value less reliable than when it would would have been gathered in an obligated numerical field.
- **Appointments:** The appointments were listed in a separate file than the patient characteristics. First, the file is grouped by patients and then by appointments, after which the duplicates of appointments per patient are count. Each appointment is added as a column to the main patient database, including the amount of each appointment per patient.
- **Care activities:** The were listed in a separate file than the patient characteristics. First, the file is grouped by patients and then by care activities, after which the duplicates of care activities per patient are count. Each appointment is added as a column to the main patient database, including the amount of each care activity per patient.
- **Cavity length:** Cavity length is extracted from the free text fields with the help of a written function. It searches through the text fields for parts of strings which indicate that the number around it provides information on the cavity length. Chances are that, due to inconsistent ways of notation, not all noted values are found. This makes the value less reliable than when it would would have been gathered in an obligated numerical field. To increase the reliability of the extracted value, a controlling function has been added. It checks whether the value is between the range of 22 and 65 mm (Canteiro et al., 2010)[2].
- **Cavity width:** Cavity width is extracted from the free text fields with the help of a written function. It searches through the text fields for parts of strings which indicate that the number around it provides information on the cavity width. Chances are that, due to inconsistent ways of notation, not all noted values are found. This makes the value less reliable than when it would would have been gathered in an obligated numerical field. To increase the reliability of the extracted value, a controlling function has been added. It checks whether the value is between the range of 7 and 55 mm (Goldstuck, 2018)[2].
- **Endometrium thickness:** Endometrium thickness is extracted from the free text fields with the help of a written function. This makes the value less reliable than when it would would have been gathered in an obligated numerical field. Endometrial thickness can take any value (Smith-Bindman, Weiss, and Feldstein, 2004)[2].
- **Reintervention:** Whether a patient has undergone a reintervention, is determined by whether the cells containing information on the specific reintervention type included information or not. When the cell is empty, the column

[2]This range has also been checked by a MMC expert.

Reintervention is given a 0, and when the cell contains information, the *Reintervention* column is assigned a 1. Depending on the type of reintervention, *Invasive* or *Non-invasive*, also these columns are assigned a 0 or 1.

- **Waiting time:** The waiting time for the operating theatre is calculated by extracting the surgery request date of the date the surgery is carried out. It is possible to calculate the waiting times for operating theatre, where surgery under the local anaesthesia takes place, and not for the treatment room. This is due to the gynaecologists not being able to provide this information, because such a free text field does not exist in the EHR.

Many feature values are missing. The exact numbers can be found in table 4.1. Simply removing patients of whom information is missing, is no option. This causes information loss, loss of accuracy and creates potential biases resulting in misinterpretation of results. Therefore, missing data is going to be handled in multiple ways:

For *BMI*, *cavity length*, *cavity width*, *endometrial thickness*, *waiting time*, *ablation duration*, and *ablation power* the missing values are replaced by doing mean imputation. This decision is based on the research by Sangra and Codina (2015). Mean imputation consists in replacing any missing data by the mean of non-missing data. In their research, they propose better methods to estimate the BMI per patient, but these methods are based on other patient features, like the sex, the overweight and obesity index and total energy intake of their patients. For all of our patients, the sex is the same and the rest of the variables are not included in the dataset, so missing BMI's are replaced by the mean BMI of the dataset.

The above-mentioned patient characteristics and process features are found in the main sheet. The dataset also includes sheets for appointments and care activities. Per patient, every appointment and care activity is noted providing information about the process. For each of the existing appointments and care activities, a new column is added to the working data file and for each patient the frequency of attending these appointments and care activities is counted. These columns are taken into account when also addressing process features.

4.5 Feature exploration

The dataset exists of 1029 patients. All patients are biologically female, since Novasure surgeries are only applied to uteri. In 129 of the 1029 cases, a reintervention of any form was carried out after the first Novasure surgery, see figure 4.2. This is a percentage of 12.54%.

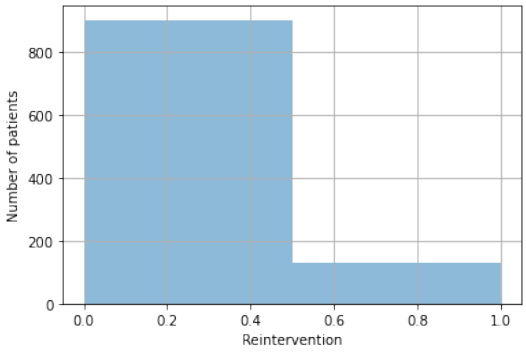


FIGURE 4.2: Histogram of population based on reintervention

The patients have an *age* between 26 and 60 years old, see figure 4.3. Pregnancy is dangerous after a Novasure surgery. Therefore, it is obligated that a patient has fulfilled the wish for children or the wish must be absent. A Novasure surgery is offered only if the patient has not had her menopause. Both constraints certify the range of age and the light negative skewness. The average age of patients is 43.83 years old. The line of patients having a reintervention based on age is quite flat, especially compared to the distribution of all patients. One peak occurs around the age of 44, the average age, and around 50 years two small peaks occur.

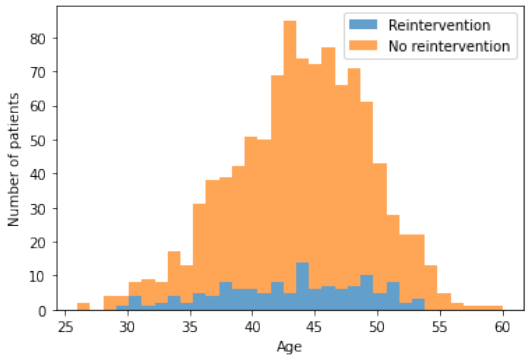


FIGURE 4.3: Histogram of population based on age

A big majority of patients undergoing a Novasure surgery started the whole process due to a *cycle disorder*, see figure 4.4. The amount of patients suffering from *benign adnexal abnormality* or *uterine fibroids* are close to zero. For 18 patients the main complaint has not been registered by the gynaecologist. The patients undergoing a reintervention also mostly suffer from *cycle disorder* in the first place, followed by *uterine fibroids* and unknown complaints.

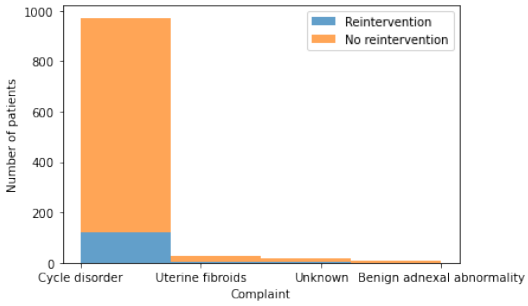
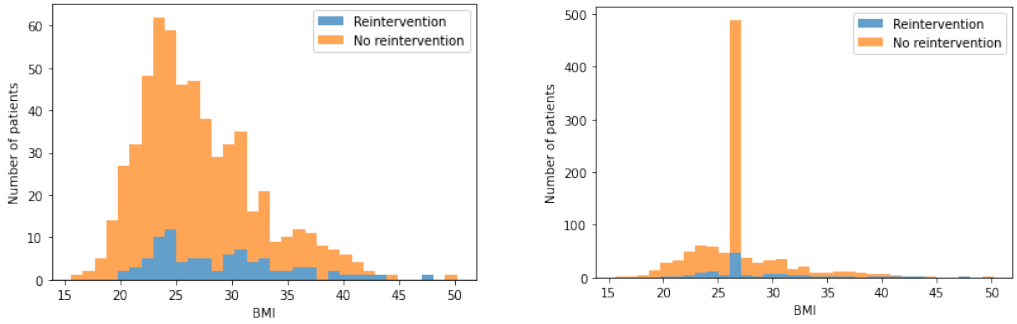


FIGURE 4.4: Histogram of population based on primary complaint

For 587 patients the *BMI* could be retrieved from the database. The *BMI* ranges between 15.57 and 50.20 and the patient population has an average *BMI* of 27.11, see figure 4.5a. The other 442 patients have been given this average *BMI*, which results in the histogram as seen in figure 4.5b. A peak in reinterventions occurs at the average *BMI*, but at first sight this seems well in relation with the amount of patients given that *BMI* score.

Of the 791 patients for whom is known how many children they delivered, 59 have delivered none, see figure 4.6. For 238 patients it is not known whether and how many children they delivered. This is presented by '-1' in the figure. When delivering a child, the cavity width is stretched due to the passing of the child. Also this activity has an influence on how the bipolar electrode array places itself on the



(A) BMI before filling NaN values

(B) BMI after filling NaN values

FIGURE 4.5: Histograms of population based on BMI

interior surface area. The amount of reinterventions seems in line with the total amount of patients.

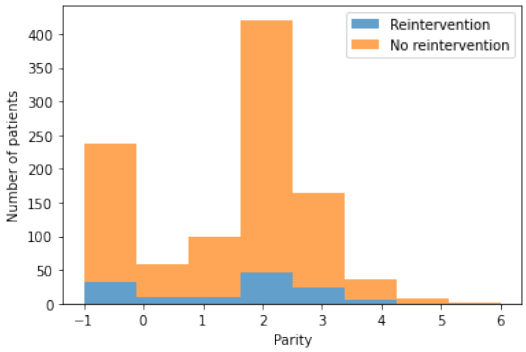


FIGURE 4.6: Histogram of population based on parity

72 patients had at least one child born with the help of a *cesarean section* (figure 4.7). A cesarean section causes damage and scars to the uterine wall, which can have an influence on how the bipolar electrode array places itself on the interior surface area. Although, literature says that having a *cesarean section* increases the chance of a reintervention, most patients undergoing a reintervention have not had a *cesarean section*.

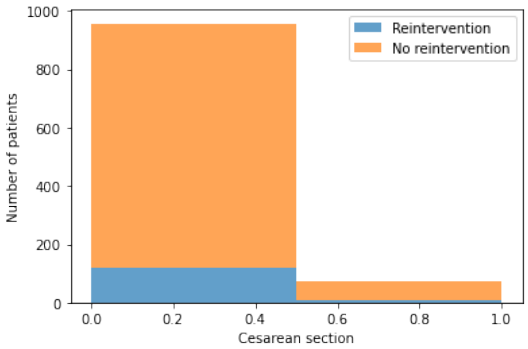


FIGURE 4.7: Histogram of population based on cesarean section

For all of the patients the *uterus position* has been researched and the results are shown in figure 4.8. The majority of the patients has an anteverted uterus (*AVF*), meaning that the uterus is turned a little forward. This group also holds the majority

of patients having reintervention, (*RVF*). 171 patients have a uterus which is turned a little backward, retroverted. For 35 patients, the uterus was found in a *stretching position*. For 99 patients the uterus has not been registered.

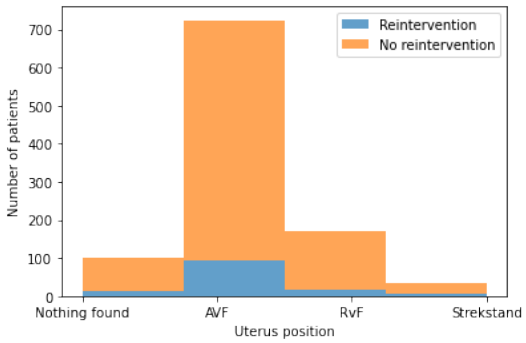
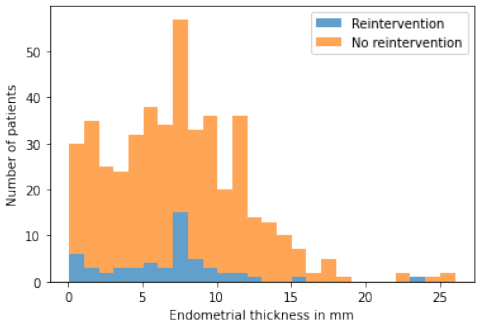
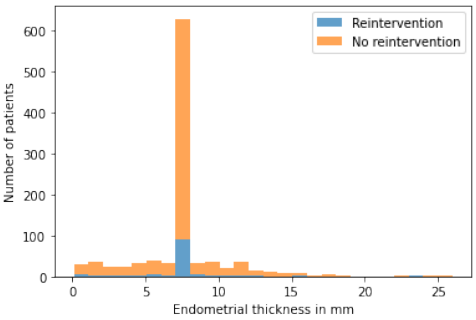


FIGURE 4.8: Histogram of population based on uterus position

The *endometrial thickness* has been registered for 458 patients, see figure 4.9a. Endometrial tissue is at least 1 mm thick. As explained on forehand, the remaining 571 patients have been given the average *endometrial thickness* of 7.74 mm, see figure 4.9b. The distribution of reinterventions follows the overall distribution of *endometrial thickness*, even after assigning the average *endometrial thickness* to the remaining patients.



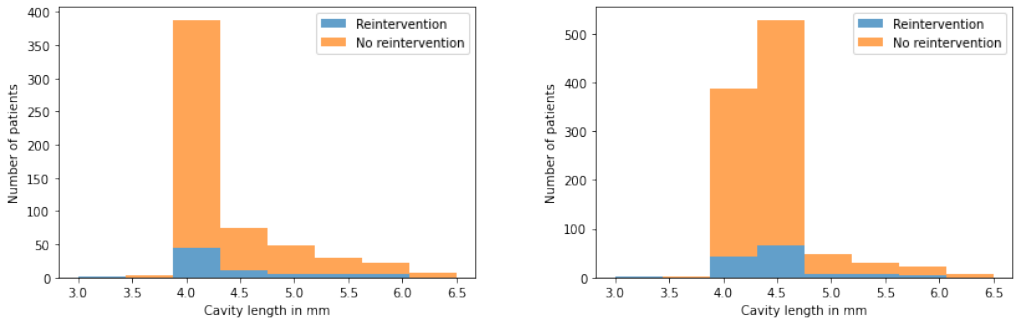
(A) Endometrial thickness before filling NaN values



(B) Endometrial thickness after filling NaN values

FIGURE 4.9: Histograms of population based on endometrial thickness

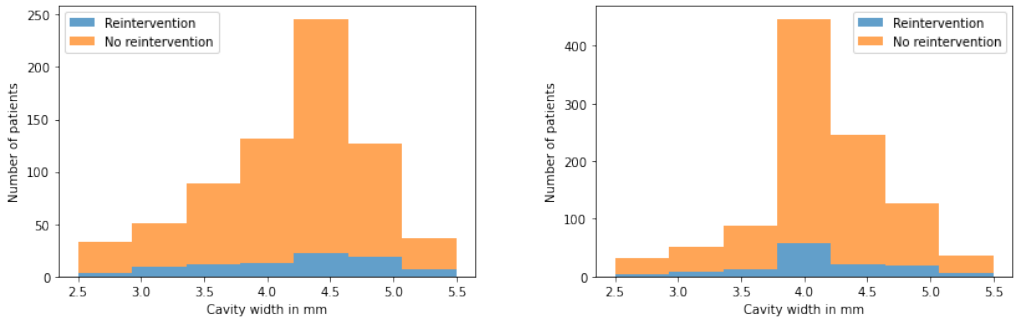
Cavity length can range between 2.2 and 6.5 cm. The *cavity length* has been registered and retrieved for 576 patients and results in the distribution presented in figure 4.10a. Like the other float variables, the remaining patients have been given the average, which for *cavity length* is 4.34 mm, see figure 4.10b. The distribution of interventions follows the trend.



(A) Cavity length before filling NaN values (B) Cavity length after filling NaN values

FIGURE 4.10: Histograms of population based on cavity length

The *cavity width* has been found for 714 patients, see figure 4.11a. It ranges between 2.5 and 5.5 mm with an average of 4.12 mm. This average has been assigned to the rest of the patients resulting in the distribution shown in figure 4.11b and the distribution of reinterventions seems to form a flatter, but sort-like curve.



(A) Cavity width before filling NaN values (B) Cavity width after filling NaN values

FIGURE 4.11: Histograms of population based on cavity width

The presence of *dysmenorrhea* has been checked for 351 patients, of which in 258 cases the presence of *dysmenorrhea* has been found. For all the unknown case, a dummy value of -1 is created (figure 4.12). The trend of reinterventions follows the main distribution of *dysmenorrhea*.

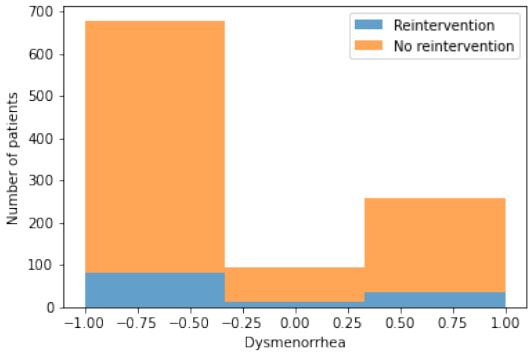


FIGURE 4.12: Histograms of population based on dysmenorrhea

For all of the patients, the presence of endometriosis has been notated. In total 55 patients suffer from endometriosis, of which 12 have undergone a reintervention. Of the 857 patients not suffering from endometriosis, 117 patients have undergone

a reintervention. The percentage of patients suffering from endometriosis is slightly higher (21,8%) than the average (12,5%).

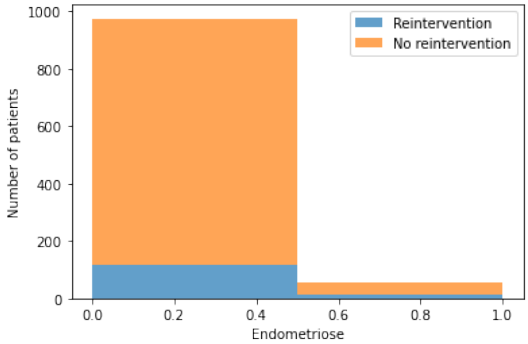


FIGURE 4.13: Histogram of population based on endometriosis

Adenomyosis is a condition which is hard to check, since it exists of tissue within the uterus wall. This dataset includes 156 patients of whom is assumed based on indications that they suffer from *adenomyosis*, 18 patients of whom *adenomyosis* is excluded, and 855 patients for whom no indications have been found or searched (figure 4.14). To this last group of patients the -1 dummy variable is assigned. Remarkable is the fact that the group of patients knowing they suffer from *adenomyosis* and undergoing a reintervention is approximately of the same size as the group of patients patients not knowing whether they suffer from *adenomyosis* and undergoing a reintervention, even though the amount of patients not knowing whether they have *adenomyosis* is four times the amount of patients knowing they have *adenomyosis*.

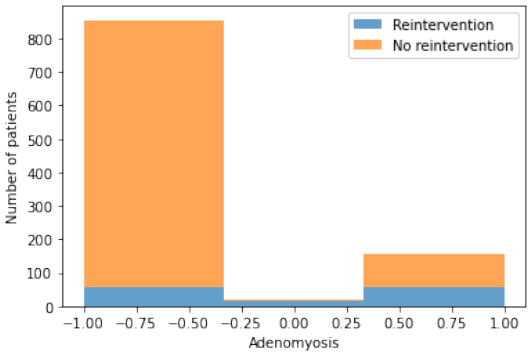


FIGURE 4.14: Histogram of population based on adenomyosis

In this dataset, 309 patients suffer from uterine fibroids and 722 do not, see figure 4.15. The amount of patients undergoing a reintervention lay relatively close to each other, even with the amount of patients not suffering form *uterine fibroids* being at least twice as much as patients suffering from *uterine fibroids*.

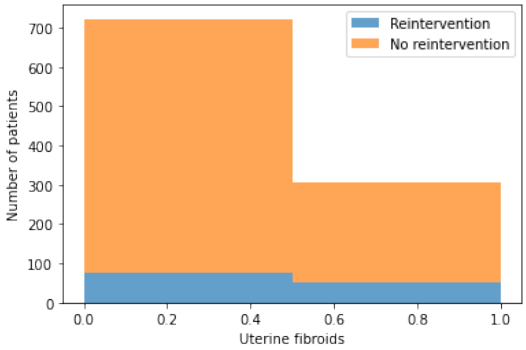


FIGURE 4.15: Histogram of population based on uterine fibroids

As seen in figure 4.16, 900 patients are not *sterilised* versus 129 patients being *sterilised*, with the amount of reinterventions relatively following this distribution.

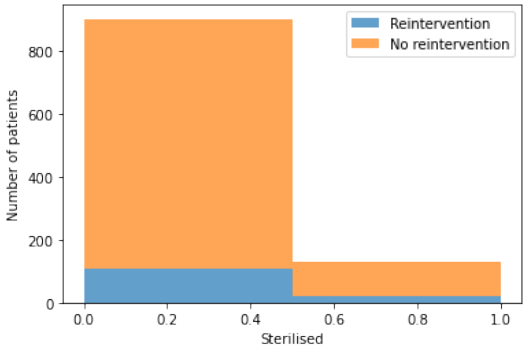
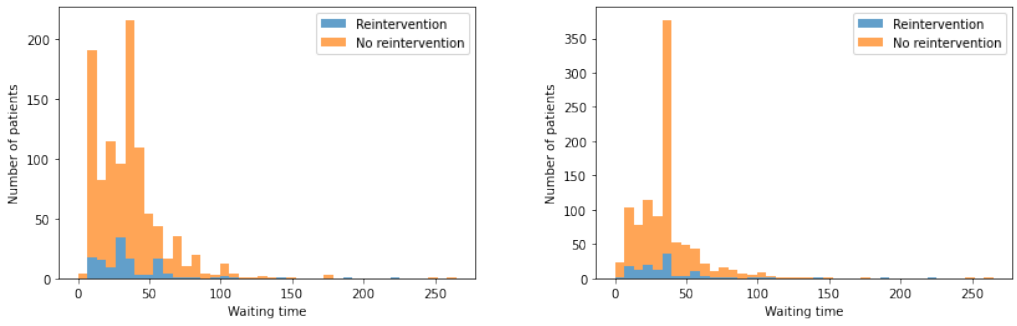


FIGURE 4.16: Histogram of population based on sterilisation

When patients undergo a *general anaesthetic* or *sedation*, the time between the request for the operation room and the Novasure surgery is tracked (figure 4.17a). Keeping track of the *waiting time* to perform a local anaesthetic is not possible. All 313 patients who received this kind of anaesthesia have been assigned the average *waiting time* (figure 4.17b). Most of the patients have their Novasure planned within 10 weeks after the request and most reinterventions occur when the Novasure was within 50 days after the request.



(A) Waiting time before filling NaN values (B) Waiting time after filling NaN values

FIGURE 4.17: Histograms of population based on waiting time

While most patients choose for a *general anaesthetic*, others choose for a *local anaesthetic* or *sedation* with their own reasons. The amount of reinterventions is also the

highest for the *general anaesthetic* (figure 4.18). The number of reinterventions after a *sedation*, is remarkably close to the number of reinterventions after *local anaesthetic*, despite their difference in amount of patients undergoing both the anaesthetics.

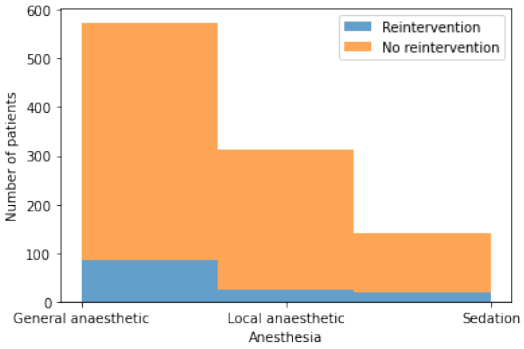
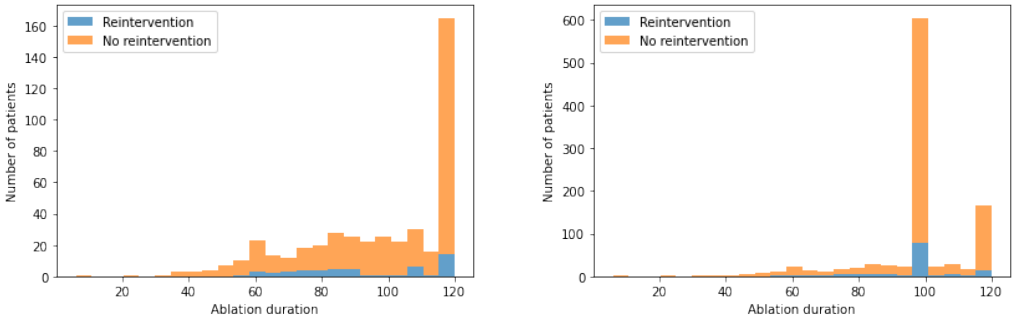


FIGURE 4.18: Histogram of population based on anaesthesia

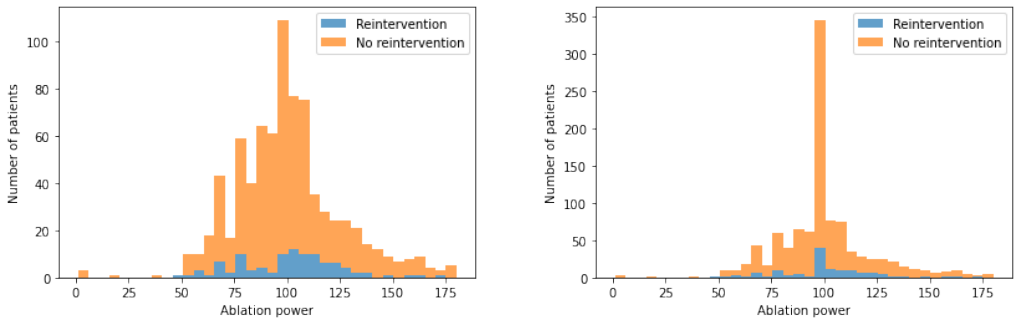
The *ablation duration* seems like a slightly growing curve, until it reaches the amount of 120 seconds (figure 4.20a). For approximately 160 patients the maximum amount of time is needed by the Novasure device. Remarkable is that before the nan-values have been filled in, no patients with an average duration of 100 seconds undergo a reintervention, but the part of patients of whom the power duration is unclear, deliver a great share in reinterventions (figure 4.20b).



(A) Ablation duration before filling NaN values (B) Ablation duration after filling NaN values

FIGURE 4.19: Histograms of population based on ablation duration

The *ablation power* is automatically determined by the Novasure device based on the endometrial thickness it finds. A slight likeliness can be found in the form of both distributions. The amount of reinterventions follow the same curve as the distribution on no reintervention (figure 4.20a), and again, the missing values have been replaced by the average (figure 4.20b). Like with the *ablation duration*, the share of reinterventions is large in the group of patients of whom the power duration is not clear.



(A) Power duration before filling NaN values (B) Power duration after filling NaN values

FIGURE 4.20: Histograms of population based on ablation duration

4.6 Data analytic strategy

After the data is preprocessed, it is used to train different models and transform it into valuable information. As one can see in figure 4.2, the data is imbalanced. There are 900 samples in the class of no reintervention and 129 in the class of reintervention. This can cause bias towards the more popular class, patients not undergoing a reintervention. To prevent this, the minority is upsampled. Using Synthetic Minority Oversampling Technique (SMOTE) (Chawla et al., 2002), a technique which acts like a data transform object, new samples are synthesised from existing samples. The minority class reintervention is oversampled to 900 samples, equalling the no reintervention class. To find out whether upsampling has impact on the outcome and taking the influence on of process features into account, this results in four experiments with different input data:

- The original set of patients with only patient characteristics
- The original set of patients with patient characteristics, appointments and care activities
- A sampled set of patients with only patient characteristics
- A sampled set of patients with patient characteristics, appointments and care activities

The dataflow of the four experiments is presented in figure 4.21. At first, the data is split into a training set of 80% and a test set of 20%. Then, the training set is sampled, after which all experiments follow equal paths. The first two experiments use the original training set. For each algorithms used, the algorithms are tuned and cross-validated using sklearn's model_selection technique GridSearchCV. In the training set, further partitions are done, to tune model parameters. Tuning these parameters is done utilising 5-fold cross-validation. The training set is used to tune algorithms, after which the test set is used to evaluate the algorithms. The lower part of the figure represents this path. For the second two experiments the training data is sampled so that both classes hold an even amount of patients. In this case there are 720 patients for each class. Then, like the other two experiments, the algorithms are tuned and cross-validated using sklearn's model_selection technique GridSearchCV. The following machine learning algorithms are included; decision trees, random forests, logistic regression, gradient boosting, neural networks, and k-nearest neighbour. Information on what these techniques include and why they

are chosen can be found in section 2.1. Model performance is determined by AUC and accuracy. The conclusion focuses on a comparison of the performance of the algorithms.

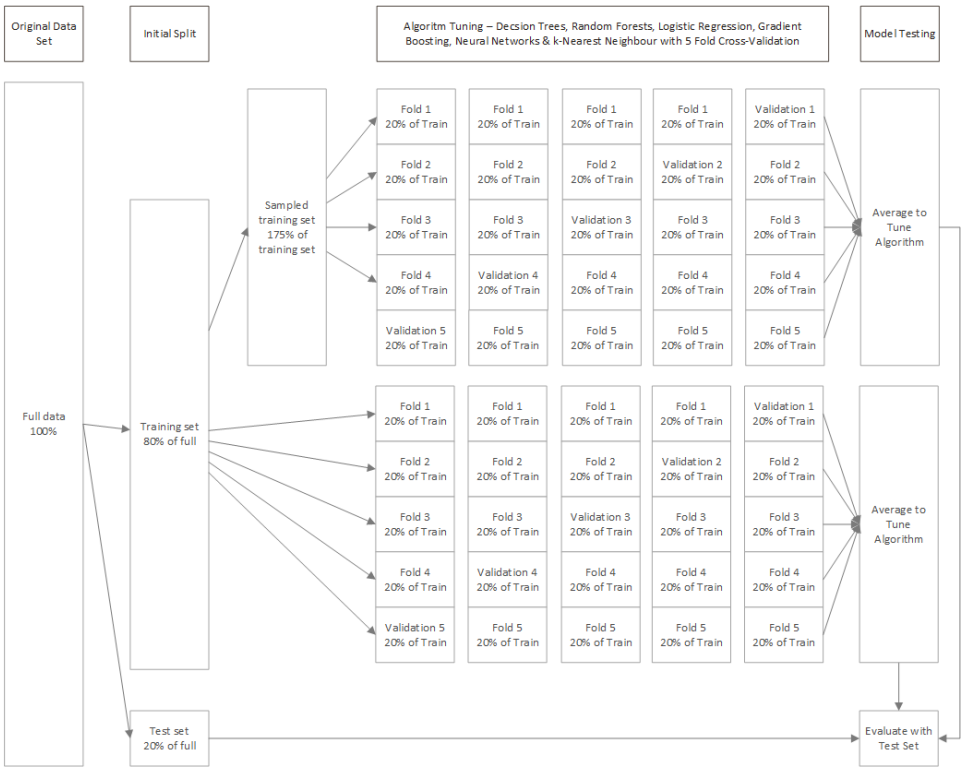


FIGURE 4.21: Tuning algorithms done for dataset including patient features and for dataset including patient features and process features - based on Suchting et al. (2018)

For each of the algorithms, we tried to find a best-fitting range. This was done during earlier experiments using trial and error. The resulting ranges are presented in table 5.1. For decision tree, all possible options for the maximum number of features are taken into account. The range of maximal tree depth is set wider to 25, when a depth of 10 did frequently not seem enough during trial and error experiments. Also the minimum number of samples required to split an internal node and the minimal samples required to be at a leaf node are found while doing trial and error experiments by broadening the range. For random forests, the number of estimators is set quite broad from 10 to 500. The maximal depth of a tree is a bit less than with decision tree. A depth of 19 seems good enough. Logistic regression is tuned taking in to account all possibilities of norm of penalty and on the maximal number of iterations. While running the algorithm with default options, the maximal number of iterations was reached before the optimisation had converged. This was not the case with 500 iterations. During later experiments was found that if this error did not occur, less iterations had a positive influence on the scoring metric. That is why the range on maximum number of iterations is between 100 and 1000 with steps of 100. Extreme gradient boosting is tuned with a range for number of estimators partly influenced by random forest and a maximum tree depth influenced by decision tree. The MLPclassifier for neural network is tuned on the hidden layer sizes, batch sizes, the maximum number of iterations. The hidden layer sizes have been kept small. From earlier experiments a hidden layer size of lower than four never seemed to satisfy, this is why the range if from 4 to 10. The batch sizes range is

somewhat bigger, being over the total number of samples with a step of 100. While the default number of iterations, 200, seems fine during the default run, the number has to increase during tuning. The range is extended to 1300 iterations with a step of 100. The maximum number of neighbours for k-nearest neighbour is based on Teinemaa et al. (2019) and Pijnenborg et al. (2021).

TABLE 4.2: Hyperparameters for algorithmic tuning per model, with n = number of features

Prediction model	**Hyperparameter**	**Range**
DT	Max features	$[1, \sqrt{n}, \log_2 n]$
DT	Max tree depth	[None, 1, 5, 10, 15, 20, 25]
DT	Min samples split	[2, 6, 10, 14, 16, 20, 24]
DT	Min samples leaf	[1, 7, 13, 19, 25, 31, 37, 43]
RF	#estimators	[10, 50 , 100, 150, 200, 250, 300, 350, 400, 450, 500]
RF	Max tree depth	[None, 1, 3, 5, 6, 7, 9, 11, 13, 15, 17, 19]
LR	Penalty	[L1, L2, Elastic-Net]
LR	Max iterations	[100, 200, 300, 400, 500, 600, 700, 800, 900, 1000]
XGBoost	#estimators	[1, 10, 50, 100, 150, 200, 250, 300, 350, 400, 450]
XGBoost	Max depth	[None, 2, 4, 6, 8, 10, 12, 14, 16, 18, 20, 22, 24, 26]
NN	Hidden layer sizes	[4, 5, 6, 7, 8, 9, 10]
NN	Batch sizes	[1, 50, 100, 200, 250, 300]
NN	Max iterations	[400, 500, 600, 700, 800, 900, 1000. 1100, 1200, 1300]
NN	Early stopping	[True, False]
kNN	#neighbours	[2, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60]

The ranges are validated during the first experiment with the original data set as input data and only focusing on patient characteristics. In case the hyperparameter tuning shows that the range should be adjusted, the adjusted range is used in the three remaining experiments.

The best parameters which result from this cross-validated model are implemented in the classifier to create the optimal model. All models are compared on their AUC, accuracy, f1 score, precision and recall. The AUC is used as strategy to evaluate the performance of the cross-validated model on the test set. The reason behind this, is that focusing on accuracy creates misleading results for this data set. The data is so imbalanced, that when every case would be labelled as 'no reintervention', the accuracy would be 0.87.

The f1 score is the harmonic mean of the precision and recall. When using f1 score as scoring metric, the focus on obtaining as many true positives as possible, instead of as many true values as possible.

From the optimal models, rankings with feature importances are obtained in order to help answer the research questions.

Chapter 5

Results

In this chapter, the results are discussed. Algorithm performance is presented for the original dataset and the sampled dataset, each only having patient characteristics as input data and later also having patient characteristics and process features as input data. Then, rankings are presented for features.

5.1 Hyperparameters

This section presents the tuning of at least one of the hyperparameters for each of the algorithms. Due to time restrictions, only the tuning process of the hyperparameter per algorithm for the first experiment (original data, only patient characteristics) is shown. For the sake of time, it is assumed that these ranges also fit for the other three experiments, including (1) original data set with patient characteristics and process features, (2) sampled data set with patient characteristics and (3) sampled data set with patient characteristics and process features.

5.1.1 Decision tree

Decision tree is tuned on the maximum number of features, the maximal tree depth, the minimal sample needed to for a split and the minimum samples need to create a leaf. All four hyperparameters are discussed.

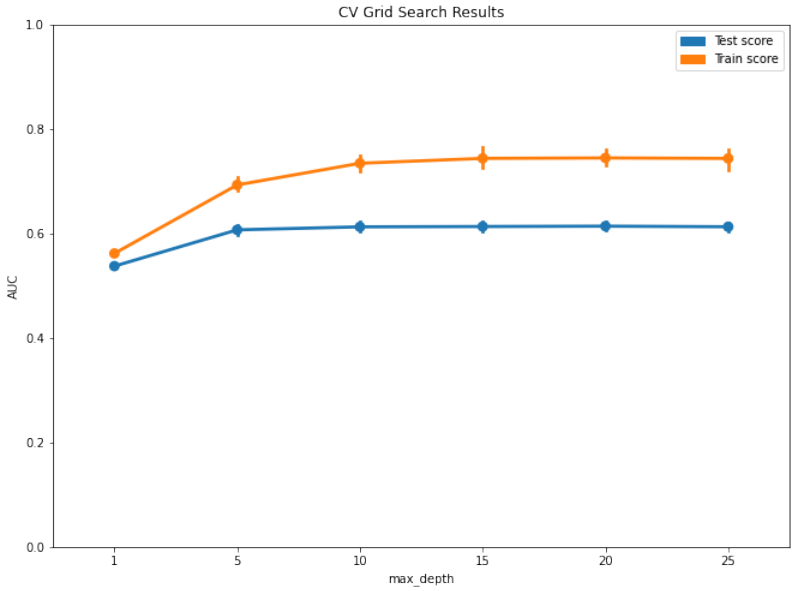


FIGURE 5.1: Decision tree algorithm tuned on maximum tree depth

Figure 5.1 shows the tuning of the maximum tree depth. The AUC increases from a maximum tree depth of one to ten, after which the curve flattens for both training data and test data. The AUC does not seem to differ much with a tree depth between 10 and 25, but to stay safe, the range is kept as [1, 5, 10, 15, 20, 25].

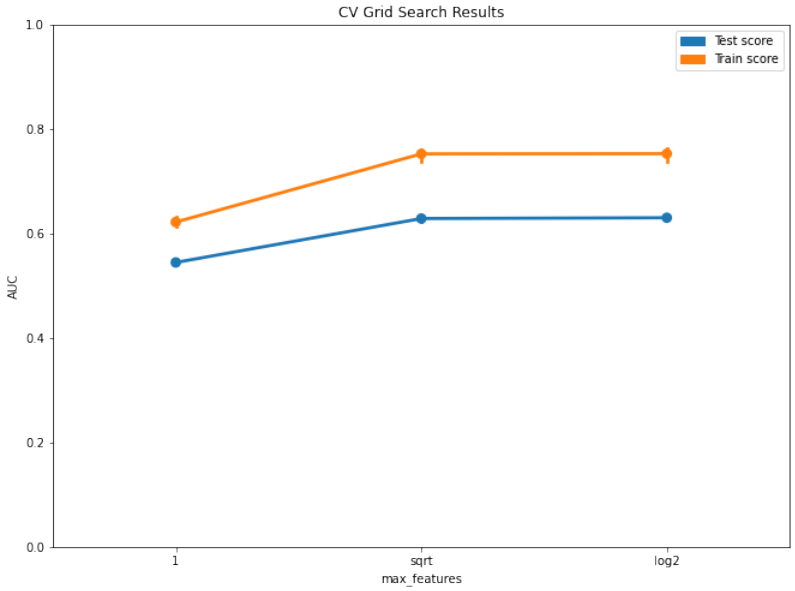


FIGURE 5.2: Decision tree algorithm tuned on maximum number of features

Figure 5.2 presents the best AUC on training and test data for number of features are hyperparameters. As one can see, there is not much difference in performance when it comes to taking $\sqrt{n}$ or $\log_2 n$ as maximum amount of features, as long as 1 is not chosen. For the sake of completeness, 1 has been kept as optional value for the hyperparameter tuning.

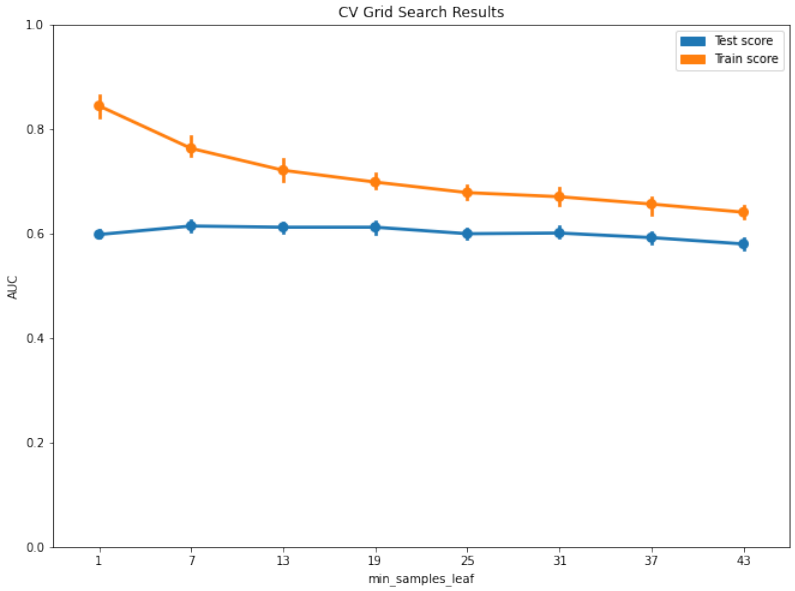


FIGURE 5.3: Decision tree algorithm tuned on minimum samples in leaf

The tuning of the minimum amount of samples in a leaf is presented in figure 5.3. The train score decreases over the whole range, decreasing less steep from 13 on. The test score slightly increases in the beginning, but also decreases from 13 on. Tuning resulted in 19 being the best value for the hyperparameter. The range for tuning is therefore shortened to [1, 5, 9, 13, 17, 21, 25, 29].

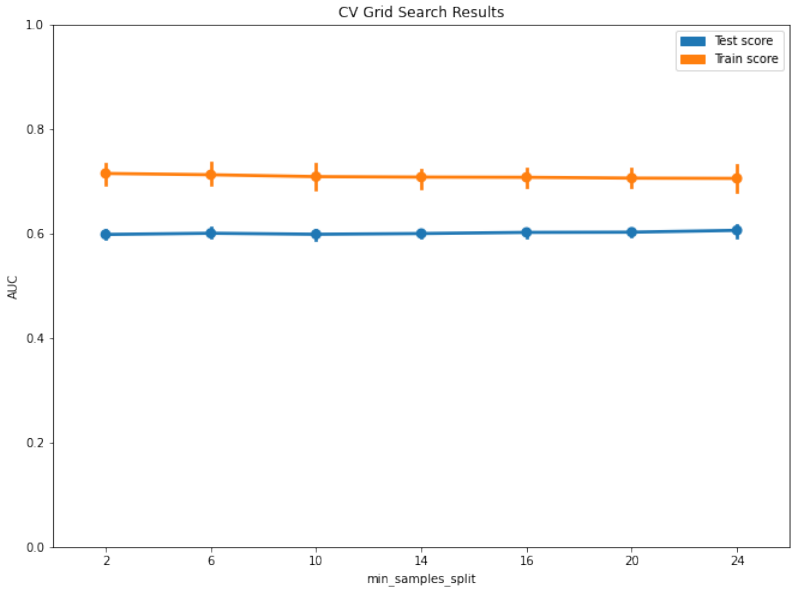


FIGURE 5.4: Decision tree algorithm tuned on minimum samples for split

Figure 5.4 shows the AUC for different values of the maximum samples for a split. The train score and test score look like two parallel lines. For a more detailed view, we refer to figure **??** presented in Appendix A, from which can be seen that the train score and test score slightly converge with the train score decreasing and test score increasing. Tuning the hyperparameter provided 16 as optimal value, and so the range has been kept between 2 and 24.

5.1.2 Random forest

Random forest is tuned on two hyperparameters: the number of estimators and the maximum tree depth.

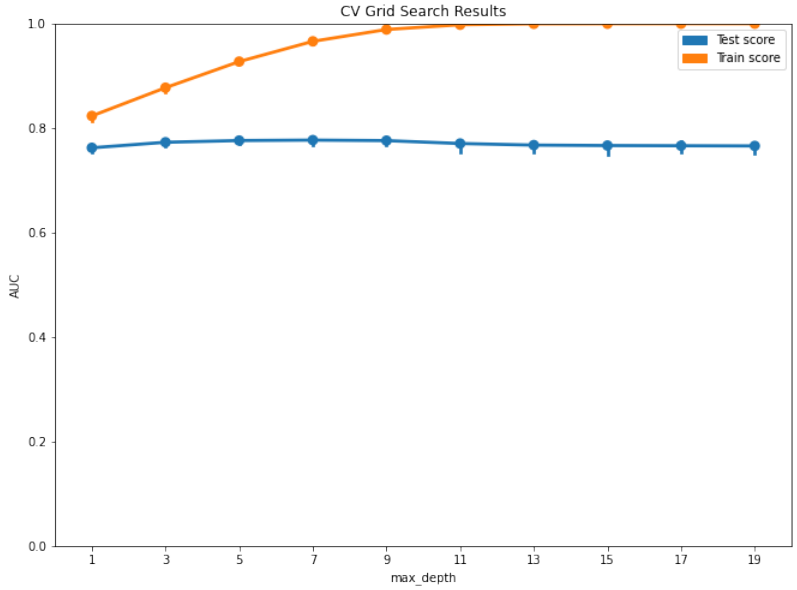


FIGURE 5.5: Random forest algorithm tuned on maximum tree depth

Figure 5.5 shows tuning the maximum tree depth. From a tree depth of 11, the train score equals 1 and the model seems to be overfit. The test score decreases from approximately the same tree depth. For tuning the other models, the range is adjusted to [None, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. The process of tuning the number

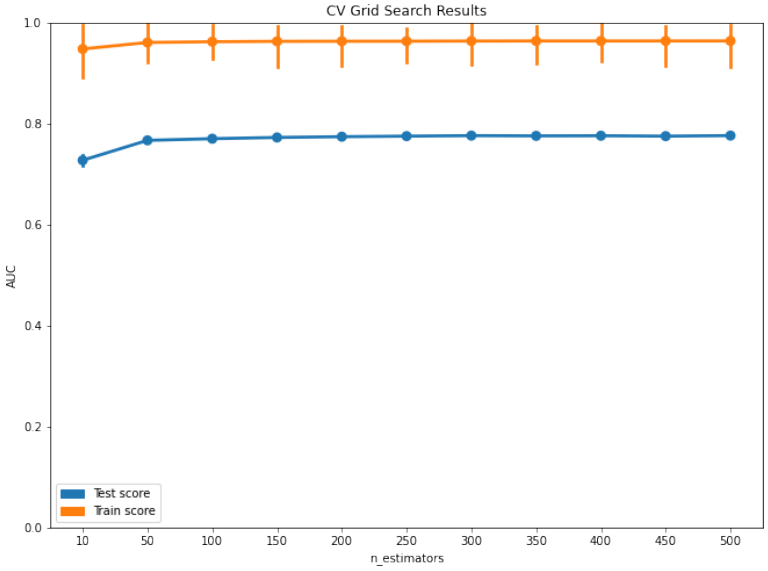


FIGURE 5.6: Random forest algorithm tuned on number of estimators

of estimators is presented in figure 5.6. From 10 to 50 estimators the curve for both train and test score increase relatively sharply, after which both curves quite stabilise. Both curves do not show a clear peak at the moment. The range of number of estimators has been rearranged to [1, 30, 60, 90, 120, 150, 180, 210, 240, 270, 300].

5.1.3 Logistic regression

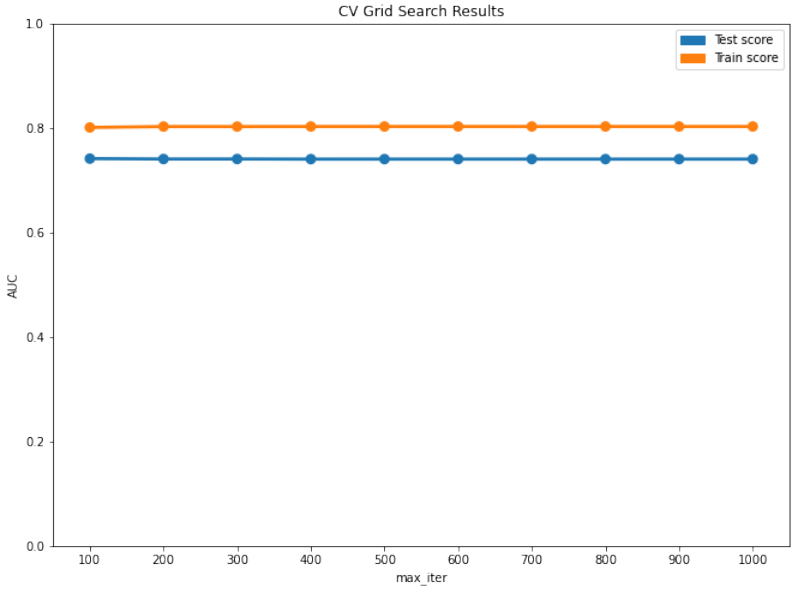


FIGURE 5.7: Logistic regression algorithm tuned on maximum number of iterations

For logistic regression, two hyperparameters are tuned: the penalty and the maximum amount of iterations. While performing the algorithm with default parameters, the error "*STOP: TOTAL NO. of ITERATIONS REACHED LIMIT. Increase the number of iterations (max_iter)*" arose until the maximum amount of iterations was set to 500. Trial and error resulted in that a number of iteration of under 500 should be included in hyperparameter tuning as well. That is why the range has been set from 500 iterations to 1000 while tuning the hyperparameters.

Figure 5.7 shows two appearing horizontal straight lines in tuning the maximum number of iterations. In Appendix A, figure A.2 provides a more informative plot, with an y-axis ranging from 0.75 to 0.81. The train score increases between 100 iterations and 200 iterations, after which it drops a little and than stabilises. The test score decreases between 100 iterations and 300 iterations. Then, it increases a little, after which it stabilises. The range for the hyperparameter has been set between 100 iterations and 1000 iterations with steps of 100, so that the error message is taken into account, as well as the number of iterations belonging to higher train and test scores.

Figure 5.8 shows that only L2 penalty creates an AUC value. This is because the default solver (lbfgs), which is used, only works with L2 penalty, and not with L1 and Elastic-Net. Therefore, the penalty hyperparameter is removed from tuning processes for the other experiments.

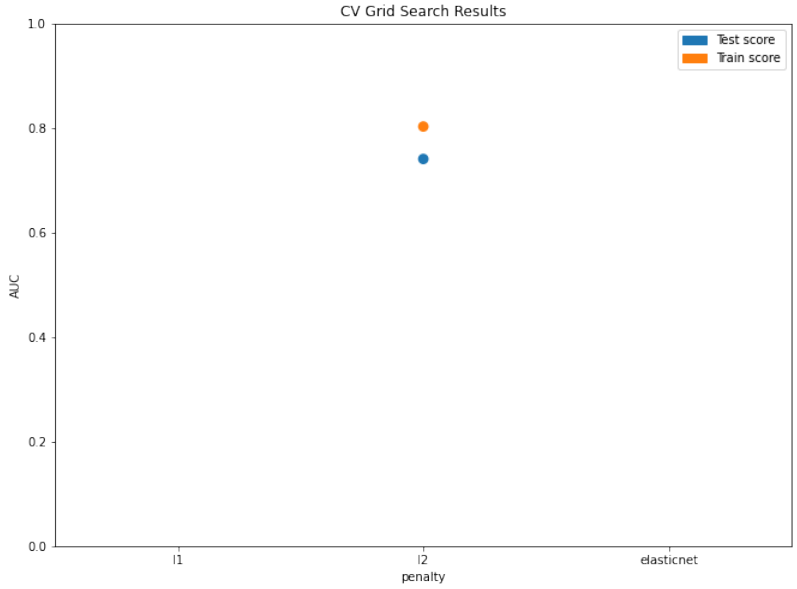


FIGURE 5.8: Logistic regression algorithm tuned on penalty

5.1.4 Extreme gradient boosting

The extreme gradient boosting algorithm is tuned on the number of estimators and the maximum tree depth.

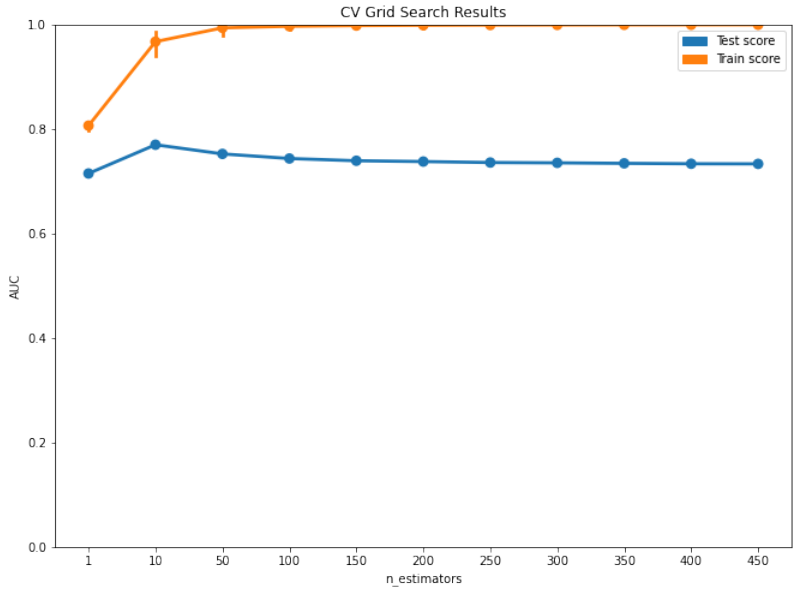


FIGURE 5.9: Extreme gradient boosting algorithm tuned on number of estimators

Figure 5.9 presents tuning on the number of estimators. The original range is widely chosen, as from 100 estimators the train score is overfitting. Also, test score is decreasing from 10 estimators and onward. Therefore, a more detailed range is chosen: [1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100].

Also the range for the maximum tree depth is estimated too wide, see figure 5.10. From a tree depth of 12, both train and test scores do not seem to increase or decrease anymore. The range is shortened to [None, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14].

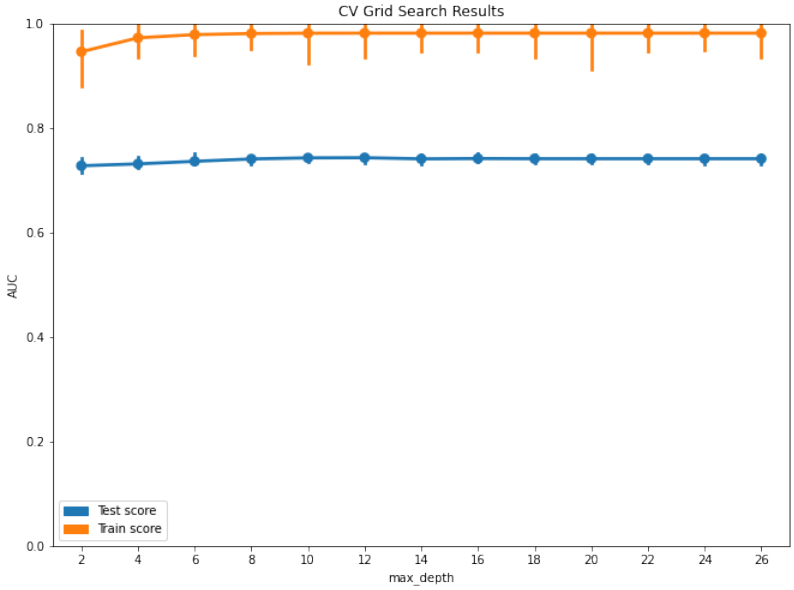


FIGURE 5.10: Extreme gradient boosting algorithm tuned on maximum tree depth

5.1.5 Neural network

The neural network algorithm is tuned on four hyperparameters: the size of the hidden layers, the batch sizes, the maximum amount of iterations and whether early stopping is valuable.

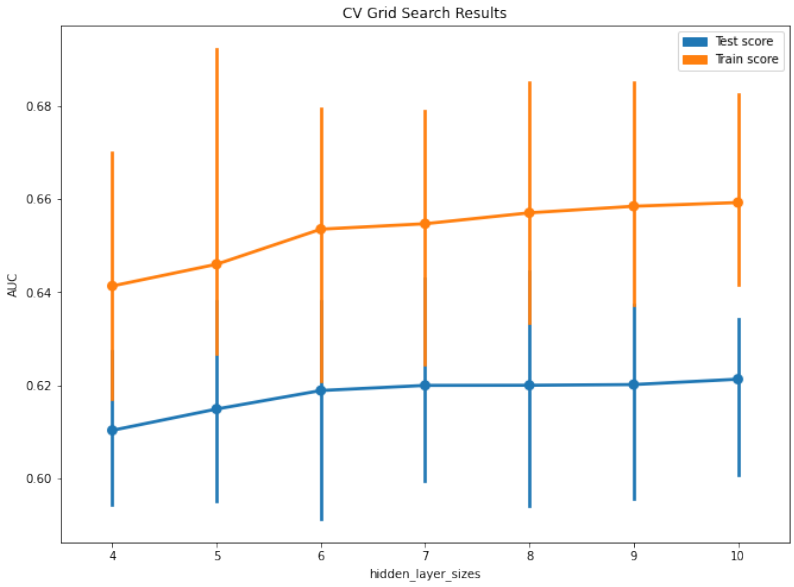


FIGURE 5.11: Neural network algorithm tuned on hidden layer sizes

Compared to figure A.3 in Appendix A, figure 5.11 provides more detailed information on the tuning of hidden layer sizes for neural network. Between the hidden layer sizes of four and six, the AUC for both train score and test score increases relatively steep, after which both curves continue increasing, but less. Despite that the optimum hidden layer size is five for this tuned model, the prediction is that the

hidden layer sizes should have been tuned on a wider range, given the continuing increasing curves.

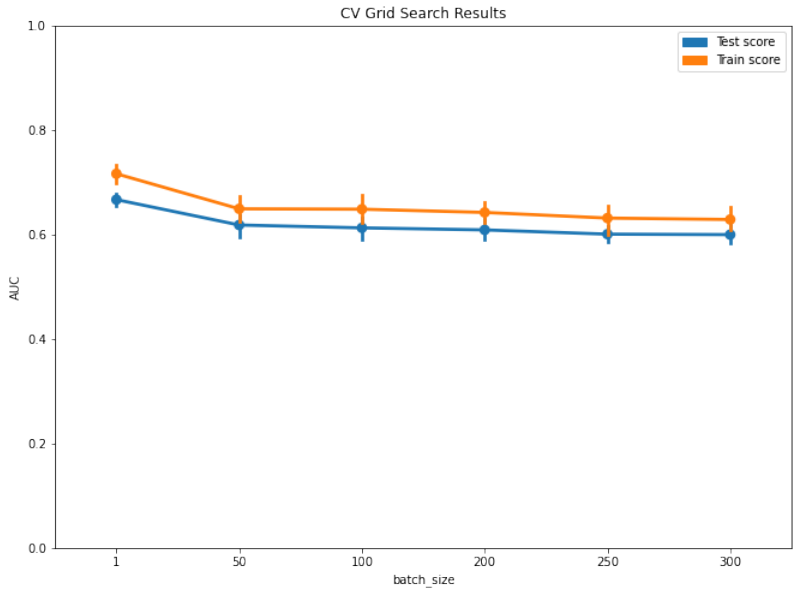


FIGURE 5.12: Neural network algorithm tuned on batch sizes

The process on tuning the hyperparameter on batch size is presented in figure 5.12. Both train and test score show a decrease in AUC from a batch size of 1 to a batch size of 250. From there on, both curves appear horizontally. With the optimum batch size being one for tuning on this range, the range is adjusted for the other three experiments to [1, 50, 100, 150].

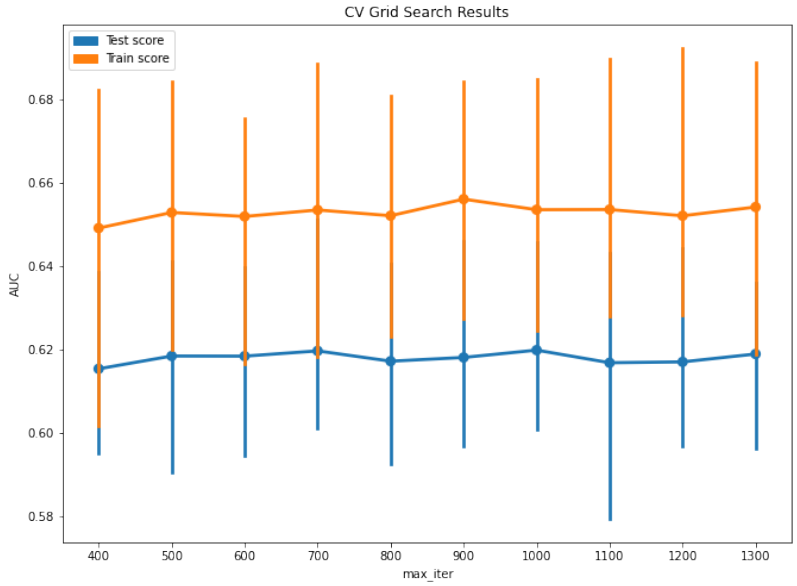


FIGURE 5.13: Neural network algorithm tuned on maximum number of iterations

In figure 5.13 provides a more detailed view on tuning the maximum amount of iterations for neural network. The original figure A.4 can be found in Appendix A. Figure 5.13 provides no clear trend on the train and test score for the maximum amount of iterations.

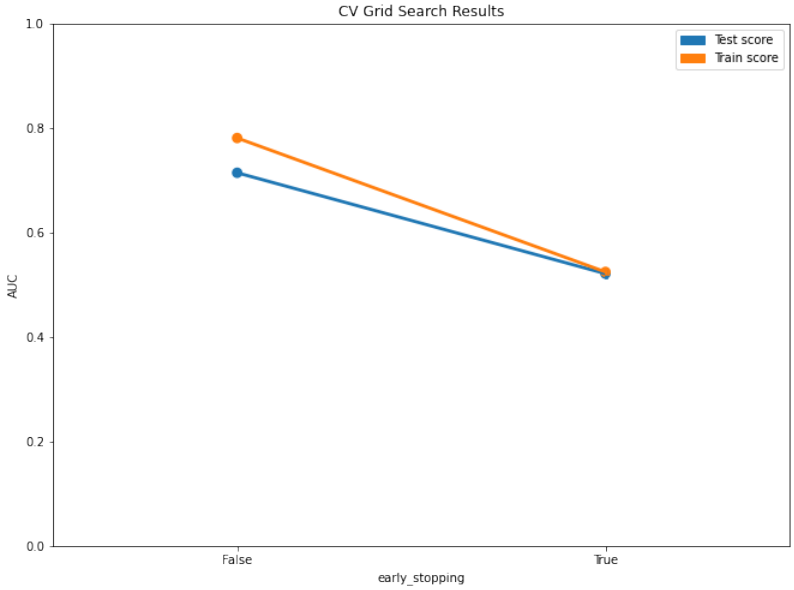


FIGURE 5.14: Neural network algorithm tuned on early stopping

Figure 5.14 shows very clear that continuing the training process is preferred over early stopping when the loss does not improve enough anymore. The difference in train and test score (over 0.20 and over 0.10, respectively) and the total amount of fits, led to the decision to remove the hyperparameter as tuning hyperparameter for the other three experiments.

5.1.6 K-nearest neighbour

K-nearest neighbour is tuned on the number of neighbours. As one can see in figure 5.15, the train score starts decreases from two neighbours, whereas the test score rises from the amount of two neighbours, to the peak at approximately 30 neighbours. After this peak, the test score also decreases. The optimal amount of neighbours for the original dataset on patient characteristics is 25. The range is kept the same for the other experiments.

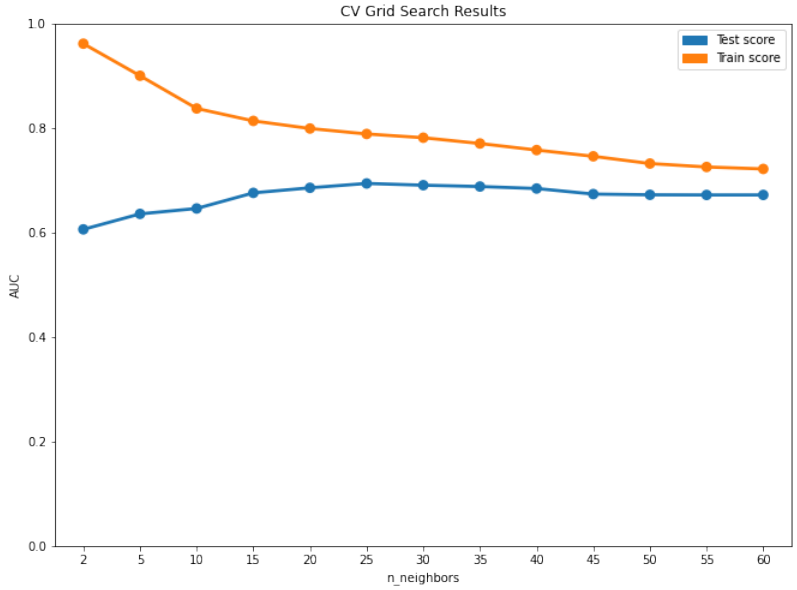


FIGURE 5.15: K-nearest neighbour algorithm tuned on number of neighbours

5.1.7 Overview

After displaying hyperparameter processes, some adjustments in ranges are made to improve hyperparameter tuning for the three remaining experiments. Table 5.1 provides an overview of the hyperparameter ranges on which the models are tuned. To recall, for decision tree the range of the minimum of samples in a leaf is more detailed, the penalty hyperparameter for logistic regression is removed, both the ranges for number of estimators and maximum tree depth for extreme gradient boosting are more detailed, the range of batch sizes for neural network is more detailed, and also for neural network, the early stopping hyperparameter is removed.

TABLE 5.1: Hyperparameters for algorithmic tuning per model, with n = number of features

Prediction model	**Hyperparameter**	**Range**
DT	Max features	$[1, \sqrt{n}, \log_2 n]$
DT	Max tree depth	[None, 1, 5, 10, 15, 20, 25]
DT	Min samples split	[2, 6, 10, 14, 16, 20, 24]
DT	Min samples leaf	[1, 5, 9, 13, 17, 21, 25, 29]
RF	#estimators	[1, 30, 60, 90, 120, 150, 180, 210, 240, 270, 300]
RF	Max tree depth	[None, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]
LR	Max iterations	[100, 200, 300, 400, 500, 600, 700, 800, 900, 1000]
XGBoost	#estimators	[1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100]
XGBoost	Max depth	[None, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]
NN	Hidden layer sizes	[4, 5, 6, 7, 8, 9, 10]
NN	Batch sizes	[1, 50, 100, 150]
NN	Max iterations	[400, 500, 600, 700, 800, 900, 1000, 1100, 1200, 1300]
kNN	#neighbours	[2, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60]

5.2 Algorithm performance

The algorithm performance is presented in this section. For each of the algorithms in each of the experiments, the tuned hyperparameters are first presented. A table including the accuracy, AUC, f1 score, precision, and recall follows. When the difference between the AUC from default hyperparameters and the AUC from tuned hyperparameters provides an interesting image, these figures are shown as well. If not, the figures can be found in appendix B.

5.2.1 Original dataset: Patient characteristics

Decision tree

The decision tree model is tuned on the following hyperparameters and resulting values and results (table 5.2):

- Max features: $\log_2 n$
- Max tree depth: 15
- Min samples split: 16
- Min samples leaf: 19

TABLE 5.2: Algorithm performance for decision tree on imbalanced dataset - patient characteristics

Using default parameters			**Using tuned parameters**	
accuracy		0.77		0.88
AUC		0.54		0.77
	0	1	0	1
F1 score	0.86	0.20	0.93	0.29
precision	0.88	0.18	0.89	0.56
recall	0.84	0.23	0.98	0.19

Tuning decision tree has a positive effect on almost all metrics. Accuracy increases with 0.11 and the AUC increases with a step of 0.23, see table 5.2. Also the f1 score increases, with 0.07 and 0.09 towards no reintervention and reintervention respectively. Precision towards no reintervention increases with 0.01 to 0.89 and towards reintervention, it increases significantly from 0.18 to 0.56. The recall towards no reintervention increases from 0.84 to 0.98, but decreases from 0.23 to 0.19 towards reintervention.

Random forest

The random forest model is tuned on the following hyperparameters and resulting values and results (table 5.3):

- Max tree depth: 7
- #estimators: 300

TABLE 5.3: Algorithm performance for random forest on imbalanced dataset - patient characteristics

Using default parameters			Using tuned parameters	
accuracy		0.87		0.87
AUC		0.75		0.72
	0	1	0	1
F1 score	0.93	0.19	0.93	0.00
precision	0.89	0.50	0.87	0.00
recall	0.98	0.12	0.99	0.00

After tuning the hyperparameters and implementing them in the model, the results have worsened compared to using default hyperparameters (table 5.3). Only one case has been predicted as a reintervention, but incorrectly. This causes f1 score, precision and recall to be 0.00 towards reintervention. The accuracy stayed 0.87, while AUC decreased from 0.75 to 0.72.

Logistic regression

The logistic regression model is tuned on the following hyperparameters and resulting values and results (table 5.4):

- Max iterations: 100
- Penalty: L2

TABLE 5.4: Algorithm performance for logistic regression on imbalanced dataset - patient characteristics

Using default parameters			Using tuned parameters	
accuracy		0.86		0.86
AUC		0.78		0.78
	0	1	0	1
F1 score	0.92	0.29	0.92	0.29
precision	0.90	0.40	0.90	0.40
recall	0.95	0.23	0.95	0.23

While running on default parameters, the maximum number of iterations was reached and should be increased. While running with a range from 500 to 1000 with a step of 100, the best value for the hyperparameter was 100. None of the measures seem to have improved or worsened, as seen in table 5.4, which is logical since the confusion matrix obtained with tuned hyperparameters is the same as using default hyperparameters.

Extreme gradient boosting

The extreme gradient boosting model is tuned on the following hyperparameters and resulting values and results (table 5.5):

- Max depth: None
- #estimators: 10

TABLE 5.5: Algorithm performance for extreme gradient boosting on imbalanced dataset - patient characteristics

Using default parameters			**Using tuned parameters**	
accuracy		0.84		0.87
AUC		0.65		0.73
	0	1	0	1
F1 score	0.91	0.27	0.93	0.23
precision	0.89	0.33	0.89	0.44
recall	0.93	0.23	0.97	0.15

Table 5.5 shows that tuning the hyperparameters for extreme gradient boosting has a positive effect on the accuracy and AUC. Accuracy increased with 0.03 to 0.87 and AUC with 0.08 to 0.73. Predictions on which patients undergoing a reinterventions have increased, causing precision to increase from 0.33 to 0.44. Overall, less patients are predicted to undergo a reintervention. With default hyperparameters 18 reinterventions were predicted and after tuning only nine in total. F1 score and recall therefore increase with 0.02 and 0.04, respectively, towards no reintervention, but decrease with 0.04 and 0.08 towards reintervention.

Neural network

The neural network model is tuned on the following hyperparameters and resulting values and results (table 5.6):

- Hidden layer sizes: 5
- Batch sizes: 1
- Max iterations: 1300
- Early stopping: False

TABLE 5.6: Algorithm performance for neural network on imbalanced dataset - patient characteristics

Using default parameters			**Using tuned parameters**	
accuracy		0.88		0.87
AUC		0.80		0.82
	0	1	0	1
F1 score	0.94	0.20	0.93	0.00
precision	0.89	0.75	0.87	0.00
recall	0.99	0.12	1.00	0.00

From table 5.6 can be seen that using tuned hyperparameters for neural network has a large negative influence on the outcome. The model seems to work good using default hyperparameters, with an accuracy of 0.88 and an AUC of 0.80. F1 score, precision and recall are all above 0.90 towards no reintervention, and 0.20, 0.75, and 0.12 towards reintervention. The fact that precision is 0.75, is due to only four reintervention predictions, of which one is wrongly predicted. Predicting using the tuned hyperparameters results in ill-defined metrics. No reinterventions are pridicted, causing f1 score, precision and recall to equal 0.00.

K-nearest neighbour

The k-nearest neighbour model is tuned on the following hyperparameters and resulting values and results (table 5.7):

- #neighbours: 25

TABLE 5.7: Algorithm performance for k-nearest neighbour on imbalanced dataset - patient characteristics

Using default parameters			Using tuned parameters	
accuracy		0.86		0.87
AUC		0.50		0.61
	0	1	0	1
F1 score	0.93	0.00	0.93	0.00
precision	0.87	0.00	0.87	0.00
recall	0.99	0.00	1.00	0.00

While from the metrics in table 5.7, it does not necessarily seem like it, different predictions are made. Using default hyperparameters, two reinterventions are wrongly predicted versus 26 wrongly predicted no reinterventions. With tuned hyperparameters, no reinterventions are predicted all. This has caused accuracy to increase with 0.01 to 0.87 and AUC with 0.11 to 0.61. Recall towards no reintervention equals one this time, because all cases are predicted to undergo no reintervention.

5.2.2 Original dataset: Patient characteristics and process features

Decision tree

The decision tree model is tuned on the following hyperparameters and resulting values and results (table 5.8):

- Max features: $\sqrt{n}$
- Max tree depth: 20
- Min samples split: 24
- Min samples leaf: 9

TABLE 5.8: Algorithm performance for decision tree on imbalanced dataset - patient characteristics and process features

Using default parameters			Using tuned parameters	
accuracy		0.78		0.87
AUC		0.56		0.56
	0	1	0	1
F1 score	0.87	0.24	0.93	0.00
precision	0.89	0.21	0.87	0.00
recall	0.85	0.27	1.00	0.00

Although tuning hyperparameters uses AUC as scoring metric, the AUC stays 0.56 after using tuned hyperparameters, see table 5.8. Also, the model using default

parameters performs quite good towards reintervention, with an f1 score of 0.24. 27% of the positive cases was correctly predicted, and 7 out of 33 positive predicted cases was correct. Using the tuned hyperparameters, no reinterventions were predicted and the metrics are again ill-defined.

Random forest

The random forest model is tuned on the following hyperparameters and resulting values and results (table 5.9):

- Max tree depth: 4
- #estimators: 210

TABLE 5.9: Algorithm performance for random forest on imbalanced dataset - patient characteristics and process features

	Using default parameters		Using tuned parameters	
accuracy		0.87		0.87
AUC		0.72		0.74
	0	1	0	1
F1 score	0.93	0.00	0.93	0.00
precision	0.87	0.00	0.87	0.00
recall	1.00	0.00	1.00	0.00

Fitting the model with default hyperparameters returns an ill-defined model. No reinterventions are predicted. Using the tuned hyperparameters, no reinterventions were predicted and the metrics are again ill-defined. Despite the same values for the other metrics, the AUC's differ with 0.02 as presented in figure 5.16.

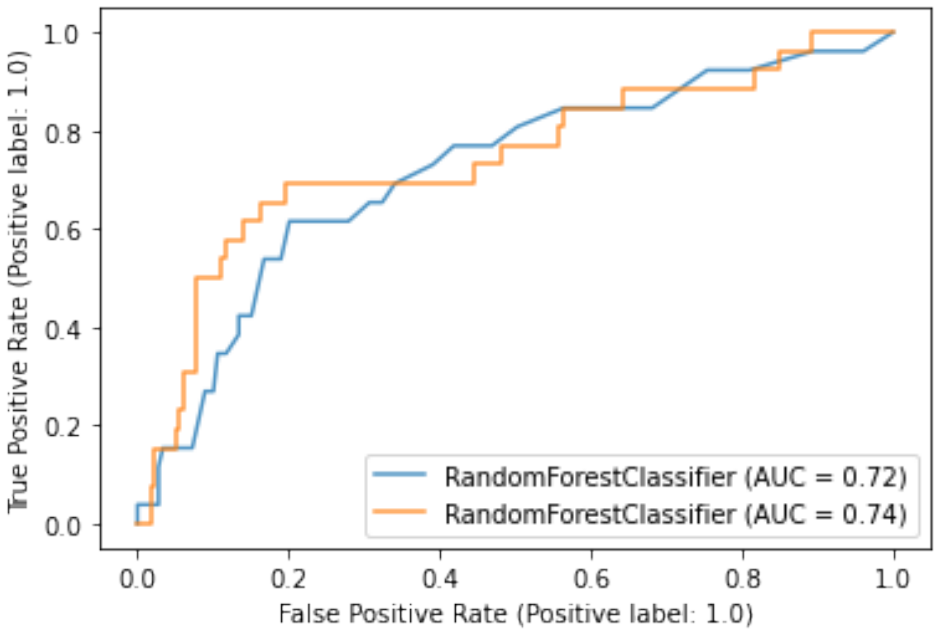


FIGURE 5.16: AUC's for random forest on imbalanced dataset - patient characteristics and process features, with the default hyperparameters in blue and the tuned hyperparameters in orange.

Logistic regression

The logistic regression model is tuned on the following hyperparameters and resulting values and results:

- Max iterations: 100
- Penalty: L2

TABLE 5.10: Algorithm performance for logistic regression on imbalanced dataset - patient characteristics and process features

Using default parameters			Using tuned parameters	
accuracy		0.88		0.85
AUC		0.75		0.73
	0	1	0	1
F1 score	0.94	0.33	0.92	0.25
precision	0.90	0.60	0.89	0.36
recall	0.98	0.23	0.95	0.19

Like with the logistic regression model using only patient characteristics as input data, the maximum number of iterations was reached and should be increased while running on default parameters. The message occurred until a maximum number of iteration of 500 was chosen. Still, tuning the hyperparameter resulted in 100 iterations being the optimal, but using this in the hyperparameter gives worse predictions than using the default parameter, see table 5.10. Accuracy drops from 0.88 to 0.85 and AUC from 0.75 to 0.73. Also, f1 score, precision, and recall decrease towards both no reintervention and reintervention, with values between 0.01 and 0.08. On top of that, precision towards reintervention drops with 0.24. With default hyperparameters, 6 out of 10 reintervention predictions were correct, with tuned hyperparameter 5 out of 14 reintervention predictions were correct.

Extreme boosting

The extreme gradient boosting model is tuned on the following hyperparameters and resulting values and results (table 5.11):

- Max depth: 2
- #estimators: 10

TABLE 5.11: Algorithm performance for extreme gradient boosting on imbalanced dataset - patient characteristics and process features

Using default parameters			Using tuned parameters	
accuracy		0.86		0.87
AUC		0.76		0.76
	0	1	0	1
F1 score	0.93	0.18	0.93	0.07
precision	0.88	0.38	0.89	0.50
recall	0.97	0.12	0.99	0.04

Although, hyperparameter tuning is done with AUC as scoring metric, the AUC with default hyperparameters is the same as the one with tuned hyperparameters, 0.76. Accuracy does increase with 0.01 to 0.87. Out of the eight reinterventions predicted using default hyperparameters, only three were correct, resulting in a precision of 0.38. Precision towards no reintervention reaches 0.88, with 174 correct no reintervention predictions of the 197. These predictions result in a recall of 0.97 towards no reintervention and of 0.12 towards reintervention. After tuning the hyperparameters, many less reinterventions were predicted. In total two, instead of eight. But precision towards reintervention increases from 0.38 to 0.50, because of the less reinterventions predicted. Out of two, one was correct and the other one was not. The decrease in recall is also caused by rise in no reintervention predictions. Earlier, three of the 26 reinterventions were correctly predicted as reintervention, with tuned hyperparamters only one.

Neural network

The neural network model is tuned on the following hyperparameters and resulting values and results (table 5.12):

- Hidden layer sizes: 5
- Batch sizes: 50
- Max iterations: 800

TABLE 5.12: Algorithm performance for neural network on imbalanced dataset - patient characteristics and process features

Using default parameters			**Using tuned parameters**	
accuracy		0.85		0.85
AUC		0.68		0.68
	0	1	0	1
F1 score	0.92	0.11	0.92	0.24
precision	0.88	0.22	0.89	0.33
recall	0.92	0.08	0.94	0.24

From the metrics accuracy and AUC, it does not seem like it, but using tuned hyperparameter settings improves the model. Both accuracy and AUC stay at their value with default hyperparameter settings, to be more specific: 0.85 and 0.68. At first nine reinterventions are predicted, of which two are correctly predicted. This results in precision towards no reintervention of 0.88 and towards reintervention of 0.22. Recall towards no reintervention is 0.92, due to 172 correctly predicted no reinterventions, and towards reintervention is 0.08, because of the two out of 26 correctly predicted reinterventions. After tuning, 15 reinterventions are predicted, of which five correctly predicted. Precision towards no reintervention increases with 0.01 to 0.89, and towards reintervention increases with 0.11 to 0.33. The recall towards no reintervention increases with 0.02 to 0.94, and recall towards reintervention triples to 0.24. This causes f1 score towards no reintervention to stay 0.92, and towards reintervention to increase from 0.11 to 0.24.

K-nearest neighbour

The k-nearest neighbour model is tuned on the following hyperparameters and resulting values and results (table 5.13):

- #neighbours: 50

TABLE 5.13: Algorithm performance for k-nearest neighbour on imbalanced dataset - patient characteristics and process features

	Using default parameters		Using tuned parameters	
accuracy		0.86		0.87
AUC		0.63		0.58
	0	1	0	1
F1 score	0.92	0.00	0.93	0.00
precision	0.87	0.00	0.87	0.00
recall	0.98	0.00	1.00	0.00

While fitting k-nearest neighbour with the default amount of neighbours, three reinterventions are predicted. These have all been wrongly predicted, causing f1 score, precision and recall towards reinterventions to equal 0.00. Towards no reinterventions, a f1 score, precision and recall of 0.92, 0.87, and 0.98 are obtained. After implementing 50 as the number of neighbours in the model, no reinterventions are predicted anymore. This gives recall towards no reintervention a value of 1.00 and towards reintervention a value of 0.00. Precision scores towards both no reintervention and reintervention stay the same and the f1 score towards no reintervention increases with 0.01 to 0.93. Accuracy increases with 0.01 to 0.87, but the AUC has worsened. Its value dropped from 0.63 to 0.58, mainly due to the ill-defined metrics.

5.2.3 Balanced dataset: Patient characteristics

Decision tree

The decision tree model is tuned on the following hyperparameters and resulting values and results:

- Max features: $\log_2 n$
- Max tree depth: 20
- Min samples split: 24
- Min samples leaf: 1

Table 5.14 shows the results of metrics for decision tree with patient characteristics as input data. All metrics improve using the tuned hyperparameters. While using default hyperparameters, 38 reinterventions are predicted, of which eight correctly resulting in precision towards reintervention of 0.21. These eight correctly predicted cases also result in a recall of 0.31 towards reinterventions. After tuning the hyperparameters and rerunning the model with these, accuracy increases wiht 0.03 to 0.80 and AUC increases with 0.14 to 0.71. Reinterventions are better predicted. In total 40 reinterventions get predicted, of which 12 correctly. This makes precision increasing to 0.30 and recall to 0.46. The f1 score increases from 0.25 to 0.36.

TABLE 5.14: Algorithm performance for decision tree on balanced dataset - patient characteristics

Using default parameters			Using tuned parameters	
accuracy		0.77		0.80
AUC		0.57		0.71
	0	1	0	1
F1 score	0.86	0.25	0.88	0.36
precision	0.89	0.21	0.92	0.30
recall	0.83	0.31	0.84	0.46

Random forest

The random forest model is tuned on the following hyperparameters and resulting values and results:

- Max tree depth: None
- #estimators: 300

TABLE 5.15: Algorithm performance for random forest on balanced dataset - patient characteristics

Using default parameters			Using tuned parameters	
accuracy		0.87		0.88
AUC		0.84		0.82
	0	1	0	1
F1 score	0.93	0.27	0.94	0.37
precision	0.89	0.45	0.90	0.58
recall	0.97	0.19	0.97	0.29

Tuning hyperparameters has a positive influence on almost every metric, except for the AUC, see table 5.15. With five correctly and 11 incorrectly reinterventions predicted with default hyperparameters, a f1 score of 0.27 towards reintervention is reached. Precision towards reintervention is 0.45 and recall 0.19. Tuning happens with the AUC as scoring metric, but this metric decreases from 0.84 to 0.82 after tuning. Accuracy increases with 0.01 to 0.88. The tuned correctly predicts seven reinterventions and incorrectly predicts five reinterventions. This cause the precision to increase with 0.13 to 0.58, the recall to increase with 0.10 to 0.29, and the f1 score as well to increase with 0.10 to 0.37.

Logistic regression

The logistic regression model is tuned on the following hyperparameters and resulting values and results (table 5.16):

- Max iterations: 700
- Penalty: L2

For the logistic regression model, tuning the hyperparameters does not seem to change any performance metrics. Even the AUC curve is exactly the same as with

TABLE 5.16: AUC's for random forest on balanced dataset - patient characteristics, with the default hyperparameters in blue and the tuned hyperparameters in orange.

Using default parameters			Using tuned parameters	
accuracy		0.87		0.87
AUC		0.73		0.73
	0	1	0	1
F1 score	0.93	0.48	0.93	0.48
precision	0.92	0.50	0.92	0.50
recall	0.93	0.46	0.93	0.46

the default hyperparameters. This is due to predictions being the same with default and tuned hyperparameters. In total, 168 no reinterventions are correctly predicted and 12 reinterventions are correctly predicted in both cases, resulting in the same confusion matrices.

Extreme gradient boosting

The extreme gradient boosting model is tuned on the following hyperparameters and resulting values and results (table 5.17):

- Max depth: 9
- #estimators: 90

TABLE 5.17: Algorithm performance for extreme gradient boosting on imbalanced dataset - patient characteristics

Using default parameters			Using tuned parameters	
accuracy		0.84		0.83
AUC		0.68		0.67
	0	1	0	1
F1 score	0.91	0.27	0.90	0.22
precision	0.89	0.32	0.89	0.26
recall	0.93	0.23	0.92	0.19

Tuning the hyperparameters has a negative effect on all of the performance metrics. Accuracy and AUC both drop with 0.01, comparing tuned hyperparameters with the default, to 0.83 and 0.67, respectively. Using default hyperparameters, six reinterventions are correctly predicted and 13 are incorrectly predicted. Using the tuned hyperparameters, only five reinterventions are correctly predicted, and 14 are wrongly predicted. This results in the precision to drop from 0.32 to 0.26, and the recall to drop from 0.23 to 0.19. F1 score decreases with 0.05 to 0.22.

Neural network

The neural network model is tuned on the following hyperparameters and resulting values and results:

- Hidden layer sizes: 10

- Batch sizes: 1
- Max iterations: 1300

TABLE 5.18: Algorithm performance for neural network on balanced dataset - patient characteristics

Using default parameters			**Using tuned parameters**	
accuracy		0.83		0.82
AUC		0.70		0.70
	0	1	0	1
F1 score	0.90	0.36	0.89	0.42
precision	0.91	0.34	0.93	0.35
recall	0.89	0.38	0.86	0.54

Although the accuracy slightly decreases with 0.01 to 0.82, tuning the hyperparameters has positive results for the neural network model, see table 5.18. At first, the model performs quite well, correctly predicting 10 reinterventions and wrongly predicting 19 reinterventions. This gives a precision of 0.34, a recall of 0.38 and a f1 score of 0.36 towards reinterventions. Tuning the model keeps the AUC at 0.70, but the predictions change. 14 reinterventions are correctly predicted and 26 are incorrectly predicted. The f1 score increases from 0.36 to 0.42, precision rises with 0.01 to 0.35 and recall increases from 0.38 to 0.54.

K-nearest neighbour

The k-nearest neighbour model is tuned on the following hyperparameters and resulting values and results (table 5.19):

- #neighbours: 5

TABLE 5.19: Algorithm performance for neural network on balanced dataset - patient characteristics

Using default parameters			**Using tuned parameters**	
accuracy		0.69		0.69
AUC		0.74		0.74
	0	1	0	1
F1 score	0.79	0.37	0.79	0.37
precision	0.95	0.25	0.95	0.25
recall	0.68	0.73	0.68	0.73

There is no difference in any of the performance measures of the model using default values or tuned hyperparameters. Accuracy and AUC stay 0.69 and 0.74. In both cases 19 reinterventions are correctly predicted, and 57 are wrongly predicted. 123 no reinterventions are correctly predicted versus seven incorrectly. The fact that the correlation matrices are the same, might be due to the tuned hyperparameter having the same value as the default hyperparameter. In both cases, five neighbours are used.

5.2.4 Balanced dataset: Patient characteristics and process features

Decision tree The decision tree model is tuned on the following hyperparameters and resulting values and results (table 5.20):

- Max features: $\sqrt{n}$
- Max tree depth: 25
- Min samples split: 16
- Min samples leaf: 1

TABLE 5.20: Algorithm performance for decision tree on balanced dataset - patient characteristics and process features

	Using default parameters		Using tuned parameters	
accuracy		0.79		0.76
AUC		0.52		0.52
	0	1	0	1
F1 score	0.88	0.15	0.86	0.14
precision	0.88	0.15	0.87	0.12
recall	0.88	0.15	0.84	0.15

Tuning the hyperparameters of the decision tree model causes most performance metrics to decrease. Using default hyperparameters, accuracy is 0.79 and the AUC is 0.52. With incorrectly predicting 22 reinterventions and as well incorrectly predicting 22 no reinterventions, f1 score, precision and recall towards no reintervention all equal 0.88 and towards reintervention all equal 0.15. Fitting the model with tuned hyperparameters slightly changes the correlation matrix. Again, only four reinterventions are correctly predicted, but 151 no reinterventions have been predicted, instead of 155 using the default parameters. This makes, apart from the AUC and recall towards reintervention, drop all performance metrics. Precision towards no reintervention drops to 0.86 and towards reintervention drops to 0.14. Precision towards no reintervention drops to 0.87 and towards reintervention drops to 0.12. Recall towards no reintervention drops to 0.84. The AUC stays 0.52 and recall towards reintervention stays 0.15, since again four reinterventions are correctly predicted and 22 reinterventions were predicted as no reintervention.

Random forest The random forest model is tuned on the following hyperparameters and resulting values and results (table 5.21):

- Max tree depth: None
- #estimators: 270

Using default hyperparameters, the random forest model performs not so good when it comes to prediction reinterventions. Accuracy is 0.86, because 176 out of 205 cases are correctly predicted, but these were all no reinterventions. The three predicted reinterventions, turned out to actually be no reintervention, causes f1 score, precision and recall towards reintervention all to equal 0.00. The AUC is 0.68, and f1 score, precision, and recall towards no reintervention equal 0.92, 0.87, and 0.98. Using the tuned hyperparameters, slightly increases the performance of the model.

TABLE 5.21: Algorithm performance for random forest on balanced dataset - patient characteristics and process features

Using default parameters			Using tuned parameters	
accuracy		0.86		0.87
AUC		0.68		0.70
	0	1	0	1
F1 score	0.92	0.00	0.93	0.07
precision	0.87	0.00	0.88	0.50
recall	0.98	0.00	0.99	0.04

Accuracy increases to 0.88 and the AUC to 0.70. Two reinterventions are predicted, of which one of them is correct, resulting in a precision towards reintervention of 0.50. With only one of the 26 reinterventions correctly predicted, recall towards reintervention equals 0.04 and the f1 score is 0.07. Towards no reintervention the metrics are higher; 0.88 for precision, 0.99 for recall and 0.93 for f1 score.

Logistic regression The logistic regression model is tuned on the following hyperparameters and resulting values and results (table 5.22):

- Max iterations: 1000
- Penalty: L2

TABLE 5.22: Algorithm performance for logistic regression on balanced dataset - patient characteristics and process features

Using default parameters			Using tuned parameters	
accuracy		0.85		0.86
AUC		0.74		0.73
	0	1	0	1
F1 score	0.92	0.31	0.92	0.33
precision	0.90	0.37	0.90	0.41
recall	0.93	0.27	0.94	0.27

Using the default hyperparameters for logistic regression results in seven correctly predicted reinterventions, 12 incorrectly predicted reinterventions, 167 correctly predicted no reinterventions and 19 reinterventions which had been predicted as no reinterventions. Performance metrics are good, with a f1 score of 0.31, a precision of 0.37 and recall of 0.27, all towards reintervention. The accuracy is 0.85 and AUC 0.74. After tuning the accuracy increases with 0.01 to 0.86 and the AUC decreases with 0.01 to 0.73. After using tuned hyperparameter, again seven reinterventions are correctly predicted, but now ten cases are predicted as reintervention, but in reality did not undergo a reintervention. Recall stays the same, but f1 score increases to 0.33 and precision to 0.41, both towards reintervention. From the metrics towards no reintervention, recall is the only one changing, with an increase of 0.01 to 0.94.

Extreme gradient boosting The extreme gradient boosting model is tuned on the following hyperparameters and resulting values and results (table 5.23):

- Max depth: 4
- #estimators: 20

TABLE 5.23: Algorithm performance for extreme gradient boosting on balanced dataset - patient characteristics and process features

	Using default parameters		**Using tuned parameters**	
accuracy		0.84		0.85
AUC		0.72		0.76
	0	1	0	1
F1 score	0.91	0.16	0.92	0.17
precision	0.88	0.25	0.88	0.30
recall	0.95	0.12	0.96	0.12

At first using default hyperparameters, the model performs quite good with an accuracy of 0.84 and an AUC of 0.72. Three reinterventions are correctly predicted, 23 reinterventions have incorrectly received the label 'no reintervention' while predicting and nine no reinterventions are incorrectly predicted as 'reintervention'. Using the model with tuned hyperparameters causes two of the incorrectly predicted as reinterventions now to be correctly predicted as no reintervention. Improvement is therefore made on precision towards reintervention. The metric increases in value from 0.25 to 0.30. Again, three reinterventions are correctly predicted, so the recall towards reintervention stays 0.12. Together, this increases the f1 score towards reintervention from 0.16 to 0.17, the accuracy from 0.84 to 0.85 and the AUC from 0.72 to 0.76.

Neural network The neural network model is tuned on the following hyperparameters and resulting values and results (table 5.24):

- Hidden layer sizes: 10
- Batch sizes: 100
- Max iterations: 1300

TABLE 5.24: Algorithm performance for neural network on balanced dataset - patient characteristics and process features

	Using default parameters		**Using tuned parameters**	
accuracy		0.85		0.84
AUC		0.64		0.68
	0	1	0	1
F1 score	0.92	0.38	0.91	0.40
precision	0.91	0.41	0.91	0.38
recall	0.93	0.35	0.90	0.42

Overall, the neural network model performs quite well. 166 no reinterventions have correctly been predicted, as well as nine reinterventions being correctly predicted. This results in an accuracy of 0.85 and an AUC of 0.64 and f1 score, precision, and recall of 0.92, 0.91, and 0.93 towards no reinterventions. Implementing

the hyperparameters results in labelling more cases as no reintervention than before. Precision towards reintervention decreases, since first, nine out of 22 predicted reinterventions were correct, while after implementing tuned hyperparameters, 11 out of 29 labelled reinterventions were actual reinterventions. Recall towards reintervention increases from 0.35 to 0.42, because cumulative a greater amount of reinterventions was labelled correctly. Accuracy drops with 0.01 to 0.84 due to 172 correctly predicted cases, instead of the earlier 175. But the AUC increases from 0.64 to 0.68.

K-nearest neighbour The k-nearest neighbour model is tuned on the following hyperparameters and resulting values and results (table 5.25):

- #neighbours: 5

TABLE 5.25: Algorithm performance for k-nearest neighbour on balanced dataset - patient characteristics and process features

	Using default parameters		**Using tuned parameters**	
accuracy		0.61		0.61
AUC		0.55		0.55
	0	1	0	1
F1 score	0.74	0.25	0.74	0.25
precision	0.90	0.16	0.90	0.16
recall	0.63	0.50	0.63	0.50

Again, the best number of neighbours to put in the hyperparameter is five, equalling the default value. Like in the previous experiment, with sampled data and only patient characteristics as input data, the models also perform the same. Accuracy stays 0.61, AUC stays 0.55, f1 score towards no reinterventions stays 0.74 and towards reintervention stays 0.25, precision towards no reintervention equals 0.90 both times and towards reintervention equals 0.16 both times, and the recall is 0.63 towards no reintervention and 0.50 towards reintervention. This is due to both correlation matrices being the same. In both cases, a reintervention was predicted 79 times, of which 13 were correct. 13 times no reintervention was predicted, while actually a reintervention took place, and 66 times a reintervention was incorrectly predicted.

5.3 Feature importances

The models with tuned hyperparameters are used to predict which patient characteristics and process features have the most impact on the outcomes of the Novasure surgery. The figures per experiment per algorithm can be found in Appendix D. Figure 5.17 provides an overview of the patient characteristics, perioperative features, and waiting time and their ranking per algorithm per experiment. For each of the model outcomes per experiment, a top 10 ranking is made by assigning the most influential factor a one, the second most influential factor a two, and so on until the tenth factor is assigned. The table is horizontally divided into four quarters. The first two quarters on the left show the rankings only based on patient characteristic, if it belongs to the ten most influential features according to the algorithm. The left half first presents the outcomes of the six models having the original dataset as input and then the outcomes with the sampled data as input. The third and fourth quarters of the table provide the ranking of the patient characteristics, the perioperative

features and the waiting time. Not all rankings between 1 and 10 have been filled, since those ranks are taken by one of the 258 appointments and care activities. A list of these appointments and care activies and their number of occurrences is presented in table 5.26. The darker the blue cell is coloured, the higher the ranking of the feature is and the more influence it delivers, according to the prediction. When a characteristic or feature has not been ranked 10 or higher in importance, the cell is left blank.

Figure 5.17 shows per patient feature per experiment at which rank between 1 and 10 the feature is ranked. *Age* occurs 11 times in the top 10 ranks. Its ranks differ from two to eight. Remarkable is that, once the data is balanced and process features are added, the characteristic does not occur anymore.

Almost the same counts for *BMI*. It occurs 13 times in the rankings, multiple times in the first three experiments, but once the data is balanced and process features are added, it only occurs once on rank 9.

The *complaints* are not that well presented in the rankings. *Benign adnexal abnormality* does not occur at all, *cycle disorder* four times, *uterine fibroids* three times and unknown complaints twice. None of the complaints occur in rankings from the experiments with sampled data.

Parity, on the other hand, is represented in the ranks from all four experiments. Especially with the data is sampled and patient features as input data all different options are at least present twice. With process features added, *parity* lowers in importance compared to other features, but most options occur in the ranking of neural network.

Cesarean section a patient feature which does not occur in any top 10 ranking.

The *uterus position* seems an important feature when only looking at patient features. All options occur three to six times in these experiment. When process features are also taken into account, the *anteverted uterus* occurs once at rank five and once at rank nine and *retroverted uterus* occurs once at position seven, but compared to the experiments with only patient features, the feature is a lot less present.

The same counts for *endometrial thickness*. In experiment 1 and 3, the feature occurs eight out of 12 times, on positions from four to nine, but is absent in the rankings of experiment 2 and 4, which also focus on process features.

Cavity length occurs ten out 24 times. Eight of these occurrences are in experiments focusing on patient characteristics and its ranks range from two to nine. *Cavity width* also occurs ten out of 24 times, but six of them are in patient characteristic focused experiments and four in the experiments also including process features. The ranks differ from two to ten.

Dymenorrhea occurs seven times in total in the ranks, ranging from fifth to tenth. The only time it makes the top 10 rank in experiments with process features is when the data is balanced. Using decision tree, it is ranked as fifth important features. *Endometriosis* occurs four times in the rankings, once at rank five, once at rank seven and twice at rank eight.

The non-interrupted blue line is from *adenomyosis*. Twenty out of 24 times it is ranked at position one and if not, the feature is still present in the rankings at position two, four or five.

Uterine fibroids occurs ten times, with rankings between two and ten. Notable is that with the original data and focusing on patient characteristics, it occurs in five of the six feature importance rankings.

Sterilised only occurs three times in the rankings. In both logistic regression models focusing on patient characteristics, it is ranked fourth. It is ranked as eighth

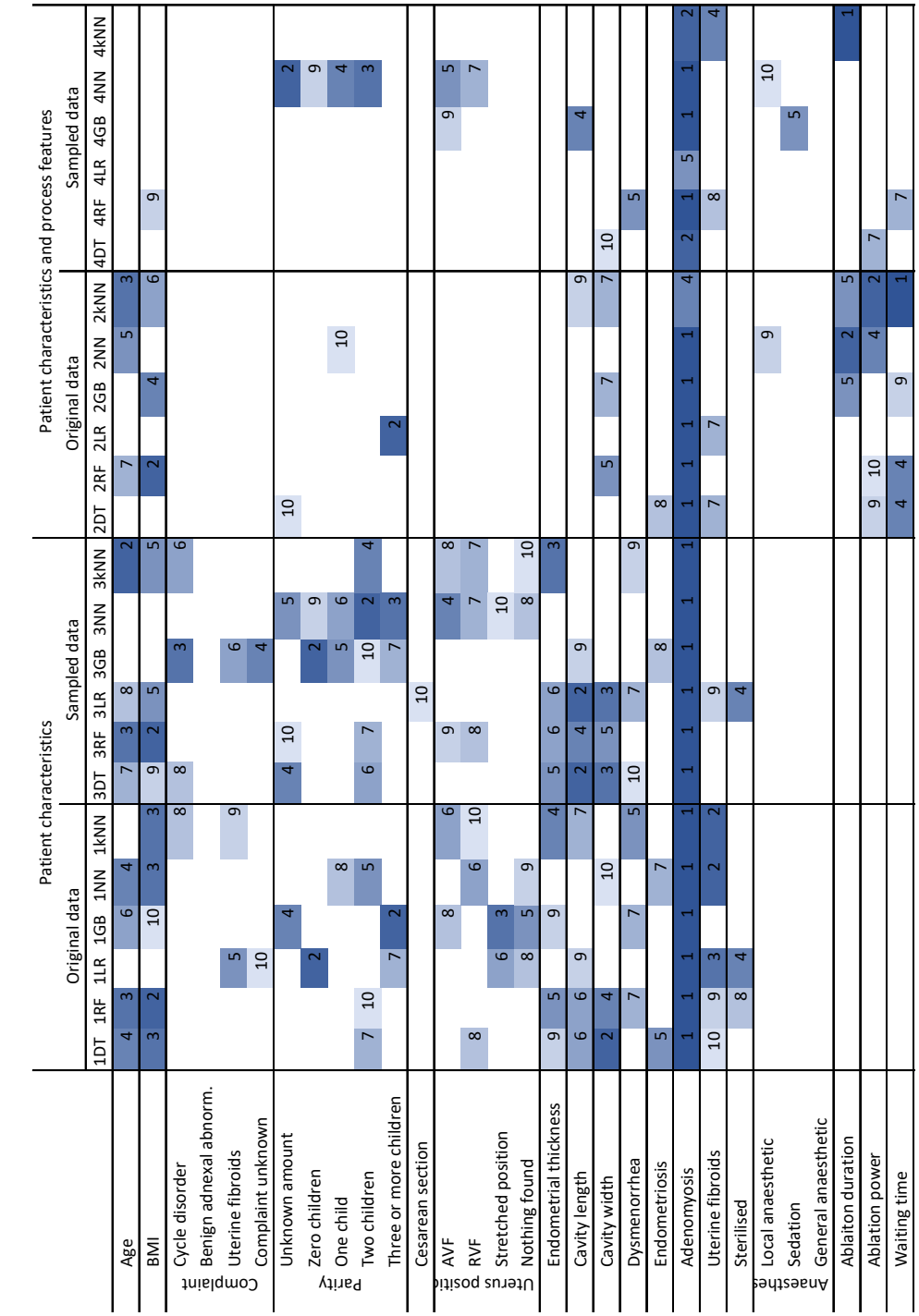


FIGURE 5.17: Top 10 ranking for patient characteristics and process features

important feature while utilising random forest with the original data focusing on patient characteristics.

Overall, the *anaesthetic* does not seem to be an important feature. *Local anaesthetic* is once ranked ninth and once tenth, and *sedation* is once ranked fifth.

Ablation duration is once ranked at position one, once at position two and twice at position five. Of course, it only occurs in the experiments with process features, but it is more present (three times) when the original data is used than when the data has been sampled (one time). *Ablation power* occurs five out of 12 times and its rankings range from two to ten.

Last, *waiting time* occurs five out of 12 times as well. Once, it is positioned as the most influential factor by k-nearest neighbour. Apart from that, the feature is twice ranked fourth, once seventh and once ninth.

TABLE 5.26: Alphabetic list of influential appointments and care activities

Appointment or care activity	**Number of occurrences**
11/21 operatief kliniek 404	2
Abnormaal bloedverlies	1
Beoordeling ecg holter inspanningsonderzoek ed	3
Controle patient	2
Controle patient bekkenbodemzorg 1	
Controle patient gynaecologie	3
Dagverpleging	3
Dagverpleging i	3
Diagnostische hyteroscopie inclusief eventuele proefexcisie(s) en/of inclusief eventuele endometriumbiopsie(en) en/of het verwijderen van een enkelvoudige poliep voor pathologisch onderzoek	7
Doelgerichte telefonische consultatie van een poortspecialist door een patiënt bij een al geopende dbc ter vervanging van een fysiek consult	5
Dotcode zorgtype 41 51 en 52	2
Echo gynaecologie via 1e lijn	2
Echo op verzoek huisarts gynaecologisch	2
Echografie van de buikorganen	2
Echoscopie gynaecologisch	5
Eerste polikliniekbezoek	2
Geen uitval standaard cyclusstoornissen geen intensieve/invasieve therapie geen klinische opname specifieke overige ingrepen	1
Geen uitval standaard cyclusstoornissen geen intensieve/invasieve therapie geen klinische opname geen specifieke overige ingrepen geen ambulant middel/ dag ambulant	2
Geen uitval standaard cyclusstoornissen intensieve/invasieve therapie geen oper groep 3 geen oper groep 2 open oper groep 2 endoscopisch geen oper groep 1 diagnostisch specifiek/ gynaecol	2

Continued on next page

Table 5.26 – continued from previous page

Appointment or care activity	Number of occurrences
Geen uitval standaard geen cyclusstoornissen geen ontstekingsprocessen vrouwelijke organen in bekken uterus en adnex intensieve/ invasieve therapie geen oper groep 3 geen oper groep 2 oper groe	1
Geen uitval standaard geen cyclusstoornissen ontstekingsprocessen vrouwelijke organen in bekken geen intensieve/ invasieve therapie geen klinische opname geen specifieke overige ingrepen geen am	2
Geen uitval standaard geen cyclusstoornissen ontstekingsprocessen vrouwelijke organen in bekken intensieve/invasieve therapie geen oper grope 2 geen oper groep 1 diagnostisch specifiek/gynaeco	1
Geen uitval standaard geen intensieve/ invasieve therapie geen klinische opname geen oper licht diagnostisch (zwaar)/ therapeutisch licht	1
Hemoglobine (incl (eventueel) hematocriet en celindices (mcv mch en mchc en erytrocyt)) 2	
Herhalingsbezoek	2
Herhaalpoliklinkiekbezoek	10
Microcurettage pipelle	1
Nieuwe patient abnormaal bloedverlies	4
Nieuwe patient plaatsen iud	1
Plaatsen spiraal (iud ius)	1
Poliklinische behandeling	2
Preassessment dagopname anesthesiemedewerker	1
Sis echo	3
Spoed	1
Spoedeisende hulp contact buiten de seh afdeling elders in het ziekenhuis	4
Telefonisch consult	10
Telefonisch consult poli medewerker	1
Telefoonpatient artsassistenten	1
Uteruscuretteage exclusief diagnostische microcurettage (endometriumsapmling zoals pipell vabra milex novak)	3
Zorgdomein abnormaal bloedverlies	2

The appointments and care activities listed in table 5.26 occurred once or more as one of the fifteen most influential process features. The table presents the appointments and care activities in alphabetical order and provides information on how many times they were one of the fifteen most influential process features in one of the algorithms. The appointments and care activities have been kept in Dutch, since the scope of this research includes answering the question on how predicting results are influenced by including process features. In total, 38 unique features occurred 136 times in the top 15 ranks. While most of them occur once or twice, some occurrences are remarkable. *Herhaalpolikliekbezoek* and *telefonisch consult* both make

it to 10 out 12 lists. The only patient characteristics occurring more often, are *adenomyosis (24)*, *age (14)*, *amount of children unknown (11)*, *BMI (15)*, *cavity length (12)*, *cavity width (13)*, *dysmenorrhea (11)*, *endometrial thickness (11)*, and *myomatosus (16)*, but these patient characteristics are used as input data in 24 models, instead of 12. Also, '*Diagnostische hyteroscopie inclusief eventuele proefexcisie(s) en/of inclusief eventuele endometriumbiopsie(en) en/of het verwijderen van een enkelvoudige poliep voor pathologisch onderzoek* occurs quite often, with seven times, even as *Doelgerichte telefonische consultatie van een poortspecialist door een patiënt bij een al geopende dbc ter vervanging van een fysiek consult* and *Echoscopie gynaecologisch*, which both occur five times. Three different appointments and care activities start with the word *Controle* and four start with the word *Echo*. The phrase *'Geen uitval standaard'* is followed seven times by mostly the same phrases being positive or negative, depending on the presence of the word '*geen*'. It looks like this is a template to fill in and the cumulative presence of 10 times in the rankings points that this phrase provides quite interesting information to investigate more.

Chapter 6

Discussion

This chapter discusses the results presented in chapter 5. First tuning the hyperparameters is addressed. For each of the algorithms (at least) one of the hyperparameters is further investigated. These tuned hyperparameters are implemented in the estimator and predictions are made. Then, the performances of the algorithms with different data inputs are discussed. The third part of this chapter focuses on the feature importances, comparing them with literature and expert opinion.

6.1 Hyperparameters

For each of the algorithms, the different tuned hyperparameters per experiment are presented and compared to each other.

6.1.1 Decision tree

TABLE 6.1: Tuned hyperparameters for decision tree, with PC for patient characteristics and PF for process features.

		Original data		**Balanced data**	
Prediction model	**Hyperparameter**	PC	PC & PF	PC	PC & PF
DT	Max features	$log_2 n$	$\sqrt{n}$	$log_2 n$	$\sqrt{n}$
DT	Max tree depth	15	20	20	25
DT	Min samples split	16	24	24	16
DT	Min samples leaf	19	9	1	1

For the decision tree model the maximum number of features is twice the square root of the features and twice the log is taken (table 6.1). Recalling figure 5.2, this does not seem odd, since the values create an AUC quite close to each other. It could be that with the changing circumstances has an influence on this. The trend which can be found is, when only patient characteristics are taken as input data, the square root is taken, and with including process features, the log is taken. This results in a maximum number of features of $log_2(24) \approx 4.58$ for experiments with patient characteristics and a maximum number of features of $\sqrt{291} \approx 17.06$. The maximal tree depth ranges from 15 to 25, which is the higher half of the tuning range. Also, the more input data or feature, the higher the maximum tree depth gets. Recalling figure 5.1, it might be interesting to investigate whether a tree depth higher than 25 or a range from 15 to 25 with smaller steps than five performs better. Another option is to optimise the maximum tree depth, while also taking the minimum samples to split a node into account, since they have great influence on each other (Harrison, 2019). Also for the tuned minimum samples split, no sure trend can be found. This

might come due to the still increasing test score, see figure ref **??**. On the other hand, the minimum samples to form a leaf has reached its optimum and is only decreasing in figure 5.3. Still, for the original dataset, more samples are needed to form a leaf, contrary to the sampled dataset where in both cases a leaf of one seems to be the optimum.

6.1.2 Random forest

TABLE 6.2: Tuned hyperparameters for random forest, with PC for patient characteristics and PF for process features.

		Original data		**Balanced data**	
Prediction model	**Hyperparameter**	PC	PC & PF	PC	PC & PF
RF	Max tree depth	7	4	None	None
RF	Number of estimators	300	210	300	270

Table 6.2 shows the optimum per hyperparameter per experiment. For maximum tree depth, there is a split in the values for original data and balanced data. Where the original data has a numerical maximum tree depth, the balanced data performs better when there is no maximum tree depth. Since the upper bound of range of the maximum depth equal 11, it might be that somewhere between 11 and no limit lies the perfect boundary. Like with decision tree, it also might been interesting to include the minimum samples to split a node, since they influence each other (Harrison, 2019). It is not clear where this different comes from, but it seems like the model only needs a flat tree when there a majority belongs to one category and a lot of depth when there is a 50% chance that the data point belongs to one category or the other. It did not seem like it from figure 5.6, but the range of number of estimators should have been increased, since the maximal number of estimators for two of the four experiments is the upper bound of the tuning range. This figure does not show a clear peak, so investigating in the number of estimators more deeply, would be interesting.

6.1.3 Logistic regression

TABLE 6.3: Tuned hyperparameters for logistic regression, with PC for patient characteristics and PF for process features.

		Imbalanced data		**Balanced data**	
Prediction model	**Hyperparameter**	PC	PC & PF	PC	PC & PF
LR	Max iterations	100	100	700	1000

During the default run, 500 iterations was the minimum to run the model without warnings. While tuning the models, it is chosen as the best value for maximum numbers of iterations for three of the four models, being the lower bound of the range (table 6.3). Still, a range from 100 to 1000 with steps of 100 is chosen. While for the first two experiments, 100 iterations are the optimum value for the hyperparameter, the model needs more iteration when sampling the input data. Taking process features into account, resulting in 291 features, provides the maximum value as optimum. When recalling figure A.2, we can conclude that, despite the warning,

the amount of iterations should be somewhere near zero at it's lowest. Investigating in what smaller steps would do to the AUC, could be interesting. From figure **??** can also be concluded that doing more iterations would not improve the AUC or do harm. It would only take more time and memory. As earlier mentioned, the penalty hyperparameter has been removed from the tuning hyperparameters, since the solver of the logistic regression was kept the default and only works with L2-penalty.

6.1.4 Extreme gradient boosting

TABLE 6.4: Tuned hyperparameters for extreme gradient boosting, with PC for patient characteristics and PF for process features.

		Original data		**Balanced data**	
Prediction model	**Hyperparameter**	PC	PC & PF	PC	PC & PF
XGBoost	#estimators	10	10	90	20
XGBoost	Max depth	None	2	9	4

Extreme gradient boosting has been tuned on two hyperparameters: the number of estimators and the maximum tree depth, see table 6.4. The number of estimators fluctuates for the four experiments. For the original dataset, it overall needs a low number of estimators of ten. For the balanced input data it once needs 90 estimators and once 20 estimators. No trend can be found one whether the amount of input data or the amount features has an influence on that. The same counts for the maximum tree depth. With values of *None*, 2, 9, and 4, it looks like the less features, the deeper the trees need to be, but this cannot be concluded with too much confidence. Both the fluctuations in hyperparameters between the different experiments might point out that the global optimum has not been found, and that further investigation, may it be extending the ranges of decreasing the steps of the ranges, is recommended.

6.1.5 Neural network

TABLE 6.5: Tuned hyperparameters for neural network, with PC for patient characteristics and PF for process features.

		Original data		**Balanced data**	
Prediction model	**Hyperparameter**	PC	PC & PF	PC	PC & PF
NN	Hidden layer sizes	5	5	10	10
NN	Batch sizes	1	50	1	100
NN	Max iterations	1300	800	1300	1300

The tuning results for the neural network hyperparamters are presented in table 6.5. The hyperparameter for hidden layer size has not reached its optimum at ten, according to the curve in figure 5.11. While the original dataset settles with five as optimum hidden layer size, the balanced data sheet needs at least ten, as seen in table 6.5. Although hyperparameter tuning for the first experiment resulted in a (local) optimum of five, it turned out to not be enough every time. It would be interesting to investigate whether adding more hidden layers would make a significant difference. The batch sizes fluctuate between one and 100, reaching over the whole chosen range. Apart from the models having patient characteristics as input data preferring

a batch size of one, no actual trend can be found between these batch sizes. The trend in maximum number of iterations is clear: the upper bound of the range has been found as optimum for three out of four times. The number of iterations should increase, as lightly confirmed by figure 5.13. It is unclear where the variations for all three hyperparameters come from and it would be, if it was not for time's sake, interesting to investigate even further than in the experiments done before.

6.1.6 K-nearest neighbour

TABLE 6.6: Tuned hyperparameters for k-nearest neighbour, with PC for patient characteristics and PF for process features.

		Original data		**Balanced data**	
Prediction model	**Hyperparameter**	PC	PC & PF	PC	PC & PF
kNN	#neighbours	25	50	5	5

For k-nearest neighbour, only one hyperparameter is tuned: the amount of neighbours (table 6.6). For the original dataset with patient characteristics as input data, the optimal value for neighbours corresponds with the tuned value of 25, as seen in figure 5.15. The amount of neighbours increases to 50, when adding process features. It seems like having imbalanced data causes noise in predicting the class. For both experiments using sampled data, the hyperparameter is set to 5 neighbours, which according to Tatsat, Puri, and Lookabaugh (2020) is in the range of good values (1 to 20 neighbours). One clarification for this might be sampling using SMOTE. Specific data points from the minority class have been duplicated, so it is possible that one or more "neighbours" are found with a total distance of 0, for the data point of which the outcome is predicted. This could result in needing less neighbours to accurately predict the outcome of the point to be predicted.

6.2 Algorithm performance

After the hyperparameters have been tuned, the algorithms are run and criticised on their performances. Recall that the algorithms are tuned on AUC and that based on AUC and accuracy, the best algorithms is determined. The section covers and repeats the results found in section 5.2 and the figures found in appendix B. This section is written towards reinterventions. Therefore all f1 scores, precisions and recall are towards reinterventions.

6.2.1 Original dataset: Patient characteristics

First, the AUC's for all models on patient characteristics with the original dataset as input data is shown in figure 6.1.

TABLE 6.7: AUC and accuracy for patient characteristics based on original dataset

	DT	RF	LR	GB	NN	kNN
AUC	0.77	0.72	0.78	0.73	**0.82**	0.61
Accuracy	**0.88**	0.87	0.86	0.87	0.87	0.87

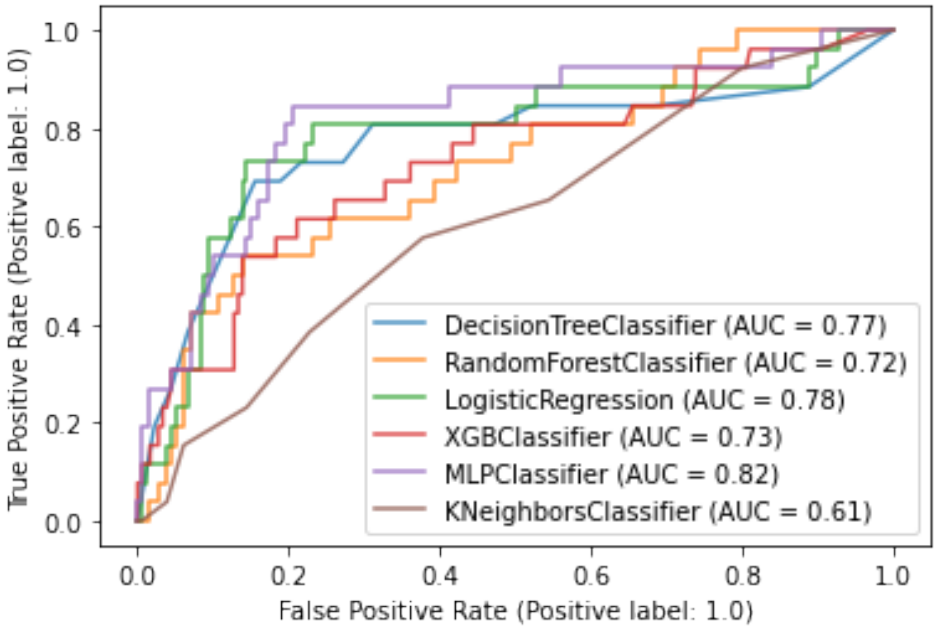


FIGURE 6.1: AUC's for algorithms on original dataset - patient characteristics

The AUC and accuracy per model for experiment 1 are shown in table 6.7. Training with the AUC as scoring metric results in neural network having the highest AUC of 0.82. Remarkable is that this AUC is achieved, even though the model has been ill-defined. The model labelled all cases as no reintervention, despite that the test data set exists of 26 reintervention cases. With an ill-defined model comes an accuracy of 0.87, being the percentage of test data with the label 'no reintervention'. While both random forest and k-nearest neighbour also have an accuracy of 0.87, only k-nearest neighbour is as well ill-defined. For random forest the accuracy of 0.87 is due to one case predicted as reintervention, while it was actually labelled as 'no reintervention'. This patient in question is 41 year old patient, who had cycle disorder as the complaint leading to the Novasure, a BMI of 24.2, one child, and suffered from dysmenorrhea, endometriosis, myamatosus, but not from adenomyosis. In this experiment, decision tree has the highest accuracy of 0.88, due to 184 correctly predicted no reinterventions and five correctly predicted reinterventions. It also has the highest precision (0.56), recall (0.19) and therefore also f1 score (0.29) from the six models in this experiment. The precision of of 0.56 is achieved by predicting nine reinterventions, of which five correctly.

6.2.2 Original dataset: Patient characteristics and process features

Figure 6.2 shows the AUC's for the experiment using the original data using patients and process features as prognostic factors.

TABLE 6.8: AUC and accuracy for patient characteristics and process features based on original dataset

	DT	RF	LR	GB	NN	kNN
AUC	0.56	0.74	0.73	0.76	**0.85**	0.58
Accuracy	**0.87**	**0.87**	0.85	**0.87**	0.68	**0.87**

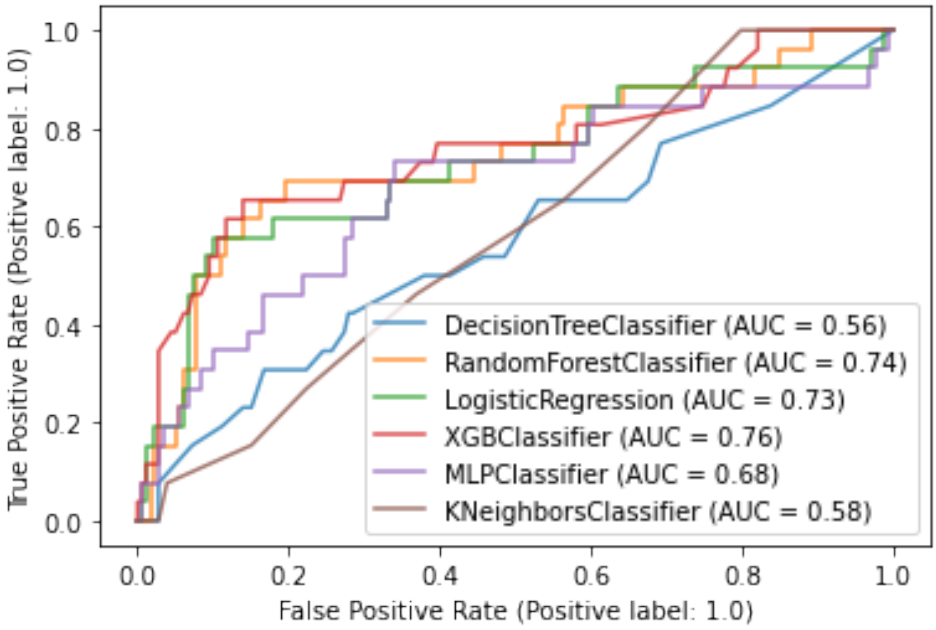


FIGURE 6.2: AUC's for algorithms on original dataset - patient characteristics and process features

The AUC values for the second experiment range from 0.56 to 0.85. Neural network reached the highest AUC, although this was no improvement compared to the model with the default hyperparameter values (recall table 5.12). Also, the accuracy of the neural network did not change, while the the f1 score, precision and recall did due to different predictions. After tuning, more reinterventions were predicted, both correctly and incorrectly, causing the increase of the f1 score, precision and recall.

The only model with an improving AUC was random forest. The model got ill-defined twice, due to no reintervention predictions. Still, the AUC increased with 0.02 and therefore has become a slightly better model with tuned hyperparameters.

Some other models reached the same AUC after tuning the hyperparameters, as before tuning the hyperparameters. These were decision tree and extreme gradient boosting. For decision tree the ill-defined model worked as good as the model using default hyperparameter, taking the AUC into account. At first, seven reinterventions were correctly predicted, while 26 cases were incorrectly labelled as reintervention. The change in this predictions caused an increase in accuracy, making the ill-defined model perform better than the model with predictions in both classes.

Gradient boosting does not suffer from an ill-defined model, but also after tuning this model, the amount of reinterventions predicted decreases from eight to two in total. From those two, one of them was correctly predicted, making precision increase to 0.50.

The AUC's of logistic regression and k-nearest neighbour even decreased. Noteworthy is that the highest accuracy received is 0.87. This is mainly due to the ill-defined models: decision tree, random forest and k-nearest neighbour. Extreme gradient boosting also predicts 179 cases right, except in this case one of them being a correctly predicted reintervention. The lower accuracy of logistic regression is due to 175 correctly predicted cases in total: 170 no reinterventions and five reinterventions. Nine cases are incorrectly labelled as reintervention, causing prediction to be 0.36. Based on accuracy, all ill-defined models and extreme gradient boosting perform best, but logistic regression and neural network predicted both predict for

five cases correctly that they would undergo a reintervention, resulting in higher f1 scores and higher recalls than the other models.

6.2.3 Balanced dataset: Patient characteristics

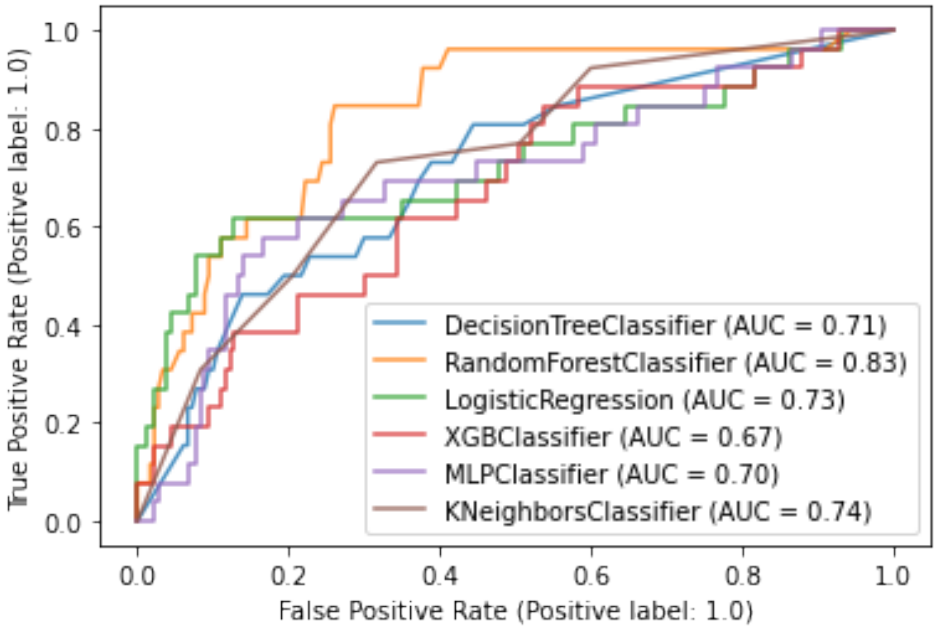


FIGURE 6.3: AUC's for algorithms on balanced dataset - patient characteristics

In experiment 3, the input data has been sampled, resulting in two classes with an equal amount of samples. The results are presented in figure 6.3 and table 6.9.

TABLE 6.9: AUC and accuracy for models with patient characteristics of balanced dataset

	DT	RF	LR	GB	NN	kNN
AUC	0.71	**0.82**	0.73	0.67	0.70	0.74
Accuracy	0.80	**0.88**	0.87	0.83	0.82	0.69

Different to the earlier done experiments, none of the models is ill-defined. On both AUC and accuracy, random forests performs best, compared to the other models in this experiment. The accuracy is higher than the accuracy for ill-defined models. In total, the random forest model predicted 182 cases right, instead of 180. From the 182 correctly predicted cases, seven were predicted as reinterventions. The accuracy of logistic regression equals the accuracy of ill-defined models, but the model is not ill-defined. According to predictions, 24 cases would undergo a reintervention, while in reality this was true for 12 of the 24. Adding these 12 to the 168 correctly predicted no reinterventions, makes 180 correct predictions and creates an accuracy of 0.87. Precision for logistic regression is 0.50 due to half of the predicted reinterventions being true. A f1 score of 0.48 is achieved, which is the highest f1 score over all experiments. Decision tree, extreme gradient boosting and neural network have accuracy values between 0.80 and 0.83. The number of correctly predicted

reinterventions fluctuate (12, 5, 14, respectively), as well as the as reintervention labelled cases (28, 14, 26). Outstanding is the relatively low accuracy from k-nearest neighbour. The predictions are the same as k-nearest neighbour using default hyperparameters, since the default hyperparameters equal the tuned hyperparameter. In total, 76 reintervention labels were predicted, but only 19 of them are right. Seven cases had incorrectly been labelled as no reintervention and 57 cases had incorrectly been labelled as reintervention. This causes the low accuracy. But despite that low accuracy, the model achieves the highest recall of all experiments. Out of the 26 actual reinterventions, 19 have been correctly predicted.

6.2.4 Balanced dataset: Patient characteristics and process features

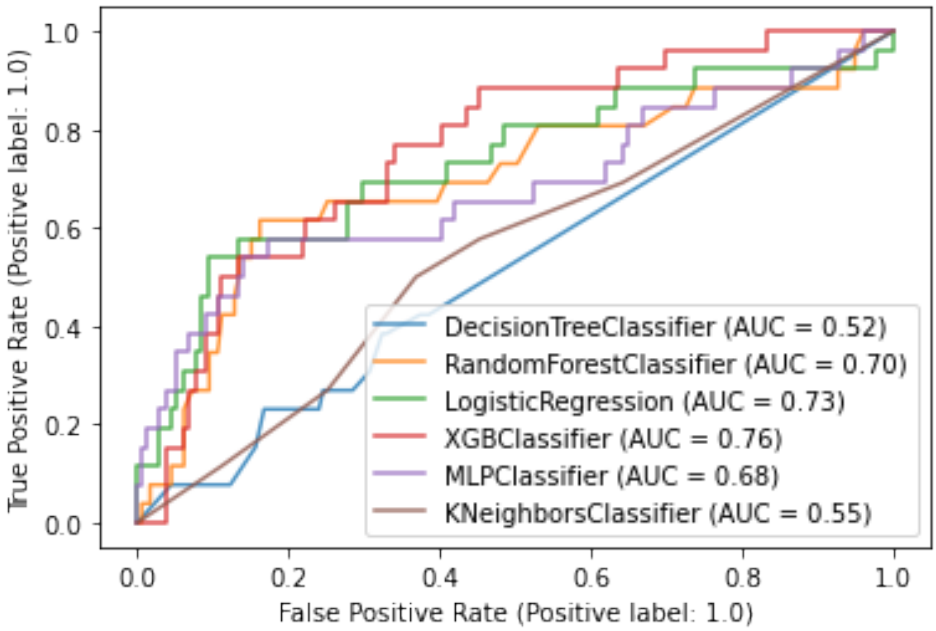


FIGURE 6.4: AUC's for algorithms on balanced dataset - patient characteristics and process features

Figure 6.4 shows the AUC's for the sampled dataset including both patient characteristics and process features. Compared to experiment 2, also taking process features into account, but not having sampled test data, none of the models in experiment 4 performs better looking at the AUC's. The AUC's of decision tree, random forest and k-nearest neighbour are lower than in experiment 2. The AUC's of logistic regression, extreme gradient boosting and neural network equal the AUC values of experiment 2. Half of the accuracies, however, is higher compared to experiment 2.

TABLE 6.10: AUC and accuracy for patient characteristics and process feature of balanced dataset

	DT	RF	LR	GB	NN	kNN
AUC	0.52	0.70	0.73	**0.76**	0.68	0.55
Accuracy	0.76	**0.87**	0.86	0.85	0.84	0.61

The accuracy of random forest is 0.87, being the highest of the experiment. In total, two reinterventions are predicted, one being correctly labelled and one incorrectly. This creates a precision of 0.50, but a recall of 0.04, and therefore a f1 score of 0.07.

Decision tree performs not as high as most other models in this experiment. The AUC sticks at 0.52 and the accuracy equals 0.76. Of the in total 161 correctly predicted labels, only four of them were reinterventions and the tuned model performs worse than the model using default hyperparameters. The tuned model predicts four correct reinterventions, but also increased the amount of incorrectly labelled reinterventions to 28 cases, while they are not.

Extreme gradient boosting receives the highest AUC in this experiment, and a thereby coming an accuracy of 0.85. The accuracy is because 172 no reinterventions and three reinterventions were predicted correctly. In total, this is no improvement compared to experiment 2 (same data input, but no sampled train data), but more reinterventions are predicted.

Both logistic regression and neural network also reach a quite high accuracy of 0.86 and 0.84. For logistic regression 169 no reinterventions and seven reinterventions were predicted correctly and for neural network 161 no reinterventions and 11 reinterventions were predicted correctly. Compared to the same models in experiment 2, more reinterventions are correctly predicted, but at small cost of correctly predicted no reinterventions. The judgement on in which experiments perform better, depends on the goal the prediction. When the main goal is to correctly predict all the patients who will be satisfied with their Novasure surgery and who will not undergo a reintervention, experiment 2 performs better. If the goal is to warn patients that they are likely to not be satisfied with their Novasure, experiment 4 performs better.

Again and due to the same reason as before, the predictions for the k-nearest neighbour mode using the tuned hyperparameters are the same as the predictions using the default hyperparameters. In total, 79 reintervention labels were predicted, but only 13 of them are right. 13 cases had incorrectly been labelled as no reintervention and 66 cases had incorrectly been labelled as reintervention. It means that accuracy stays 0.61, the AUC stays 0.55, f1 score stays 0.25, precision stays 0.16, and recall stays 0.50. This all, because both the default and tuned hyperparameter use the five nearest neighbours.

6.2.5 Algorithm performance

Overall, there is no one way to conclude that one model performs best. Four experiments have been carried out and in most cases, the models do not perform exactly the same as in the other experiments. For example, where decision tree has the highest accuracy in experiment 1, extreme gradient boosting has the highest accuracy in experiment 2 (apart from the ill-defined models also having this accuracy) and random forests has the highest accuracy in experiment 3 and 4. Neural network performs best on AUC in experiment 1 and 2, but random forest has the highest AUC in experiment 3 and extreme gradient boosting in experiment 4.

Tuning the hyperparameters was done with the AUC as scoring metric, instead of accuracy. The reason behind this, is that accuracy in these experiments is misleading. There are only two cases, in which a higher accuracy than 0.87 has been reached. In the first experiment with the original data set en focused on patient characteristics, decision tree achieved an accuracy of 0.88 with 176 correctly predicted no reinterventions and five correctly predicted reinterventions. The same accuracy is

reached by random forest in experiment 3, also focusing on patient characteristics, but with sampled test data. In this case, 175 no reinterventions were correctly labelled, as well as seven reinterventions. One of the main intentions of the research is to find out which patients are susceptible on undergoing a reintervention. In that case, an accuracy of 0.87 with predicted reinterventions should not be one of the well-performing models. The next subsection looks at what influence f1 score has as the scoring metric.

TABLE 6.11: Comparison of AUC's for random forest and logistic regression to Stevens et al. (2021) with S for Stevens et al. (2021) and H for the results found in this research.

Experiment number	**Random forest** S	H	**Logistic regression** S	H
1	0.65	0.72	0.71	0.78
2	-	0.74	-	0.73
3	-	0.82	-	0.73
4	-	0.70	-	0.73

Table 6.11 compares the AUC's obtained in the research by Stevens et al. (2021) and to the AUC's found in this research. Stevens et al. did not calculate accuracies, since their database has an imbalanced ratio of 1:8. Two points should be noted. While in the research by Stevens et al, logistic regression performs better than random forest, this is never the case in this research. Based on the AUC, random forest performs better than logistic regression. Also, both random forest and logistic regression in this research reach a higher AUC than the models did in the the research by Stevens et al. This softens their conclusion that machine learning algorithms do not perform better than predicting with logistic regression.

6.3 Ill-defined cases

Five out of 24 predictions using a model with tuned hyperparameters are ill-defined. They predict that all patients do not need to undergo an intervention, while in reality always 26 of the patients from the test data are reintervention cases. One problem with ill-defined models is that due the data structure, the accuracy is always 0.87, which is the second highest accuracy reached by all models. One possibility to circumvent this, is to use f1 score instead of AUC as scoring metric, while tuning.

As a try-out, the experiment is redone on the five ill-defined models, but with f1 score as scoring metric this time. This includes the ill-defined models from experiment 1 (neural network and k-nearest neighbour) and experiment 2 (decision tree, random forest and k-nearest neighbour). The exact results are presented in Appendix E. The try-out experiment included tuning the hyperparameters in the exact same way as during the previous experiments, expect the scoring metric AUC is replaced with f1 score. This results in none of the five models being ill-defined again. The accuracies of the five models range between 0.81 and 0.87, which is still in line with experiments using the AUC as scoring metric. The same counts for the resulting AUC's, which range between 0.50 and 0.77. Only for k-nearest neighbour with patient characteristics and process features as input data the f1 score, precision and recall are again all 0.00. As seen in previous experiments, with 5 neighbours as tuned

hyperparameter, the results for the tuned models are the same as for the model using default hyperparameters. In these cases, three reinterventions were predicted, but incorrectly.

For all the models, using f1 score as scoring metric seemed to improve the results. With the purpose of not predicting what the outcome of the Novasure surgery would be, but more specifically predicting which cases are likely to undergo a reintervention, it might be more interesting to tune hyperparameters with f1 score as scoring metric. The focus is than laid on improving reintervention predictions, instead of solely predicting the most cases correctly. For further research, this might be interesting to investigate.

6.4 Sampling

The last two experiments had different input data than the first two experiments. The training data was split in the two classes there are and the minority class of reintervention is enlarged such that the two classes are of the same size. This resulted in no more ill-defined models. In each of the cases, reinterventions were predicted. Instead, 380 reinterventions are predicted in experiment 3 and 4 in total versus the 68 reinterventions predicted in experiment 1 and 2, without sampled data. The ratio of true positives decreases slightly from 0.38 (26 out of 68) to 0.28 (108 out of 380). Table 6.12 shows how for both experiments, the average accuracy and AUC decrease, and the f1 score, precision and recall increase when the training data is sampled.

TABLE 6.12: Average f1 score, precision and recall per experiment, with PC for patient characteristics and PF for process features

	Original data		**Balanced data**	
Performance metric	PC	PC and PF	PC	PC and PF
accuracy	0.87	0.84	0.82	0.80
AUC	0.72	0.64	0.73	0.66
f1 score	0.18	0.15	0.37	0.24
precision	0.36	0.26	0.38	0.34
recall	0.12	0.12	0.45	0.26

To conclude, for the purpose of finding patient who are likely to undergo a reintervention, it is recommended to sample the training data. It results in more reinterventions being predicted, than happens without sampling the training set.

Currently, the strategy only includes oversampling the minority case so that it has the same size as the majority case, but other oversampling strategies can be tried, as well as including an undersampling strategy.

6.5 Feature importances

In this section the most highly and often ranked patient characteristics and process features are discussed. First, the most highly ranked feature importances are summarised in table 6.13, after which they are discussed more deeply. The discussion is set up with help of medical experts and literature. The paper *Prognostic Factors for the Failure of Endometrial Ablation* by Beelen et al. (2019) is used as a starting point and extended with additional papers when needed. This research provides an overview

of prognostic factors predicting failure of endometrial ablation, having considered 990 studies.

Table 6.13 lists the most important features. All features, if occurring in the experiment, have been ranked from 1 to 15. If the feature occurred as input in the experiment, but was not ranked 15 or higher, it was assigned with 16. Then the total average per feature is calculated. Table 6.13 first names the feature, then the relative position in ranking compared to other features and behind that the total average rank is presented. The important features are discussed in order of raking, except of the appointments and care activities. These phrases are in Dutch and presented in italics. These phrases are discussed at the end of the chapter together. For the full names of the shortened phrases, we refer to appendix A.3.

6.5.1 Adenomyosis

Adenomyosis is the most influential feature according to this research. It is ranked with position 1 by 20 of 24 predictions and is also present in the top 10 for the other four models. The average rank is 1.8, which is enormously high, especially compared to the follow-up average rank of 8.4. According to the expert, this is an interesting result, since the existence of *adenomyosis* is a controversial variable for several reasons. First of all, in 863 cases it has not been checked whether there might be *adenomyosis*. Purely looking at this definition, it does not provide us any information. *Adenomyosis* is a complaint hard to find since it is located within the uterus wall and even when there is a cause for its existence, this cannot be guaranteed for a hundred percent. Also, a Novasure does not cure this complaint. The device is not able to reach the *adenomyosis* in the uterus wall and therefore, it is not removed after the Novasure surgery. But on the other side, when there are no indications to check whether *adenomyosis* is present, it probably does not occur at the patients, since they do not suffer from it. In literature, *adenomyosis* is mentioned as one of the importance causes of *dysmenorrhea*, which was found to have the most strongly correlated risk factor for receiving a reintervention (Beelen et al., 2019). In this this research, *dysmenorrhea* is ranked as 15^{th} with an average rank of 14.3. McCausland and McCausland (1998) advice to offer patients with *adenomyosis* a hysterectomy over repeat ablation, one of the other surgeries offered as a reintervention in the MMC.

6.5.2 Ablation power

One of the process features highly present in multiple rankings is *ablation power*. It is ranked at position four in the relative ranking and has an average rank of 11.4. According to Abbott et al. (2003), the *ablation power* is automatically based on the cavity's size. When looking at figure 5.17, this is slightly hinted. In experiment 2, ablation power is also ranked top 10 if *cavity width* is. To what extend this is grounded or a coincidence, can be further investigated.

6.5.3 Waiting time

Waiting time is relatively ranked fourth with an average rank of 11.4. Despite the fact that the feature is this important according to this research, literature only mentions *waiting time* with the purpose of decreasing, a purpose being more important since the consequences of the Covid pandemic (Ghoubara et al., 2021).

TABLE 6.13: Most important patient characteristics and process features based on their average ranking

Feature	Relative rank	Average rank
Adenomyosis	1	1.8
Telefonisch consult	2	8.4
Herhaalpoliklinkiekbezoek	3	8.8
Ablation power	4	11.4
Echoscopie gynaecologisch	4	11.4
Waiting time	4	11.4
Diagnostische hyteroscopie inclusief eventuele...	5	11.6
Ablation duration	6	11.8
BMI	7	12.4
Spoedeisende hulp contact buiten de seh afdeling...	7	12.4
Uterine fibriods	7	12.4
Age	8	12.8
Cavity width	9	13.1
Doelgerichte telefonische consultatie van een ...	10	13.4
11/21 operatief kliniek 404	11	13.8
Sis echo	12	13.9
Controle patient gynaecologie	13	14.0
Dagverpleging i	14	14.2
Amount of children unknown	15	14.3
Dysmenorrhea	15	14.3
Herhalingsbezoek	15	14.3
Cavity length	16	14.4
Dagverpleging	16	14.4
Three or more children	16	14.4
Anteverted uterus	17	14.5
One child	17	14.5
Controle patiënt	18	14.6
Echo op verzoek huisarts gynaecologisch	19	14.7
Nieuwe patient abnormaal bloedverlies	19	14.7
Zorgdomein abnormaal bloedverlies	19	14.7
Local anaesthetic	20	14.8
Beoordeling ecg holter inspanningsonderzoek ed	20	14.8
Poliklinische behandeling	20	14.8
Two children	21	14.9
Uteruscuretteage exclusief diagnostische ...	21	14.9
Geen uitval standaard cyclusstoornissen ...	22	15.0
Nieuwe patient plaatsen iud	22	15.0

6.5.4 Ablation duration

Ablation duration is ranked as sixth important feature with and average ranking of 11.8. The expert is curious about the maximum duration of the ablation. 120 seconds is a prescribed limit, but the question arises what happens if the ablation is allowed to take, for example, five seconds longer. Their hypothesis is that maybe those five seconds are determining in the patient having to undergo a reintervention or not. Unfortunately, previous researches on this feature also do not exceed these 120 seconds (Cooper et al., 2002).

6.5.5 BMI

BMI is ranked seventh in the relative rank with an average rank of 12.4. It is generally known that the higher the patient's *BMI* is, the harder it gets for doctors to investigate and cure patients. But literature does not share the idea of *BMI* having a direct influence on the outcome of the Novasure surgery. Only one study (Smithling et al., 2014) found an association between *BMI* and reintervention, while by the other twelve studies discussed by Beelen et al. (2019), no association was found. But Beelen et al. (2019) also presents that the pooled data of three studies (Wishall et al., 2014; Madsen et al., 2013; Smith, Karpate, and Clark, 2016) show that patients with a *BMI* higher than 30 are at higher risk of receiving a reintervention. 152 patients from the dataset used in this research have a *BMI* higher than 30.

6.5.6 Uterine fibroids

Like *BMI*, *uterine fibroids* are ranked seventh in the relative rank with an average of 12.4. Earlier research shows conflicting results on *uterine fibroids* (Beelen et al., 2019). This research only focuses on the presence of *uterine fibroids*, whereas other researches distinguish in size, the amount of fibroids or type. Of the 14 studies in which the presence of *uterine fibroids* was analysed as a prognostic factor, six studies (Peeters et al., 2013; Riley, Davies, and Harkins, 2013; Bansi-Matharu et al., 2013; Soini et al., 2017; Nakamura et al., 2017; Glasser, Heinlein, and Hung, 2009) found an association between the presence and reintervention. The other eight studies did not find an association (El-Nashar et al., 2009; Simon et al., 2015; Amso et al., 2003; Hachmann-Nielsen and Rudnicki, 2012; Shavell et al., 2012; Klebanoff et al., 2017; Kdous et al., 2008; Smithling et al., 2014).

6.5.7 Age

With an average rank of 12.8, *age* appears to be the eight most influential factor for a Novasure surgery. According to the expert, this seems logical. At some point the patient addresses her menopause and will, even if they had not had a Novasure surgery, not suffer from AUB anymore. The older the patient is, the sooner the menopause arrives and the less time complaints have to reoccur. Among earlier research, *age* had widely been analysed as a prognostic factor Beelen et al. (2019). 11 out of 19 studies (Simon et al., 2015; Longinotti et al., 2008; Kdous et al., 2008; Nakamura et al., 2017; Soini et al., 2017; Bansi-Matharu et al., 2013; Riley, Davies, and Harkins, 2013; Shavell et al., 2012; Hachmann-Nielsen and Rudnicki, 2012; Kopeika et al., 2011; Klebanoff et al., 2017) concluded that younger women are at higher risk of receiving a surgical reintervention compared with older women.

6.5.8 Cavity length and width

Cavity length and *width* occur botch 10 times. *Cavity width* is relatively ranked ninth and *cavity length* 16^{th}. Their average ranking differ less, with 13.1 for *cavity width* and 14.4 for *cavity length*. During the Novasure surgery the device should cover the full internal side of the uterus, covering the total length of the cavity, as well as the width. After process features have been added as input data, both *cavity length* and *width* do not occur in the rankings anymore. This outcome contributes to what is already found in literature. Thiel et al. (2014) researched, among other things, surgical reinterventions on one group with a sounded uterus length over 10 cm and one group with a sounded uterus length under 10 cm, and found no serious

procedure-related adverse events in either group. Also Beelen et al. (2019) noted that one out of six studies (Peeters et al., 2013) researching uterus length found an association between the uterus length and reinterventions.

6.5.9 Parity

When taking patient features into account, *parity* is a feature which occurs in the rankings quite often for most models. Each of the options, *amount of children unknown*, *no children*, *one child*, *two children*, and *three or more children*, occur at least four times and at most nine times in the rankings. The options are more present when the focus is on patient characteristics, but for some reason, all options, except for *three or more children* are very relevant according to neural network in experiment 4. To current knowledge, no association has been found between *parity* and a patient receiving a reintervention (Beelen et al., 2019), which might explain the loss of importance when process features are added. In literature, there are different ways to represent *parity* as a variable. In multiple studies *parity* is seen as a float variable. Others use nulliparous versus parous, and make it a binary variable. Then, there are the studies which take categories with the amount of children, as done in this research. Madsen et al. (2013) and El-Nashar et al. (2009) handled *parity* as a categorical value and found that only if a patient had five or more deliveries, it correlates with undergoing a reintervention.

6.5.10 Dysmenorrhea

Dysmenorrhea is relatively ranked as 15^{th} with an average rank of 14.3, which literature agrees with. Beelen et al. (2019) describes that out of nine studies, seven showed that preexisting *dysmenorrhea* was a predictive factor for surgical reintervention. Out of eight studies with reported outcome measures, three studies described that both present or preexisting but now absent *dysmenorrhea* is a prognostic factor for failure of endometrial ablation, like the Novasure surgery.

6.5.11 Uterus position

All four positions of the uterus, including *anteverted uterus retroverted uterus stretched position* and *nothing found*, occur three to seven times in the rankings, but only *anteverted uterus* receives an average ranking of 14.5 and is averagely ranked 17^{th}. Contrary to that, the only position being associated with reintervention is a retroverted uterus, according to literature (Bongers, Mol, and Brölmann, 2002; Amso et al., 2003; Gervaise et al., 1999; Lok et al., 2003; Agarwal et al., 2011). Sidenote to this research is that a balloon device is used, which is a different ablation device than the Novasure.

6.5.12 Endometrial thickness

Even though *endometrial thickness* has not been ranked 15 or more influential, it is a feature to consider before performing a Novasure surgery. According to literature, *endometrial thickness* is correlated to other prognostics factors, like *age* and *BMI*. Beelen et al. (2019) mention the thought that endometrium regenerates over the years, which puts older women at lower risk of failure because they have reached menopause by that time, introducing both *age* and *endometrial thickness* as prognostic factors. Hapangama and Bulmer (2016) hypothese that the excessive amount of fat tissue in obese women causes hyperestrogenic state, which facilitates rebuilding of

the endometrium after ablation. This has the consequence of the patient being more likely to receive a reintervention.

6.5.13 Appointments and care activities

In total, 38 unique appointments and care activities were added to the features occurring in the rankings for patient characteristics and process features, as can be seen in table 5.26. The meaning of the appointments and care activities is left out of scope, but prominences are discussed. *Diagnostische hyteroscopie inclusief eventuele proefexcisie(s) en/of inclusief eventuele endometriumbiopsie(en) en/of het verwijderen van een enkelvoudige poliep voor pathologisch onderzoek* appeared in seven of the models. Only *adenomyosis*, *age*, *amount of children unknown*, *BMI*, *cavity length*, *cavity width*, *dysmenorrhea*, *endometrial thickness*, and *myomatosus* occur more often in the rankings, all being patient characteristics, and thus having twice as much chance of appearing in the rankings. The phrase *'Geen uitval standaard'* is followed seven times by mostly the same phrases being positive or negative, depending on the presence of the word '*geen*'. It looks like this is a template to fill in and the cumulative presence of ten times in the rankings points out that this phrase provides quite interesting information to investigate more. *Herhalingsbezoek* and *Herhaalpolikliniekbezoek* occur two and ten times, and, although they do not provide any extra information on what the appointment was on, these features might provide insights on the process and the following chance of a reintervention, as suggested by the expert. Taking the terms *Telefonisch consult* and *Telefonisch consult poli medewerker* together, they occur 11 times in the rankings. Investigating in these different appointments and care activities might be interesting for further research.

6.6 Threats of validity

Like every other research, this research has threats of validity which influence the quality of the research, but also open opportunities to investigate and improve.

First, preparing the data can cause bias in this research. The data set used came from the hospital and includes raw data written by the genealogists while they were seeing patients. This results in a dishevelled data set, making it hard to retrieve some values. For some features, information about more than half of the total amount of patients was missing. With the total amount of patients being 1029, all living in Eindhoven and its surroundings, the dataset seems quite small and specific. Taking this into account, the research may be influenced by sampling bias.

During the project, the patient and process features chosen are mostly based on a paper on prognostic factors found with the help of a literature study and a medical expert's opinion. Both the author of the paper and the medical expert work at the gynaecologist department of the MMC. After results have been obtained, validation was done by the same literature and a MMC expert. This causes a construct validity in the form of mono-operation bias. Added to that, expert validation has been done by only one expert. It could be seen as a limitation that this research has not been externally validated in another cohort or by an expert from a different hospital providing the same surgery.

While creating the models, hyperparameter tuning has been performed to best knowledge, but the possibility exists that a combination of an untested classification technique and a classifier outperforms the settings used in this research. Also, there might exist parameter settings which outperform the current settings, even though

the hyperparameter settings are optimised using multiple iterations for optimilisation or a different optimisation algorithm. Later was found that, for example, the Bayesian hyperparameter optimisation performs better than the used GridSearchCV (Binder et al., 2020; Pijnenborg et al., 2021). Construct validity occurs here.

One last thing which should be taken into account is that this research is performed by a Business Informatics student. Like with every data science project, my scope of knowledge is limited to data science and does not include any form of medical knowledge apart from personal experiences. Although, we have seen and learnt about the genealogist domain to a great extend during the project, we are not the expert in this field and we have tried to bridge this gap with surrounding experts.

Chapter 7

Conclusion

In this research, we investigated the use of historical data of Novasure patients to provide evidence-based insights into current treatments and their impacts on the outcome of the Novasure surgery per patient. The Novasure surgery is a minimally invasive procedure that aims to destroy or remove endometrial tissue in order to solve heavy menstrual bleeding complaints. Related work on both machine learning techniques and patient-level predicting techniques in healthcare have been discussed with as most important work '*Prediction of unsuccessful endometrial ablation: random forest vs logistic regression*' by Stevens et al. (2021). Six algorithms have been taken into account: decision tree, random forest, logistic regression, extreme gradient boosting, neural network, and k-nearest neighbour. All six algorithms have been run four times. Once with the original dataset and patient characteristics, once adding process features as input data with the original dataset, once with a sampled dataset only using patient characteristics and once focusing on patient characteristics and process features using the sampled dataset.

The first research question to answer is *To what extent do patient characteristics have an influence on the outcome of a Novasure surgery?* With the help of literature and the experiments, this question is answered. From literature can be concluded that *age* and *dysmennorrhoea* are important patient characteristics influencing the outcome of a Novasure surgery. From the experiments can be concluded that *adenomyosis* is the feature which is the most important when it comes to prognostic factors. It has been ranked very high by all 12 models. Both *age* and *BMI* contribute in the determination of the outcome of a Novasure surgery. Also, *cavity length* and *cavity width* are features influencing the outcome of the Novasure surgery.

Then, experiments are carried out to answer the second research question: *How are predicting results of the Novasure success outcome influenced by including process features?* Including process features to the input data, changed the outcome drastically. Next to perioperative features *ablation power*, *waiting time*, *ablation duration*, and *local anaesthetic*, 38 appointments and care activities raised as important features. One can conclude that taking process features into account can increase the accuracy of predicting whether a reintervention is wished after undergoing a Novasure surgery. As described above, patient characteristics have a direct influence on process features and thereby an impact on the result of the surgery. This research left further investigation out of scope, but with multiple appointments and care activities also appearing multiple times in the rankings, it is valuable to give attention to process features. Very interesting is that important features provide information of the uterus or are connected in another way. It includes the *cavity width* and *cavity length*, which directly influence the form of the conformable bipolar electrode array, and the *endometrial thickness*, which has an influence on the ablation power, and the presence of *uterine fibroids*.

The third question is *How do the following predictive machine learning techniques perform compared to each other?* The six algorithms to compare are decision tree, random forest, logistic regression, extreme gradient boosting, neural network, and k-nearest neighbour. To predict surgery outcomes on patient-level, not one algorithm is to point as the best. On the original data, decision tree and extreme gradient boosting perform best with respect to accuracy. In both cases, neural network reached the highest AUC. When the dataset is sampled, random forest performs best with respect to accuracy. Both random forest and extreme gradient boosting perform best with respect to the AUC in these two experiments.

So, to answer the main research question: *Given patient characteristics and process features, which machine learning algorithm(s) can predict the outcome of Novasure surgery with highest accuracy?*: Purely looking at the accuracy, decision tree and random forest both reached the highest accuracy of 0.88 of all experiments. This was under the conditions of only taking into account patient characteristics and sampling the original data to get two equally sized groups for no intervention. Due to the impact of process features on the Novasure surgery, it is valuable to take those into account. Under these circumstances, the answer would be that random forest, together with extreme gradient boosting and neural network are able to predict the outcome of Novasure surgery with the highest accuracy. Taking possible improvements into account, it is valuable to perform research including multiple machine learning algorithms when predicting the outcome of the Novasure surgery.

This research also softens the conclusions drawn by Stevens et al. (2021). With a comparable research, the that logistic regression does not necessarily outperform other machine learning algorithms. In fact, for all four experiments, the AUC of random forest was higher than for logistic regression, and all AUC's were higher than the AUC's retrieved by that study. Due to results being so close to each other, this research shows that it is valuable to investigate in multiple machine learning models for predicting Novasure success outcomes.

7.1 Future work

After carrying out this research on the prediction of the outcome of Novasure surgery and which predictive machine learning model performs best to do this, we think that there are some factors which would contribute in delivering an improved or extended research and valuable results.

Applying the method in which both patient characteristics and process features are taken into account while predicting surgery success outcomes can validate whether process features also have an influence on surgery outcomes and contribute stronger in generalisation of this hypothesis. When focusing on the Novasure surgery, other algorithms could be added to the research and trying different techniques of hyperparameter tuning.

As a response to the results, it is interesting to first investigate in redoing this research, with the aim of finding reinterventions as precisely as possible. This would include tuning models with other scoring metrics. Also, it would be very interesting to investigate in different appointments and care activities, to find out what specific actions influence the process and the outcome of the Novasure surgery or are prior indications that a Novasure will not be sufficient and a reintervention is needed. This research has proven that process features have a significant influence on the outcome of the Novasure surgery, but has not investigated in which specific process features take the lead. This might be good starting point for further research.

Medically, the most interesting future work is building a tool in which genealogists can fill in patients characteristics and receive a statistical report with information on whether a Novasure surgery is suitable for the patient or whether an other type of ablation or intervention is desired. This could help them with patient-level decision making and improve expectation management for the patient.

Bibliography

Abbott, Jason, Jed Hawe, David Hunter, and Ray Garry (2003). "A double-blind randomized trial comparing the Cavaterm™ and the NovaSure™ endometrial ablation systems for the treatment of dysfunctional uterine bleeding". In: *Fertility and sterility* 80.1, pp. 203–208.

Abedjan, Ziawasch, Nozha Boujemaa, Stuart Campbell, Patricia Casla, Supriyo Chatterjea, Sergio Consoli, Cristobal Costa-Soria, Paul Czech, Marija Despenic, Chiara Garattini, Dirk Hamelinck, Adrienne Heinrich, Wessel Kraaij, Jacek Kustra, Aizea Lojo, Marga Martin Sanchez, Miguel A. Mayer, Matteo Melideo, Ernestina Mensalvas, Frank Moller Aarestrup, Elvira Narro Artigot, Milan Petkoveć, Diego Reforgiato Recupero, Alejandro Rodriguez Gonzalez, Gisele Roesems Kerremans, Roland Roller, Mario Romao, Stefan Ruping, Felix Sasaki, Wouter Spek, Nenad Stojanovic, Jack Thoms, Andrejs Vasiljevs, Wilfried Verachtert, and Roel Wuyts (Jan. 2019). "Data science in healthcare: benefits, challenges and opportunities". English. In: *Data Science for Healthcare*. Ed. by S. Consoli, D. Reforgiato Recupero, and M. Petković. Germany: Springer, pp. 3–38. ISBN: 978-3-030-05248-5. DOI: 10.1007/978-3-030-05249-2_1.

Agarwal, Swarnima, Adarsh Bhargava, Nimmi Chutani, and Pushpa Nagar (2011). "Uterine balloon therapy for the treatment of menorrhagia". In: *The Journal of Obstetrics and Gynecology of India* 61.1, pp. 67–71.

Altman, Naomi S (1992). "An introduction to kernel and nearest-neighbor nonparametric regression". In: *The American Statistician* 46.3, pp. 175–185.

Amso, Nazar N, Herve Fernandez, George Vilos, Claude Fortin, Peter McFaul, Monika Schaffer, PFM Van der Heijden, Marlies Y Bongers, Barry Sanders, and Bernard Blanc (2003). "Uterine endometrial thermal balloon therapy for the treatment of menorrhagia: long-term multicentre follow-up study". In: *Human Reproduction* 18.5, pp. 1082–1087.

Baijens, Jeroen, Remko Helms, and Deniz Iren (2020). "Applying scrum in data science projects". In: *2020 IEEE 22nd Conference on Business Informatics (CBI)*. Vol. 1. IEEE, pp. 30–38.

Bansi-Matharu, L, I Gurol-Urganci, TA Mahmood, A Templeton, JH Van der Meulen, and DA Cromwell (2013). "Rates of subsequent surgery following endometrial ablation among E nglish women with menorrhagia: population-based cohort study". In: *BJOG: An International Journal of Obstetrics & Gynaecology* 120.12, pp. 1500–1507.

Beelen, Pleun, Imke MA Reinders, Wessel FW Scheepers, Malou C Herman, Peggy MAJ Geomini, Sander MJ van Kuijk, and Marlies Y Bongers (2019). "Prognostic factors for the failure of endometrial ablation: a systematic review and meta-analysis". In: *Obstetrics & Gynecology* 134.6, pp. 1269–1281.

Binder, Martin, Julia Moosbauer, Janek Thomas, and Bernd Bischl (2020). "Multi-objective hyperparameter tuning and feature selection using filter ensembles". In: *Proceedings of the 2020 Genetic and Evolutionary Computation Conference*, pp. 471–479.

Bongers, Marlies Y (2007). "Second-generation endometrial ablation treatment: Novasure". In: *Best Practice & Research Clinical Obstetrics & Gynaecology* 21.6, pp. 989–994.

— (2015). "Hysteroscopy and heavy menstrual bleeding (to cover TCRE and second-generation endometrial ablation)". In: *Best Practice & Research Clinical Obstetrics & Gynaecology* 29.7, pp. 930–939.

Bongers, MY, BWJ Mol, and HAM Brölmann (2002). "Prognostic factors for the success of thermal balloon ablation in the treatment of menorrhagia". In: *Obstetrics & Gynecology* 99.6, pp. 1060–1066.

Borak, Jordan S (1999). "Feature selection and land cover classification of a MODIS-like data set for a semiarid environment". In: *International Journal of Remote Sensing* 20.5, pp. 919–938.

Bose, RP Jagadeesh Chandra, Ronny S Mans, and Wil MP Van der Aalst (2013). "Wanna improve process mining results?" In: *2013 IEEE symposium on computational intelligence and data mining (CIDM)*. IEEE, pp. 127–134.

Breiman, Leo (2001). "Random forests". In: *Machine learning* 45.1, pp. 5–32.

Canteiro, Renata, M Valeria Bahamondes, Arlete dos Santos Fernandes, Ximena Espejo-Arce, Nadia M Marchi, and Luis Bahamondes (2010). "Length of the endometrial cavity as measured by uterine sounding and ultrasonography in women of different parities". In: *Contraception* 81.6, pp. 515–519.

Cazacu, Mihaela and Emilia Titan (2021). "Adapting CRISP-DM for social sciences". In: *BRAIN. Broad Research in Artificial Intelligence and Neuroscience* 11.2Sup1, pp. 99–106.

Chapelle, Olivier, Patrick Haffner, and Vladimir N Vapnik (1999). "Support vector machines for histogram-based image classification". In: *IEEE transactions on Neural Networks* 10.5, pp. 1055–1064.

Chapman, Pete, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinartz, Colin Shearer, Rudiger Wirth, et al. (2000). "CRISP-DM 1.0: Step-by-step data mining guide". In: *SPSS inc* 9, p. 13.

Chawla, Nitesh V, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer (2002). "SMOTE: synthetic minority over-sampling technique". In: *Journal of artificial intelligence research* 16, pp. 321–357.

Cooper, Jay, Richard Gimpelson, Philippe Laberge, Donald Galen, Jose Gerardo Garza-Leal, Josef Scott, Nicholas Leyland, Paul Martyn, and James Liu (2002). "A randomized, multicenter trial of safety and efficacy of the NovaSure system in the treatment of menorrhagia". In: *The Journal of the American Association of Gynecologic Laparoscopists* 9.4, pp. 418–428.

Cortes, Corinna and Vladimir Vapnik (1995). "Support-vector networks". In: *Machine learning* 20.3, pp. 273–297.

Coulter, Angela, Viv Peto, and Crispin Jenkinson (1994). "Quality of life and patient satisfaction following treatment for menorrhagia". In: *Family Practice* 11.4, pp. 394–401.

Daniels, Jane P (2013). "The long-term outcomes of endometrial ablation in the treatment of heavy menstrual bleeding". In: *Current Opinion in Obstetrics and Gynecology* 25.4, pp. 320–326.

Dhar, Vasant (2013). "Data science and prediction". In: *Communications of the ACM* 56.12, pp. 64–73.

El-Nashar, Sherif A, Matthew R Hopkins, Douglas J Creedon, Jennifer L St Sauver, Amy L Weaver, Michaela E McGree, William A Cliby, and Abimbola O Famuyide (2009). "Prediction of treatment outcomes after global endometrial ablation". In: *Obstetrics and gynecology* 113.1, p. 97.

Foody, Giles M and Ajay Mathur (2004). "A relative evaluation of multiclass image classification by support vector machines". In: *IEEE Transactions on geoscience and remote sensing* 42.6, pp. 1335–1343.

Fraser, Ian S, Sue Langham, and Kerstin Uhl-Hochgraeber (2009). "Health-related quality of life and economic burden of abnormal uterine bleeding". In: *Expert Review of Obstetrics & Gynecology* 4.2, pp. 179–189.

Fraser, Ian S, Diana Mansour, Christian Breymann, Camille Hoffman, Anna Mezzacasa, and Felice Petraglia (2015). "Prevalence of heavy menstrual bleeding and experiences of affected women in a European patient survey". In: *International Journal of Gynecology & Obstetrics* 128.3, pp. 196–200.

Gahegan, M and G West (2001). *The classification of Complex Data Sets: An operational Comparison of Artificial Neural Networks and Decision Tree Classifiers, Eylül (1998). Han, J. and Kamber M.," Data Mining: Concepts and Techniques"*.

Gervaise, A, H Fernandez, S Capella-Allouc, S Taylor, S La Vieille, J Hamou, and V Gomel (1999). "Thermal balloon ablation versus endometrial resection for the treatment of abnormal uterine bleeding". In: *Human Reproduction* 14.11, pp. 2743–2747.

Ghoubara, Ahmed, Seuvandhi Gunasekera, Lavanya Rao, and Ayman Ewies (2021). "Re-intervention and patient satisfaction rates following office radiofrequency endometrial ablation: a comparative retrospective study of 408 cases". In: *Journal of Obstetrics and Gynaecology*, pp. 1–7.

Glasser, Mark H, Peter K Heinlein, and Yun-Yi Hung (2009). "Office endometrial ablation with local anesthesia using the HydroThermAblator system: comparison of outcomes in patients with submucous myomas with those with normal cavities in 246 cases performed over 51/2 years". In: *Journal of Minimally Invasive Gynecology* 16.6, pp. 700–707.

Goldstuck, Norman D (2018). "Dimensional analysis of the endometrial cavity: how many dimensions should the ideal intrauterine device or system have?" In: *International Journal of Women's Health* 10, p. 165.

Gowin, Joshua L, Tali M Ball, Marc Wittmann, Susan F Tapert, and Martin P Paulus (2015). "Individualized relapse prediction: Personality measures and striatal and insular activity during reward-processing robustly predict relapse". In: *Drug and alcohol dependence* 152, pp. 93–101.

Hachmann-Nielsen, Elise and Martin Rudnicki (2012). "Clinical outcome after hydrothermal ablation treatment of menorrhagia in patients with and without submucous myomas". In: *Journal of Minimally Invasive Gynecology* 19.2, pp. 212–216.

Hallberg, Leif, AM Hogdahl, Lennart Nilsson, Göran Rybo, et al. (1966). "Menstrual blood loss and iron deficiency." In: *Acta medica scandinavica* 180, pp. 639–650.

Hapangama, Dharani K and Judith N Bulmer (2016). "Pathophysiology of heavy menstrual bleeding". In: *Women's Health* 12.1, pp. 3–13.

Harrison, Matt (2019). *Machine learning pocket reference: working with structured data in python*. O'Reilly Media.

Hong, Pengyu, Qi Tian, and Thomas S Huang (2000). "Incorporate support vector machines to content-based image retrieval with relevance feedback". In: *Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101)*. Vol. 3. IEEE, pp. 750–753.

Huber, Steffen, Hajo Wiemer, Dorothea Schneider, and Steffen Ihlenfeldt (2019). "DMME: Data mining methodology for engineering applications–a holistic extension to the CRISP-DM model". In: *Procedia Cirp* 79, pp. 403–408.

Kaufmann, Michael, Tobias Eljasik-Swoboda, Christian Nawroth, Kevin Berwind, Marco X Bornschlegl, and Matthias L Hemmje (2017). "Modeling and Qualitative Evaluation of a Management Canvas for Big Data Applications." In: *DATA*, pp. 149–156.

Kdous, Moez, Denis Jacob, Amelie Gervaise, Elie Risk, and Eric Sauvanet (2008). "Thermal balloon endometrial ablation for dysfunctional uterine bleeding: technical aspects and results. A prospective cohort study of 152 cases". In: *La Tunisie Medicale* 86.5, pp. 473–478.

Klebanoff, Jordan, Gretchen E Makai, Nima R Patel, and Matthew K Hoffman (2017). "Incidence and predictors of failed second-generation endometrial ablation". In: *Gynecological Surgery* 14.1, pp. 1–6.

Koorn, Jelmer, Xixi Lu, Henrik Leopold, and Hajo A Reijers (2019). "Towards Understanding Aggressive Behavior in Residential Care Facilities Using Process Mining". In: *International Conference on Conceptual Modeling*. Springer, pp. 135–145.

Koorn, Jelmer J, Iris Beerepoot, Vinicius Stein Dani, Xixi Lu, Inge Van de Weerd, Henrik Leopold, and Hajo A Reijers (2021). "Bringing Rigor to the Qualitative Evaluation of Process Mining Findings: An Analysis and a Proposal". In: *2021 3rd International Conference on Process Mining (ICPM)*. IEEE, pp. 120–127.

Koorn, Jelmer J, Xixi Lu, Henrik Leopold, and Hajo A Reijers (2020). "Looking for meaning: Discovering action-response-effect patterns in business processes". In: *International Conference on Business Process Management*. Springer, pp. 167–183.

Koorn, Jelmer Jan, Xixi Lu, Felix Mannhardt, Henrik Leopold, and Hajo A Reijers (2022). "Uncovering Complex Relations in Patient Pathways based on Statistics: the Impact of Clinical Actions". In: *Proceedings of the 55th Hawaii International Conference on System Sciences*.

Kopeika, Julia, Simon E Edmonds, Gautam Mehra, and Mohamed A Hefni (2011). "Does hydrothermal ablation avoid hysterectomy? Long-term follow-up". In: *American journal of obstetrics and gynecology* 204.3, 207–e1.

Krause, Josua, Adam Perer, and Enrico Bertini (2016). "Using visual analytics to interpret predictive machine learning models". In: *arXiv preprint arXiv:1606.05685*.

Kuhn, Max, Kjell Johnson, et al. (2013). *Applied predictive modeling*. Vol. 26. Springer.

Lenz, Richard, Thomas Elstner, Hannes Siegele, and Klaus A Kuhn (2002). "A practical approach to process support in health information systems". In: *Journal of the American Medical Informatics Association* 9.6, pp. 571–585.

Lethaby, Anne, Cynthia Farquhar, Alvaro Sarkis, Helen Roberts, Ruth Jepson, and David Barlow (2004). "Hormone replacement therapy in postmenopausal women: endometrial hyperplasia and irregular bleeding". In: *Cochrane Database of Systematic Reviews* 2.

Lewis, Sarah (2015). "Qualitative inquiry and research design: Choosing among five approaches". In: *Health promotion practice* 16.4, pp. 473–475.

Liu, Zhimei, Quan V Doan, Paul Blumenthal, and Robert W Dubois (2007). "A systematic review evaluating health-related quality of life, work impairment, and health-care costs and utilization in abnormal uterine bleeding". In: *Value in health* 10.3, pp. 183–194.

Livingstone, Mark and Ian S Fraser (2002). "Mechanisms of abnormal uterine bleeding". In: *Human reproduction update* 8.1, pp. 60–67.

Lok, Ingrid Hung, Pui Ling Leung, Pui Shan Ng, and Pong Mo Yuen (2003). "Life-table analysis of the success of thermal balloon endometrial ablation in the treatment of menorrhagia". In: *Fertility and sterility* 80.5, pp. 1255–1259.

Longinotti, Mindyn K, Gavin F Jacobson, Yun-Yi Hung, and Lee A Learman (2008). "Probability of hysterectomy after endometrial ablation". In: *Obstetrics & Gynecology* 112.6, pp. 1214–1220.

Lukes, Andrea S, Jeffrey Baker, Scott Eder, and Tammie L Adomako (2012). "Daily menstrual blood loss and quality of life in women with heavy menstrual bleeding". In: *Women's Health* 8.5, pp. 503–511.

Madsen, Annetta M, Sherif A El-Nashar, Matthew R Hopkins, Zaraq Khan, and Abimbola O Famuyide (2013). "Endometrial ablation for the treatment of heavy menstrual bleeding in obese women". In: *International Journal of Gynecology & Obstetrics* 121.1, pp. 20–23.

Mans, Ronny S, MH Schonenberg, M Song, WMP Van der Aalst, and PJM Bakker (2015). "Process mining in healthcare". In: *International Conference on Health Informatics (HEALTHINF'08)*, pp. 118–125.

Martinez, Iñigo, Elisabeth Viles, and Igor G Olaizola (2021). "Data science methodologies: Current challenges and future approaches". In: *Big Data Research* 24, p. 100183.

Martínez-Plumed, Fernando, Lidia Contreras-Ochando, Cesar Ferri, José Hernández Orallo, Meelis Kull, Nicolas Lachiche, Maréa José Ramírez Quintana, and Peter A Flach (2019). "CRISP-DM twenty years later: From data mining processes to data science trajectories". In: *IEEE Transactions on Knowledge and Data Engineering*.

Matteson, Kristen A, Lori A Boardman, Malcolm G Munro, and Melissa A Clark (2009). "Abnormal uterine bleeding: a review of patient-based outcome measures". In: *Fertility and sterility* 92.1, pp. 205–216.

Maxwell, Aaron E, Timothy A Warner, and Fang Fang (2018). "Implementation of machine-learning classification in remote sensing: An applied review". In: *International Journal of Remote Sensing* 39.9, pp. 2784–2817.

McCausland, Vance and Arthur McCausland (1998). "The response of adenomyosis to endometrial ablation/resection". In: *Human reproduction update* 4.4, pp. 350–359.

Mercier, Grégoire and Marc Lennon (2003). "Support vector machines for hyperspectral image classification with spectral-based kernels". In: *IGARSS 2003. 2003 IEEE International Geoscience and Remote Sensing Symposium. Proceedings (IEEE Cat. No. 03CH37477)*. Vol. 1. IEEE, pp. 288–290.

Myles, Anthony J, Robert N Feudale, Yang Liu, Nathaniel A Woody, and Steven D Brown (2004). "An introduction to decision tree modeling". In: *Journal of Chemometrics: A Journal of the Chemometrics Society* 18.6, pp. 275–285.

Nakamura, Kohei, Kentaro Nakayama, Kaori Sanuki, Toshiko Minamoto, Tomoka Ishibashi, Emi Sato, Hitomi Yamashita, Masako Ishikawa, and Satoru Kyo (2017). "Long-term outcomes of microwave endometrial ablation for treatment of patients with menorrhagia: A retrospective cohort study". In: *Oncology Letters* 14.6, pp. 7783–7790.

Natekin, Alexey and Alois Knoll (2013). "Gradient boosting machines, a tutorial". In: *Frontiers in neurorobotics* 7, p. 21.

Niaksu, Olegas (2015). "CRISP data mining methodology extension for medical domain". In: *Baltic Journal of Modern Computing* 3.2, p. 92.

Pal, Mahesh and Paul M Mather (2003). "An assessment of the effectiveness of decision tree methods for land cover classification". In: *Remote sensing of environment* 86.4, pp. 554–565.

Passos, Ives Cavalcante, Benson Mwangi, Bo Cao, Jane E Hamilton, Mon-Ju Wu, Xiang Yang Zhang, Giovana B Zunta-Soares, Joao Quevedo, Marcia Kauer-Sant'Anna, Flavio Kapczinski, et al. (2016). "Identifying a clinical signature of suicidality

among patients with mood disorders: A pilot study using a machine learning approach". In: *Journal of affective disorders* 193, pp. 109–116.

Peeters, Jos AH, Josien PM Penninx, Ben Willem Mol, and Marlies Y Bongers (2013). "Prognostic factors for the success of endometrial ablation in the treatment of menorrhagia with special reference to previous cesarean section". In: *European Journal of Obstetrics & Gynecology and Reproductive Biology* 167.1, pp. 100–103.

Penninx, Josien PM, Malou C Herman, Ben W Mol, and Marlies Y Bongers (2011). "Five-year follow-up after comparing bipolar endometrial ablation with hydrothermablation for menorrhagia". In: *Obstetrics & Gynecology* 118.6, pp. 1287–1292.

Pijnenborg, Pam, Rob Verhoeven, Murat Firat, Hanneke van Laarhoven, and Laura Genga (2021). "Towards Evidence-Based Analysis of Palliative Treatments for Stomach and Esophageal Cancer Patients: a Process Mining Approach". In: *2021 3rd International Conference on Process Mining (ICPM)*. IEEE, pp. 136–143.

Poplin, Ryan, Avinash V Varadarajan, Katy Blumer, Yun Liu, Michael V McConnell, Greg S Corrado, Lily Peng, and Dale R Webster (2018). "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning". In: *Nature Biomedical Engineering* 2.3, pp. 158–164.

Provost, Foster and Tom Fawcett (2013). *Data Science for Business: What you need to know about data mining and data-analytic thinking*. " O'Reilly Media, Inc."

Reps, Jenna M, Martijn J Schuemie, Marc A Suchard, Patrick B Ryan, and Peter R Rijnbeek (2018). "Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data". In: *Journal of the American Medical Informatics Association* 25.8, pp. 969–975.

Riley, Kristin A, Matthew F Davies, and Gerald J Harkins (2013). "Characteristics of patients undergoing hysterectomy for failed endometrial ablation". In: *JSLS: Journal of the Society of Laparoendoscopic Surgeons* 17.4, p. 503.

Rodriguez, Magdalena Bofill, Anne Lethaby, Mihaela Grigore, Julie Brown, Martha Hickey, and Cindy Farquhar (2019). "Endometrial resection and ablation techniques for heavy menstrual bleeding". In: *Cochrane Database of Systematic Reviews* 1.

Sangra, Rosa Abellana and Andreu Farran Codina (2015). "The identification, impact and management of missing values and outlier data in nutritional epidemiology". In: *Nutrición Hospitalaria* 31.3, pp. 189–195.

Shalev-Shwartz, Shai and Shai Ben-David (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Shavell, Valerie I, Michael P Diamond, James P Senter, Michael L Kruger, and D Alan Johns (2012). "Hysterectomy subsequent to endometrial ablation". In: *Journal of minimally invasive gynecology* 19.4, pp. 459–464.

Simon, Rochelle A, M Ruhul Quddus, W Dwayne Lawrence, and C James Sung (2015). "Pathology of endometrial ablation failures: a clinicopathologic study of 164 cases". In: *International Journal of Gynecological Pathology* 34.3, pp. 245–252.

Smith, Paul P, Shilpaja Karpate, and T Justin Clark (2016). "Prognostic factors that predict success in office endometrial ablation: a retrospective study". In: *Gynecological Surgery* 13.2, pp. 83–87.

Smith, Paul P, Sadia Malick, and T Justin Clark (2014). "Bipolar radiofrequency compared with thermal balloon ablation in the office: a randomized controlled trial". In: *Obstetrics & Gynecology* 124.2 PART 1, pp. 219–225.

Smith-Bindman, R, E Weiss, and V Feldstein (2004). "How thick is too thick? When endometrial thickness should prompt biopsy in postmenopausal women without vaginal bleeding". In: *Ultrasound in Obstetrics and Gynecology: The Official*

Journal of the International Society of Ultrasound in Obstetrics and Gynecology 24.5, pp. 558–565.

Smithling, Katelyn R, Gina Savella, Christina A Raker, and Kristen A Matteson (2014). "Preoperative uterine bleeding pattern and risk of endometrial ablation failure". In: *American Journal of Obstetrics and Gynecology* 211.5, 556–e1.

Soini, Tuuli, Matti Rantanen, Jorma Paavonen, Seija Grénman, Johanna Mäenpää, Eero Pukkala, Mika Gissler, and Ritva Hurskainen (2017). "Long-term follow-up after endometrial ablation in Finland: cancer risks and later hysterectomies". In: *Obstetrics & Gynecology* 130.3, pp. 554–560.

Stevens, Kelly Yvonne Roger, Liesbet Lagaert, Tom Bakkes, Malou Evi Gelderblom, Saskia Houterman, Tanja Gijsen, and Benedictus C Schoot (2021). "Prediction of unsuccessful endometrial ablation: random forest vs logistic regression". In: *Gynecological Surgery* 18.1, pp. 1–9.

Suchting, Robert, Charles E Green, Stephen M Glazier, and Scott D Lane (2018). "A data science approach to predicting patient aggressive events in a psychiatric hospital". In: *Psychiatry research* 268, pp. 217–222.

Tatsat, Hariom, Sahil Puri, and Brad Lookabaugh (2020). *Machine Learning and Data Science Blueprints for Finance*. O'Reilly Media.

Teinemaa, Irene, Marlon Dumas, Marcello La Rosa, and Fabrizio Maria Maggi (2019). "Outcome-oriented predictive process monitoring: Review and benchmark". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 13.2, pp. 1–57.

Thiel, John A, M Martha Briggs, Scott Pohlman, and Darrien Rattray (2014). "Evaluation of the NovaSure endometrial ablation procedure in women with uterine cavity length over 10 cm". In: *Journal of Obstetrics and Gynaecology Canada* 36.6, pp. 491–497.

Thomas, David R (2006). "A general inductive approach for analyzing qualitative evaluation data". In: *American journal of evaluation* 27.2, pp. 237–246.

Van der Aalst, Wil MP (2012). "Process mining: Overview and opportunities". In: *ACM Transactions on Management Information Systems (TMIS)* 3.2, pp. 1–17.

— (2016). "Process mining: data science in action". In.

Van Eck, Maikel L, Xixi Lu, Sander JJ Leemans, and Wil MP Van Der Aalst (2015). "PM^2: a process mining project methodology". In: *International conference on advanced information systems engineering*. Springer, pp. 297–313.

Venter, Jacobus, Alta de Waal, and Cornelius Willers (2007). "Specializing CRISP-DM for evidence mining". In: *IFIP International Conference on Digital Forensics*. Springer, pp. 303–315.

Verenich, Ilya, Marlon Dumas, Marcello La Rosa, Fabrizio Maria Maggi, and Irene Teinemaa (2019). "Survey and cross-benchmark comparison of remaining time prediction methods in business process monitoring". In: *ACM Transactions on Intelligent Systems and Technology (TIST)* 10.4, pp. 1–34.

Wang, Jiaojiao, Dongjin Yu, Chengfei Liu, and Xiaoxiao Sun (2019). "Outcome-oriented predictive process monitoring with attention-based bidirectional LSTM neural networks". In: *2019 IEEE International Conference on Web Services (ICWS)*. IEEE, pp. 360–367.

Warner, Pamela, Hilary OD Critchley, Mary Ann Lumsden, Mary Campbell-Brown, Anne Douglas, and Gordon Murray (2001). "Referral for menstrual problems: cross sectional survey of symptoms, reasons for referral, and management". In: *BMJ* 323.7303, pp. 24–28.

Wirth, Rüdiger and Jochen Hipp (2000). "CRISP-DM: Towards a standard process model for data mining". In: *Proceedings of the 4th international conference on the*

practical applications of knowledge discovery and data mining. Vol. 1. Manchester, pp. 29–40.

Wishall, Kayla M, Joan Price, Nigel Pereira, Samantha M Butts, and Carl R Della Badia (2014). "Postablation risk factors for pain and subsequent hysterectomy". In: *Obstetrics & Gynecology* 124.5, pp. 904–910.

Wu, Jionglin, Jason Roy, and Walter F Stewart (2010). "Prediction modeling using EHR data: challenges, strategies, and a comparison of machine learning approaches". In: *Medical care*, S106–S113.

Appendix A

Deprecated table

A.1 From Data description

- Table A.1: Summary of dataset including their feature, range, or possible values and contingent notes, still including different kind of reinterventions.

uitleg toevoegen

A.2 From Results

- Figure **??**: Tuning on minimum number of samples to split for decision tree, with y-axis form 0 to 1.
- Figure 5.7: Tuning on maximum number of iterations for logistic regression, with y-axis from 0 to 1.
- Figure A.3: Tuning on maximum number of hidden layer sizes for neural network, with y-axis from 0 to 1.
- Figure A.4: Tuning on maximum number of hidden layer sizes for neural network, with y-axis from 0 to 1.

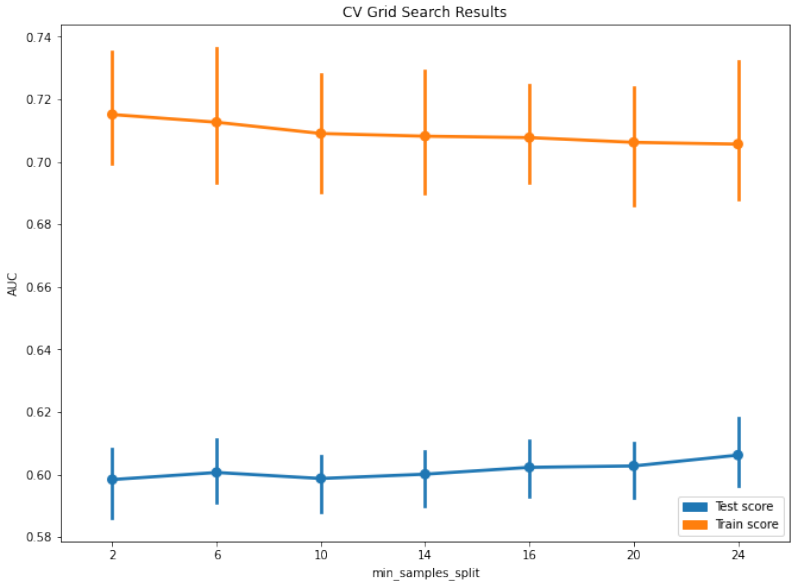


FIGURE A.1: Decision tree algorithm tuned on minimum samples for split

TABLE A.1: Summary of dataset, including their feature, range, or possible values and contingent notes.

Feature	**Range/Values**	**Notes**
Patient features - predictor variables		*Amount of missing values*
Age	$[26, 60]$	-
Complaint	*Cycle disorder*, *benign adnexal abnormality* or *uterine fibriods*	18
BMI	$[15.57, 50.20]$	442
Parity	$[0, 6]$ or *unknown*	238
Cesarean section	$0 \vee 1$	-
Uterus position	*AVF*, *Nothing found*, *RVF* or *Stretch*	-
Endometrial thickness (mm)	$[1,)$	571
Cavity length (mm)	$[22, 65]$	453
Cavity width (mm)	$[25, 55]$	315
Dysmenorrhea	$0 \vee 1$ or *unknown*	678
Endometriosis	$0 \vee 1$	-
Adenomyosis	$0 \vee 1$ or *unknown*	863
Uterine fibroids	$0 \vee 1$	-
Sterilisation	$0 \vee 1$	-
Process features - predictor variables		
Appointments 2 years pre Novasure		
Care activities 2 years pre Novasure	$[0, 100]$	
Waiting time (days)	[0,265]	313
Perioperative features - predictor variables		
Anaesthesia	*Local anaesthetic*, *Sedation* or *General anaesthetic*	-
Ablation duration (sec)	[6,120]	580
Power ablation (watt)	[1,180]	237
Reintervention information - outcome variables		*Reintervention type*
Reintervention	$0 \vee 1$	
Invasive reintervention	$0 \vee 1$	
Non-invasive reintervention	$0 \vee 1$	
Laparoscopic Hysterectomy	$0 \vee 1$	Invasive
AUE	$0 \vee 1$	Invasive
Uterus amputation	$0 \vee 1$	Invasive
VUE	$0 \vee 1$	Invasive
Fibroids resection	$0 \vee 1$	Invasive
Sonata	0	Invasive
Balloon ablation	$0 \vee 1$	Invasive
Contraceptive implant	0	Invasive
Intrauterine device	$0 \vee 1$	Invasive
Contraceptive injection	0	Invasive
Tranexamic acid	$0 \vee 1$	Non-invasive

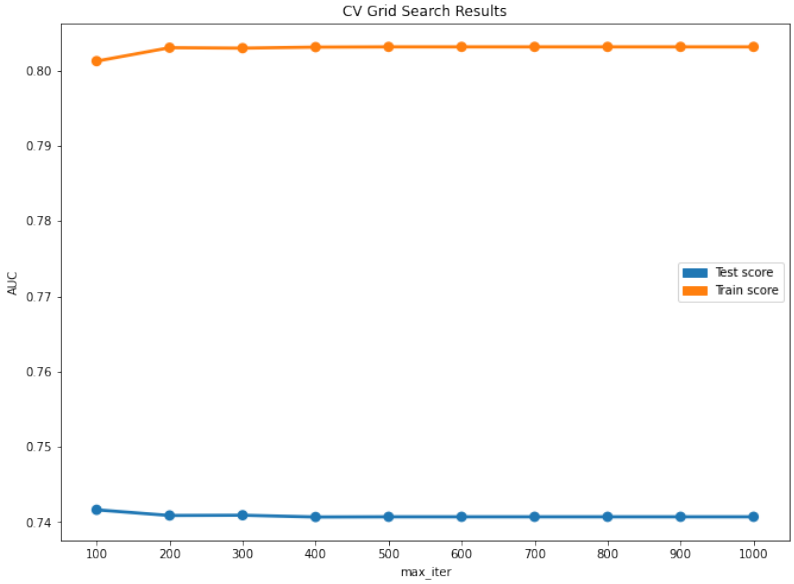


FIGURE A.2: Logistic regression algorithm tuned on maximum number of iterations

A.3 From Discussion

- 1: Adenomyosis
- 2: Telefonisch consult
- 3: Herhaalpoliklinkiekbezoek
- 4: Ablation power
- 4: Echoscopie gynaecologisch
- 4: Waiting time
- 5: Diagnostische hyteroscopie inclusief eventuele proefexcisie(s) en/of inclusief eventuele endometriumbiopsie(en) en/of het verwijderen van een enkelvoudige poliep voor pathologisch onderzoek
- 6: Ablation duration
- 7: BMI
- 7: Spoedeisende hulp contact buiten de seh afdeling elders in het ziekenhuis
- 7: Uterine fibriods
- 8: Age
- 9: Cavity width
- 10: Doelgerichte telefonische consultatie van een poortspecialist door een patiënt bij een al geopende dbc ter vervanging van een fysiek consult
- 11: 11/21 operatief kliniek 404
- 12: Sis echo

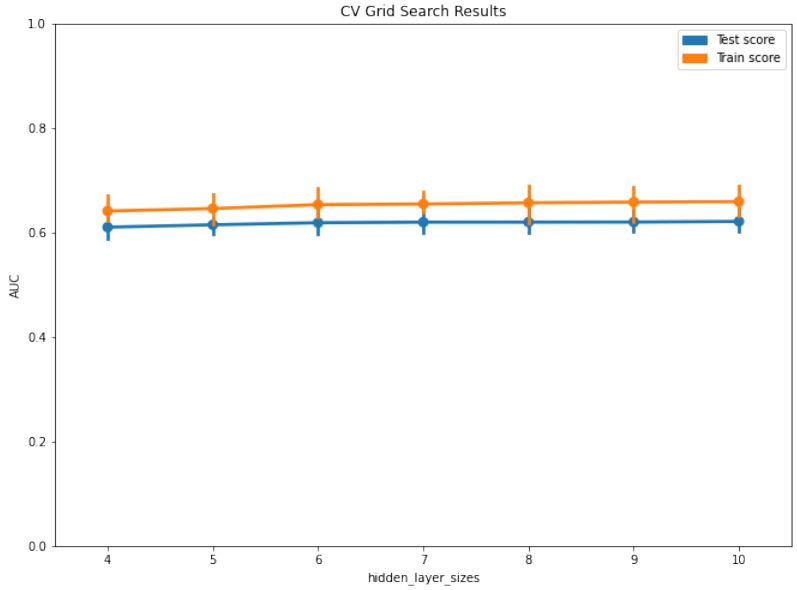


FIGURE A.3: Neural network algorithm tuned on hidden layer sizes

- 13: Controle patient gynaecologie
- 14: Dagverpleging i
- 15: Amount of children unknown
- 15: Dysmenorrhea
- 15: Herhalingsbezoek
- 16: Cavity length
- 16: Dagverpleging
- 16: Three or more children
- 17: Anteverted uterus
- 17: One child
- 18: Controle patiënt
- 19: Echo op verzoek huisarts gynaecologisch
- 19: Nieuwe patient abnormaal bloedverlies
- 19: Zorgdomein abnormaal bloedverlies
- 20: Local anaesthetic
- 20: Beoordeling ecg holter inspanningsonderzoek ed
- 20: Poliklinische behandeling
- 21: Two children
- 21: Uteruscuretteage exclusief diagnostische microcurettage (endometrium-sapmling zoals pipell vabra milex novak)

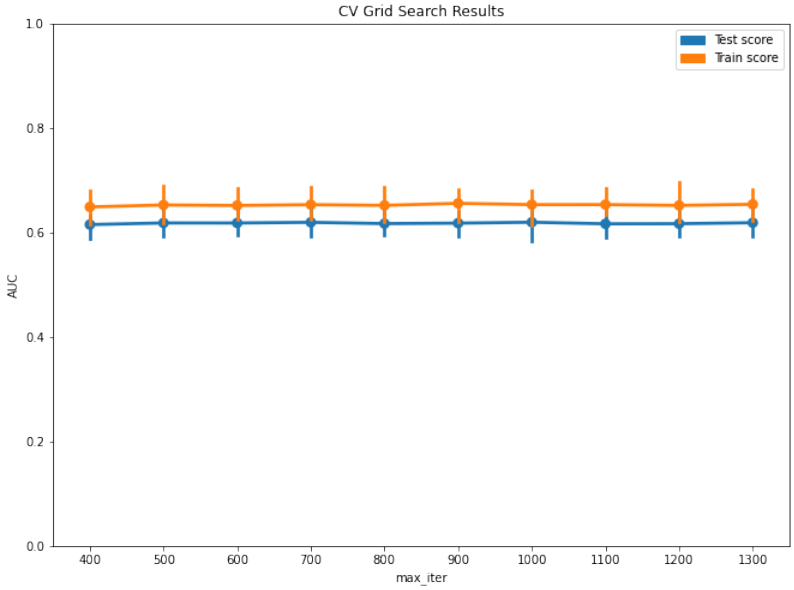


FIGURE A.4: Neural network algorithm tuned on maximum number of iterations

- 22: Geen uitval standaard cyclusstoornissen intensieve/invasieve therapie geen oper groep 3 geen oper groep 2 open oper groep 2 endoscopisch geen oper groep 1 diagnostisch specifiek/ gynaecol
- 22: Nieuwe patient plaatsen iud

Appendix B

Feature importances per algorithm per experiment

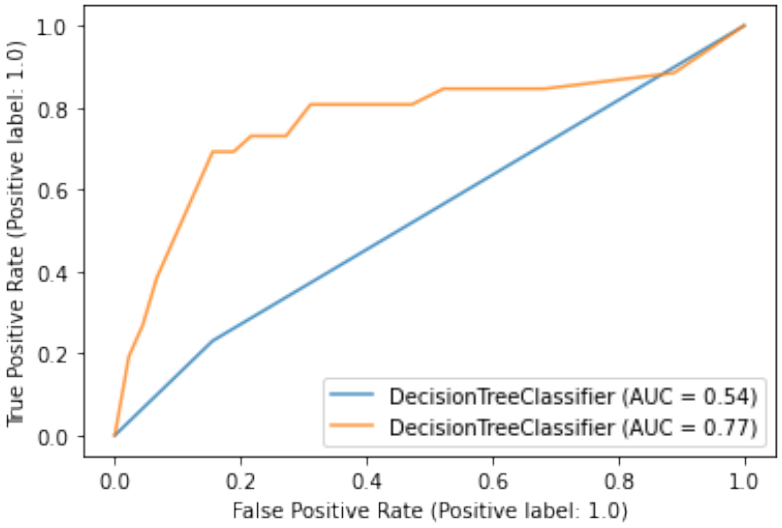


FIGURE B.1: AUC's for decision tree on imbalanced dataset - patient characteristics, with the default hyperparameters in blue and the tuned hyperparameters in orange.

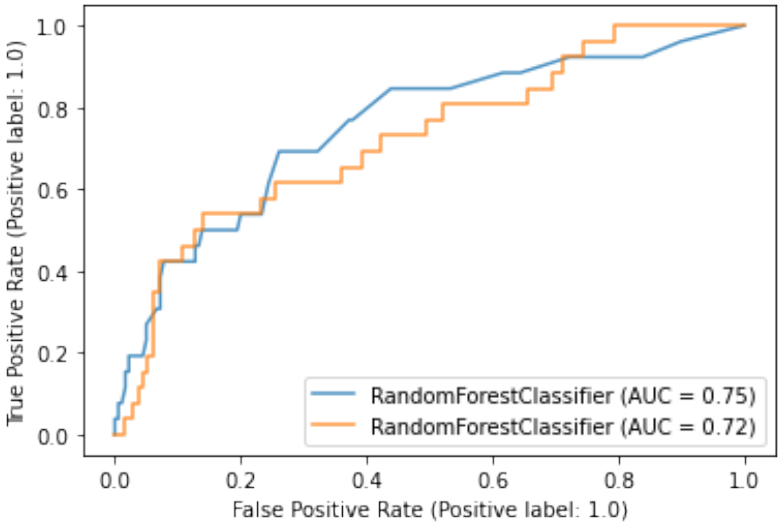


FIGURE B.2: AUC's for random forest on imbalanced dataset - patient characteristics, with the default hyperparameters in blue and the tuned hyperparameters in orange.

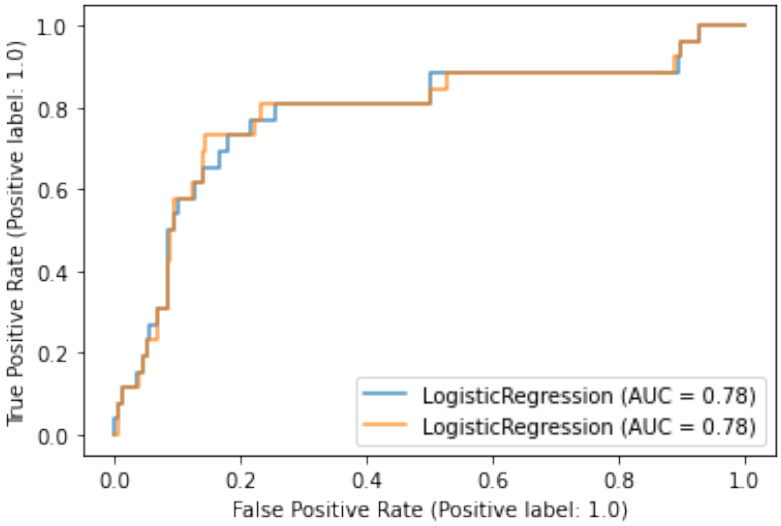


FIGURE B.3: AUC's for logistic regression on imbalanced dataset - patient characteristics, with the default hyperparameters in blue and the tuned hyperparameters in orange.

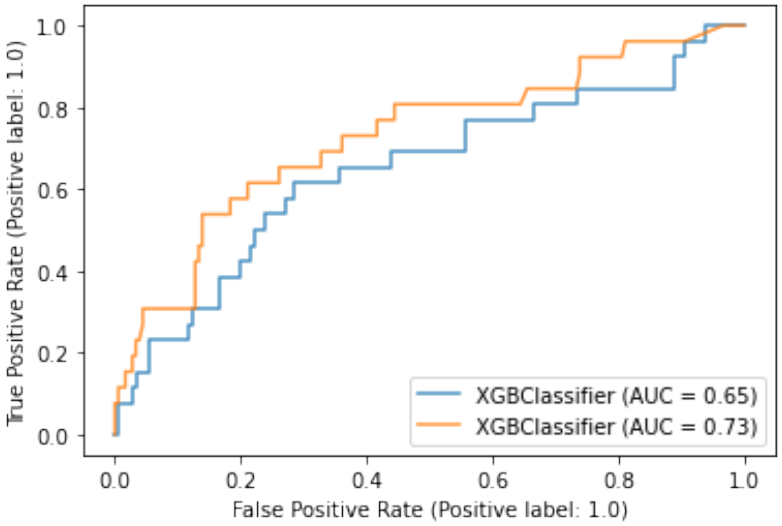


FIGURE B.4: AUC's for extreme gradient boosting on imbalanced dataset - patient characteristics, with the default hyperparameters in blue and the tuned hyperparameters in orange.

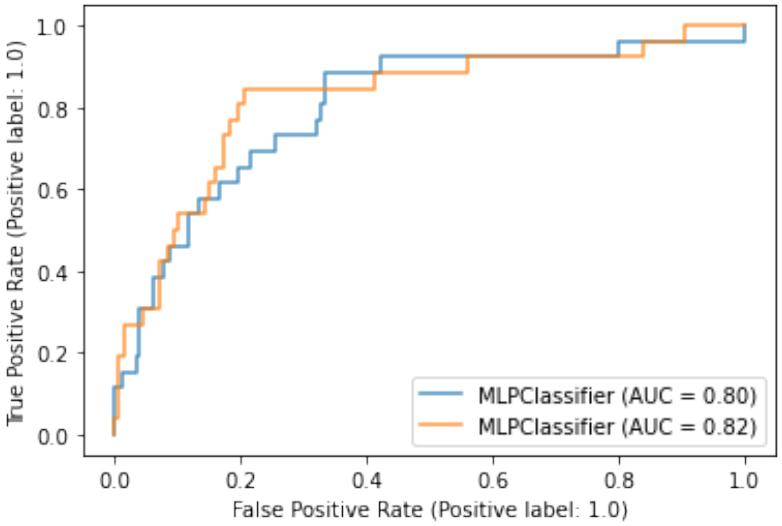


FIGURE B.5: AUC's for neural network on imbalanced dataset - patient characteristics, with the default hyperparameters in blue and the tuned hyperparameters in orange.

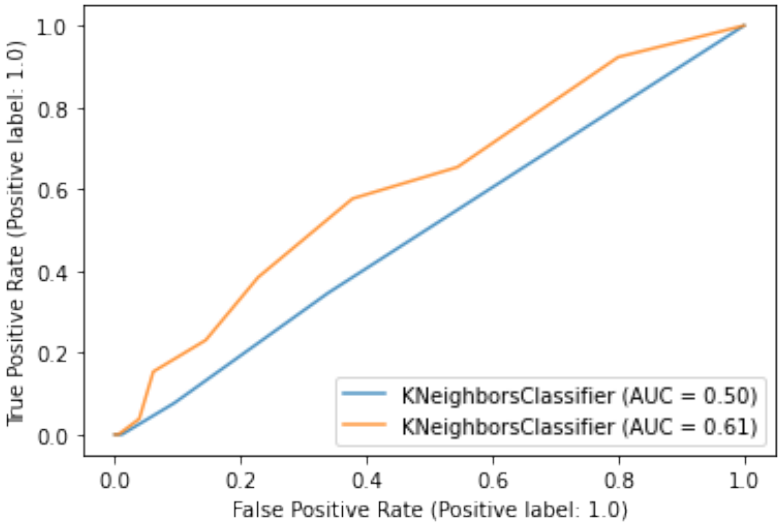


FIGURE B.6: AUC's for k-nearest neighbour on imbalanced dataset - patient characteristics, with the default hyperparameters in blue and the tuned hyperparameters in orange.

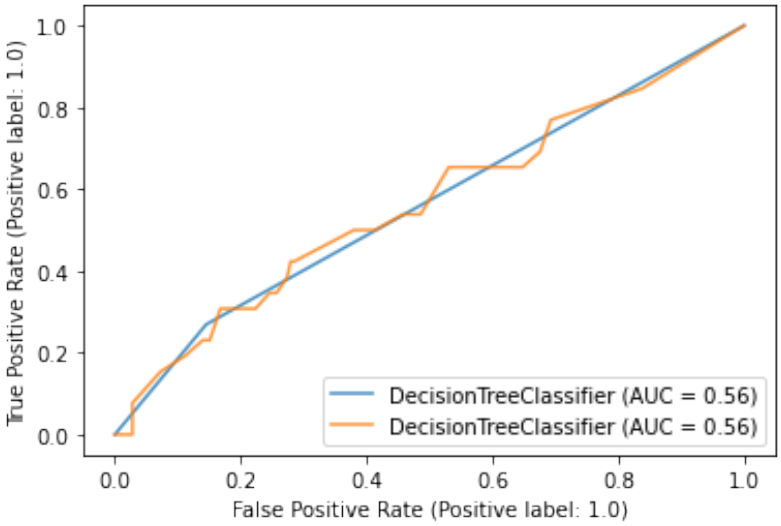


FIGURE B.7: AUC's for decision tree on imbalanced dataset - patient characteristics and process features, with the default hyperparameters in blue and the tuned hyperparameters in orange.

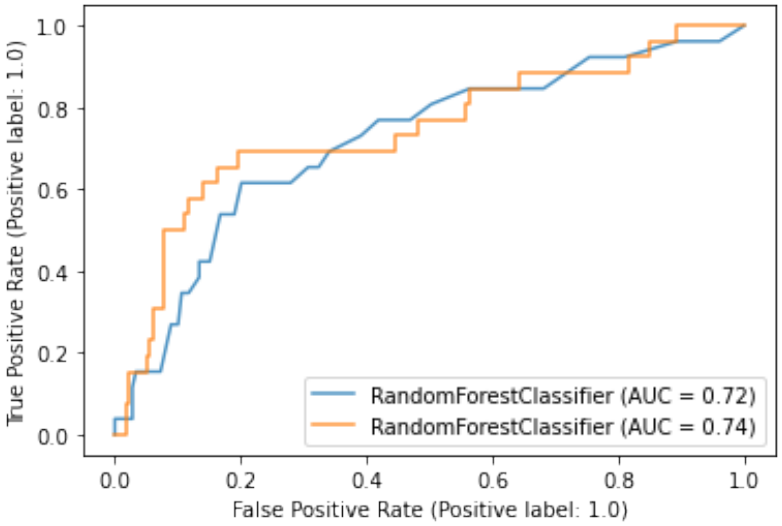


FIGURE B.8: AUC's for random forest on imbalanced dataset - patient characteristics and process features, with the default hyperparameters in blue and the tuned hyperparameters in orange.

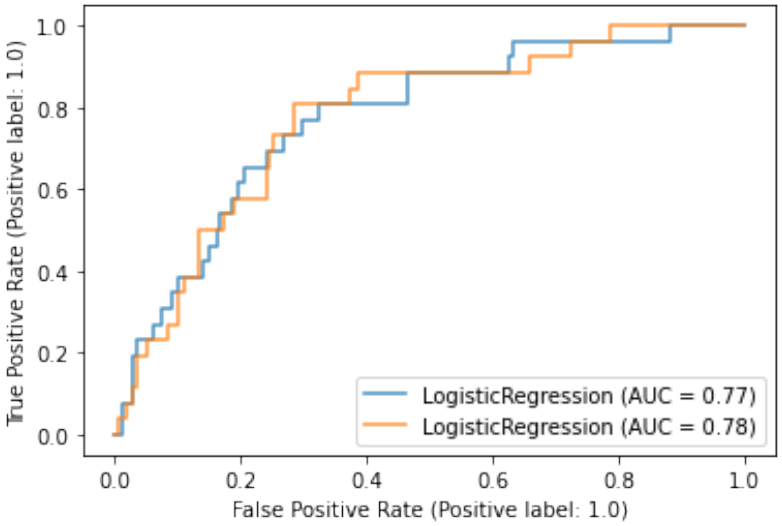


FIGURE B.9: AUC's for logistic regression on imbalanced dataset - patient characteristics and process features, with the default hyperparameters in blue and the tuned hyperparameters in orange.

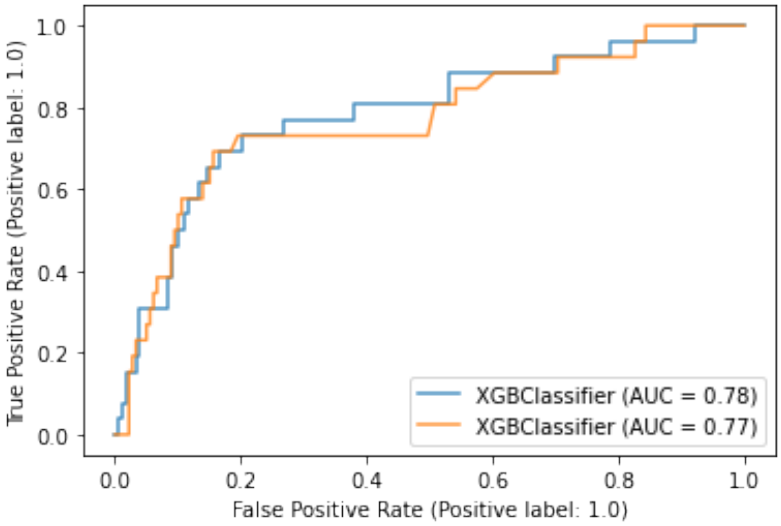


FIGURE B.10: AUC's for extreme gradient boosting on imbalanced dataset - patient characteristics and process features, with the default hyperparameters in blue and the tuned hyperparameters in orange.

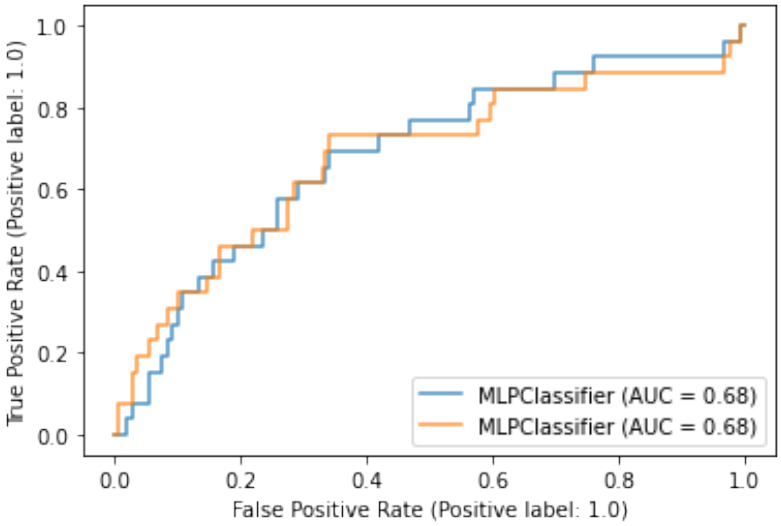


FIGURE B.11: AUC's for neural network on imbalanced dataset - patient characteristics and process features, with the default hyperparameters in blue and the tuned hyperparameters in orange.

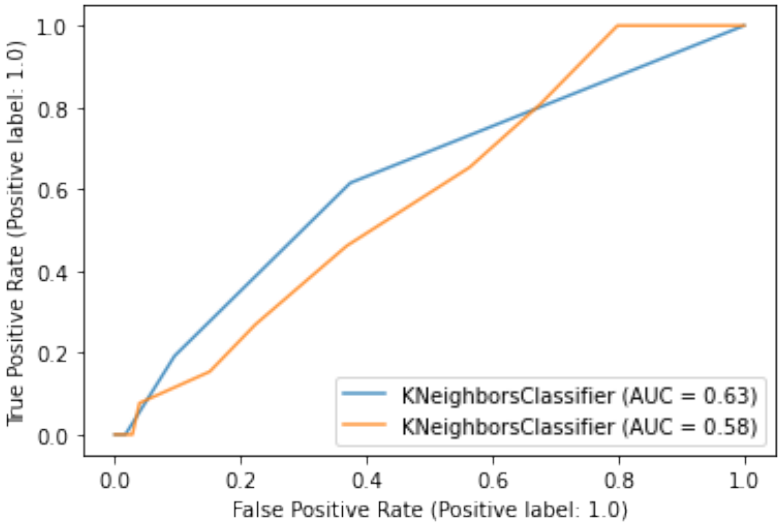


FIGURE B.12: AUC's for k-nearest neighbour on imbalanced dataset - patient characteristics and process features, with the default hyperparameters in blue and the tuned hyperparameters in orange.

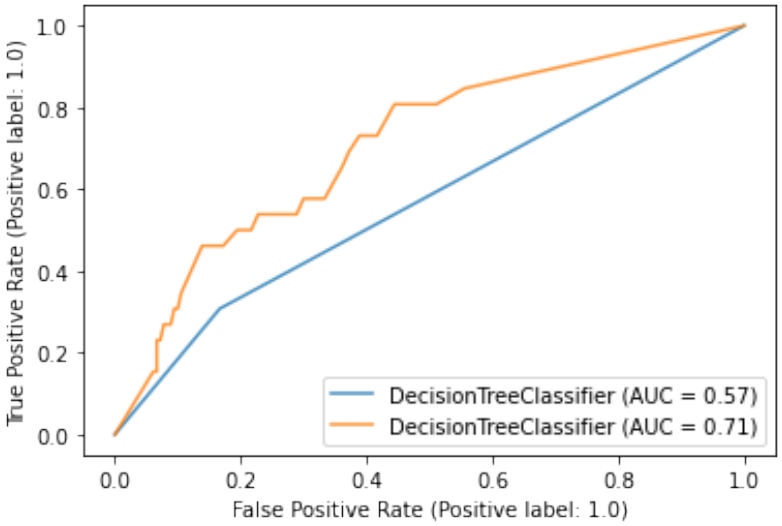


FIGURE B.13: AUC's for decision tree on balanced dataset - patient characteristics, with the default hyperparameters in blue and the tuned hyperparameters in orange.

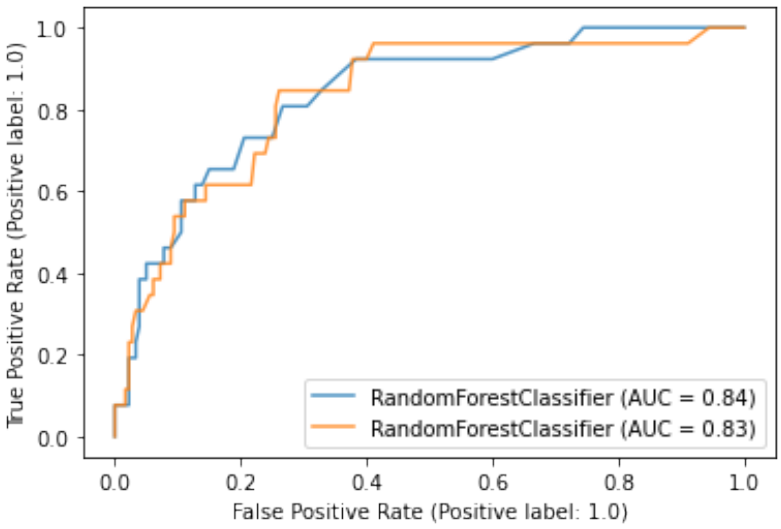


FIGURE B.14: AUC's for random forest on balanced dataset - patient characteristics

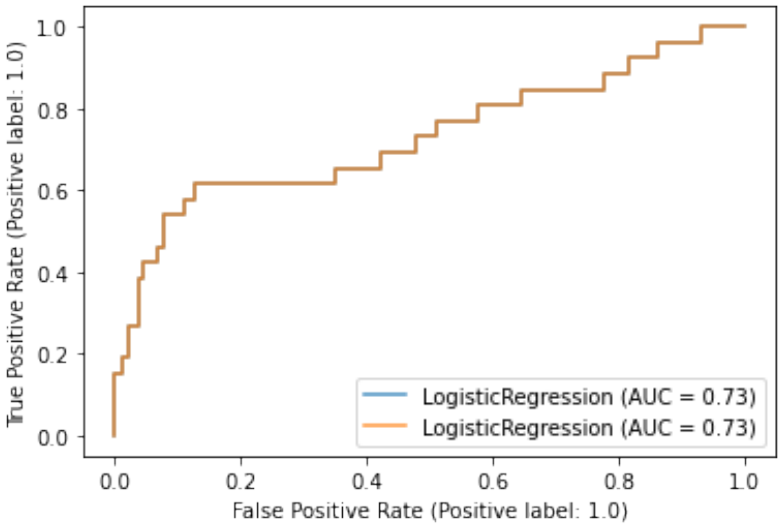


FIGURE B.15: AUC's for logistic regression on balanced dataset - patient characteristics, with the default hyperparameters in blue and the tuned hyperparameters in orange.

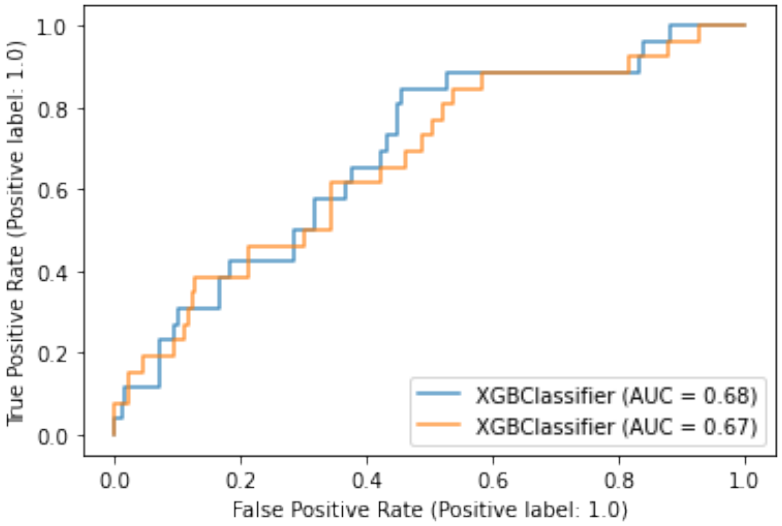


FIGURE B.16: AUC's for extreme gradient boosting on balanced dataset - patient characteristics, with the default hyperparameters in blue and the tuned hyperparameters in orange.

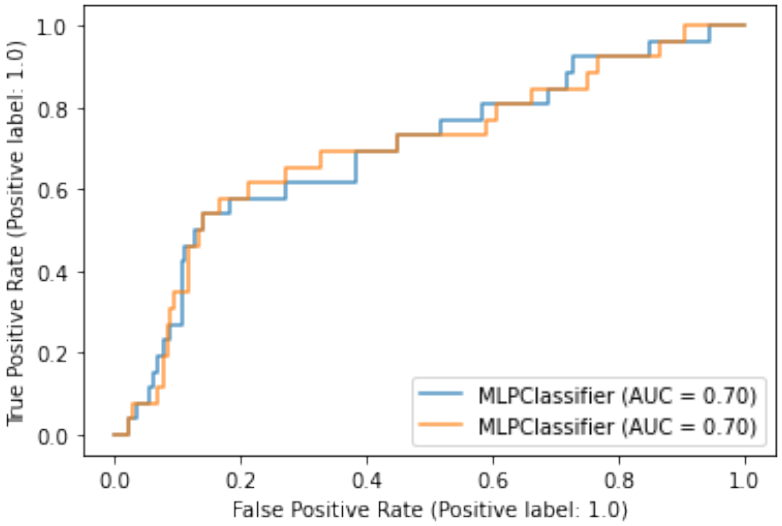


FIGURE B.17: AUC's for neural network on balanced dataset - patient characteristics, with the default hyperparameters in blue and the tuned hyperparameters in orange.

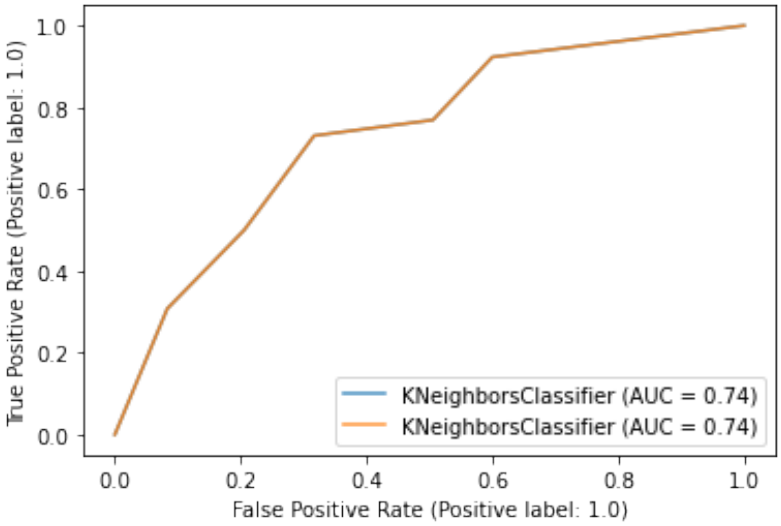


FIGURE B.18: AUC's for k-nearest neighbour on balanced dataset - patient characteristics, with the default hyperparameters in blue and the tuned hyperparameters in orange.

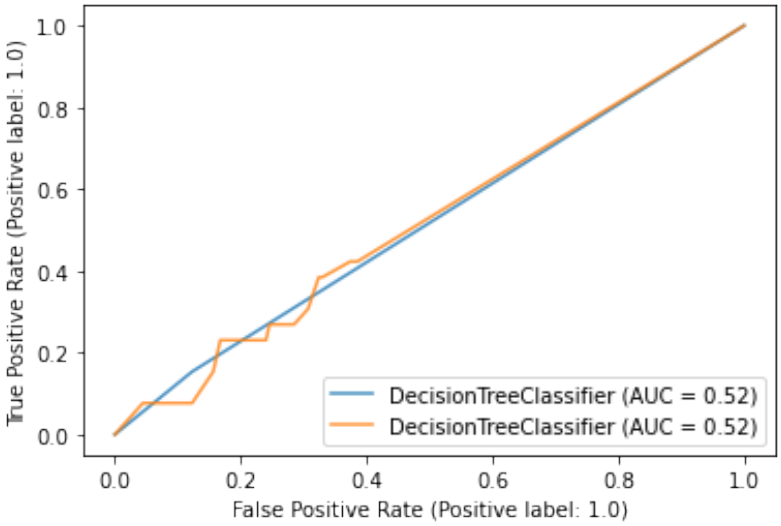


FIGURE B.19: AUC's for decision tree on balanced dataset - patient characteristics and process features, with the default hyperparameters in blue and the tuned hyperparameters in orange.

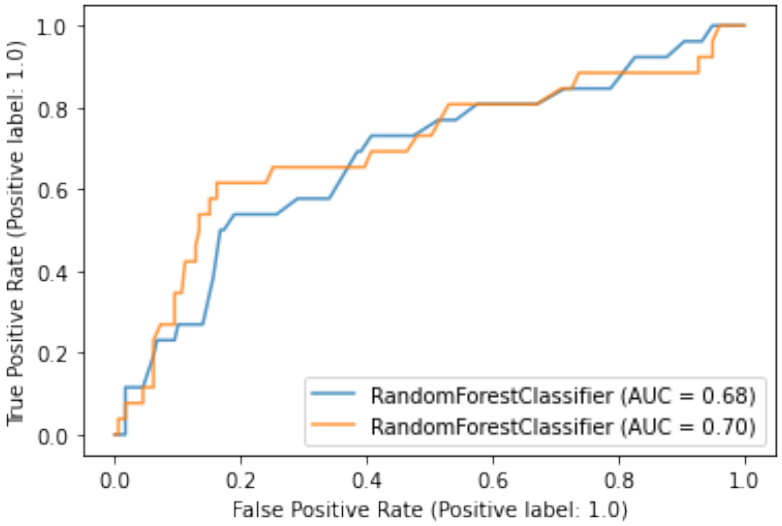


FIGURE B.20: AUC's for random forest on balanced dataset - patient characteristics and process features, with the default hyperparameters in blue and the tuned hyperparameters in orange

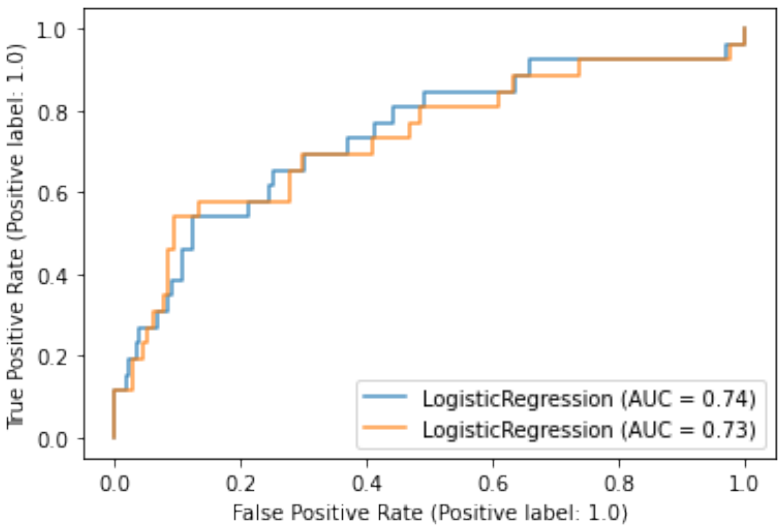


FIGURE B.21: AUC's for logistic regression on balanced dataset - patient characteristics and process features, with the default hyperparameters in blue and the tuned hyperparameters in orange

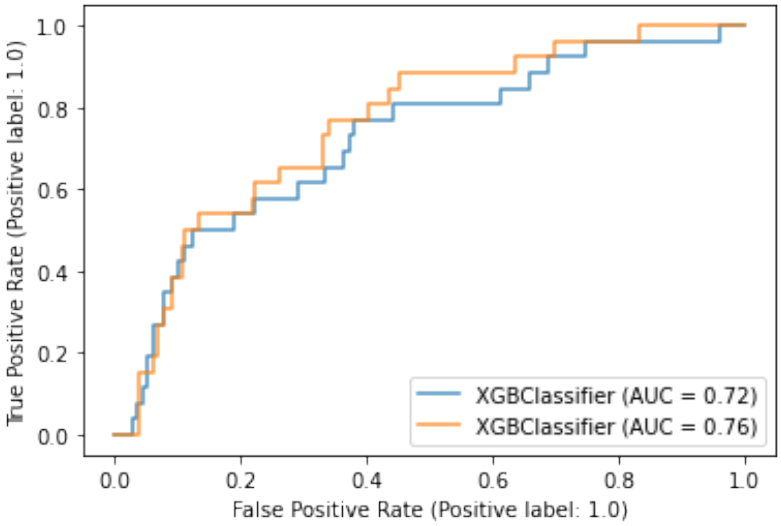


FIGURE B.22: AUC's for extreme gradient boosting on balanced dataset - patient characteristics and process features, with the default hyperparameters in blue and the tuned hyperparameters in orange

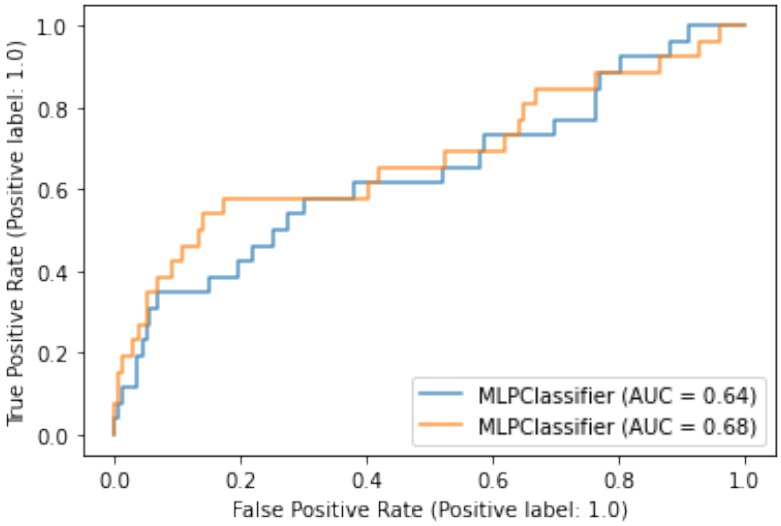


FIGURE B.23: AUC's for neural network on balanced dataset - patient characteristics and process features, with the default hyperparameters in blue and the tuned hyperparameters in orange

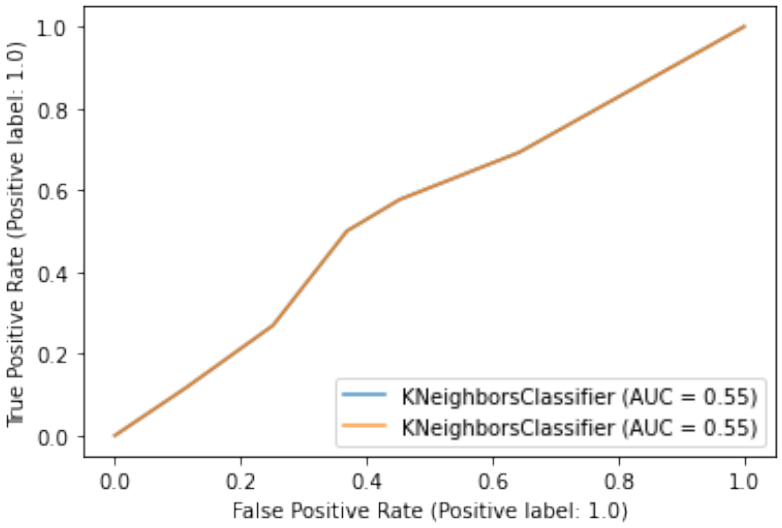
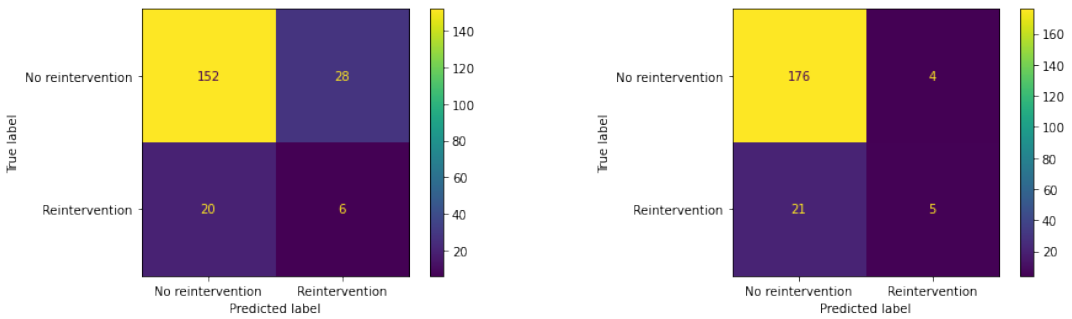


FIGURE B.24: AUC's for k-nearest neighbour on balanced dataset - patient characteristics and process features, with the default hyperparameters in blue and the tuned hyperparameters in orange

Appendix C

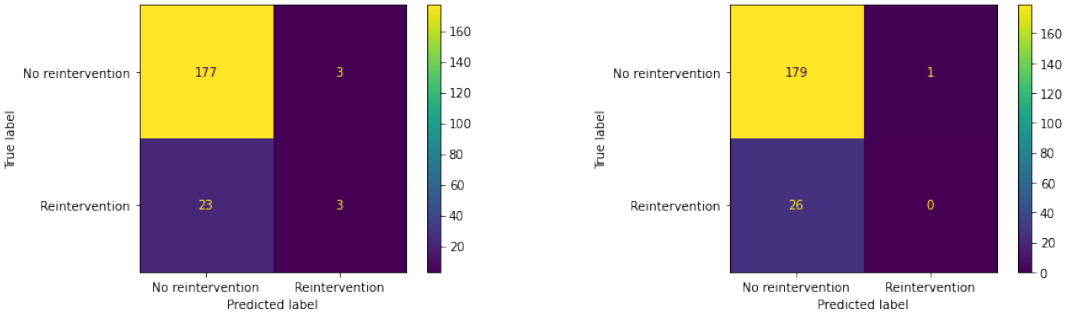
Confusion matrices



(A) Default hyperparameters

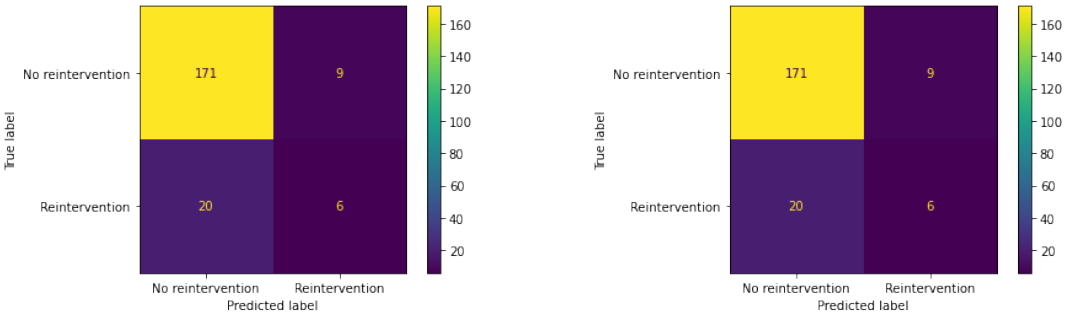
(B) Tuned hyperparameters

FIGURE C.1: Confusion matrices for decision tree on original dataset - patient characteristics



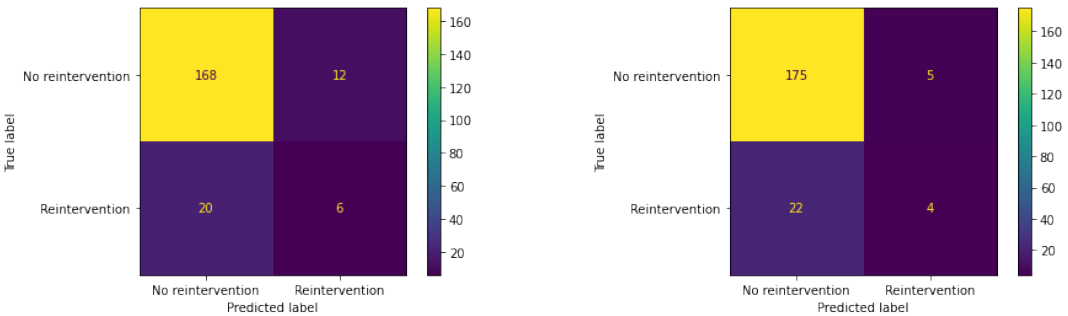
(A) Default hyperparameters (B) Tuned hyperparameters

FIGURE C.2: Confusion matrices for random forest on original dataset - patient characteristics



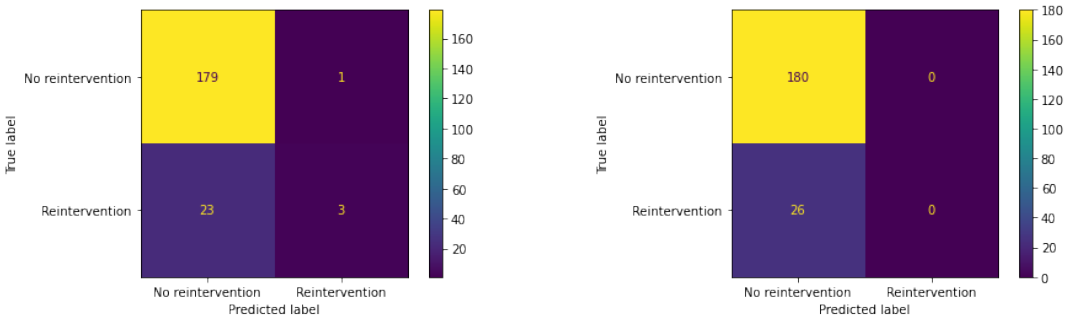
(A) Default hyperparameters (B) Tuned hyperparameters

FIGURE C.3: Confusion matrices for logistic regression on original dataset - patient characteristics



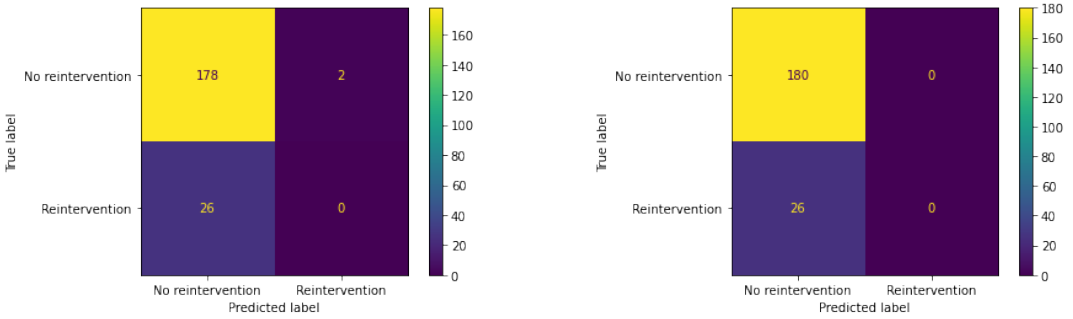
(A) Default hyperparameters (B) Tuned hyperparameters

FIGURE C.4: Confusion matrices for extreme gradient boosting on original dataset - patient characteristics



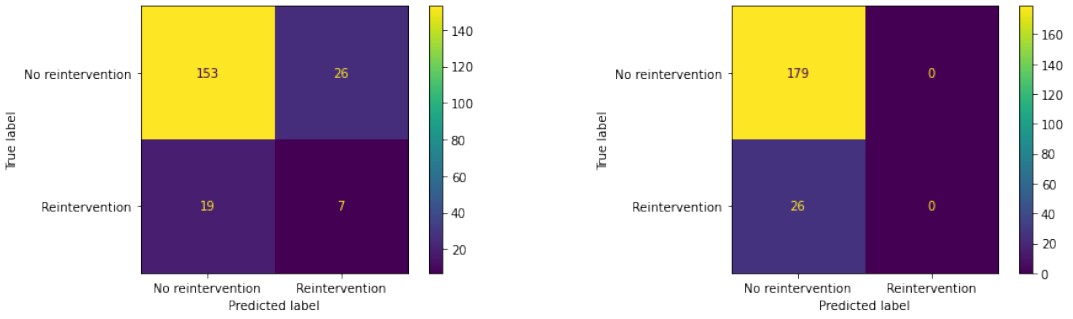
(A) Default hyperparameters (B) Tuned hyperparameters

FIGURE C.5: Confusion matrices for neural network on original dataset - patient characteristics



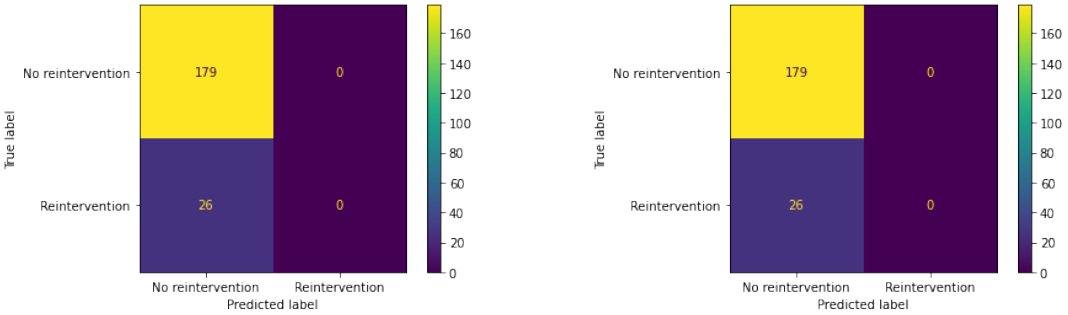
(A) Default hyperparameters (B) Tuned hyperparameters

FIGURE C.6: Confusion matrices for k-nearest neighbour on original dataset - patient characteristics



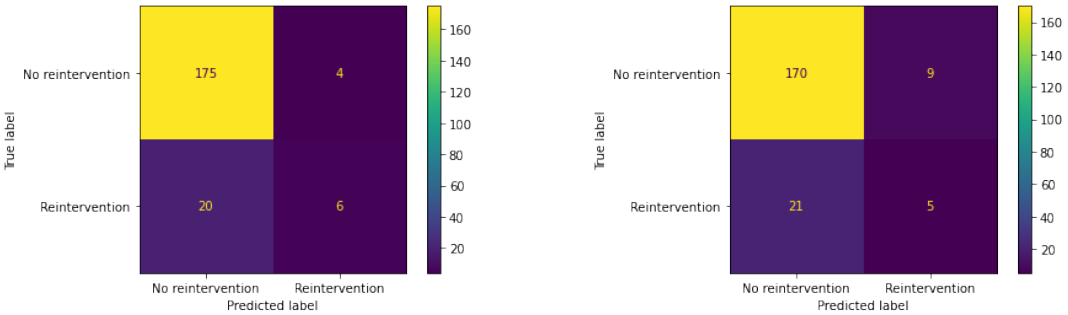
(A) Default hyperparameters (B) Tuned hyperparameters

FIGURE C.7: Confusion matrices for decision tree on original dataset - patient characteristics and process features



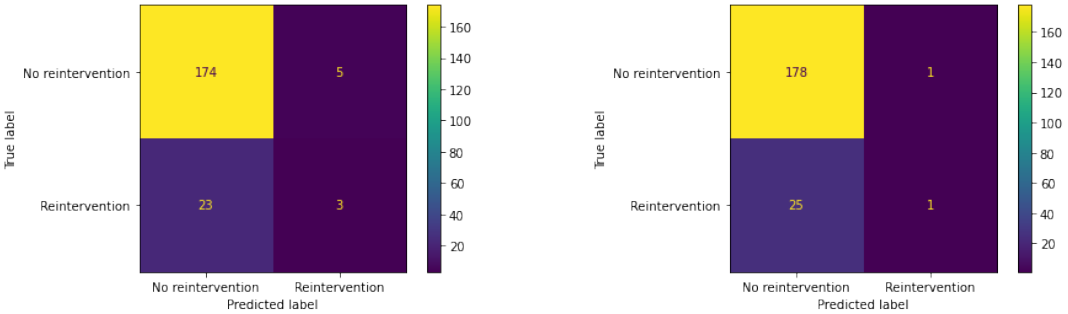
(A) Default hyperparameters (B) Tuned hyperparameters

FIGURE C.8: Confusion matrices for random forest on original dataset - patient characteristics and process features



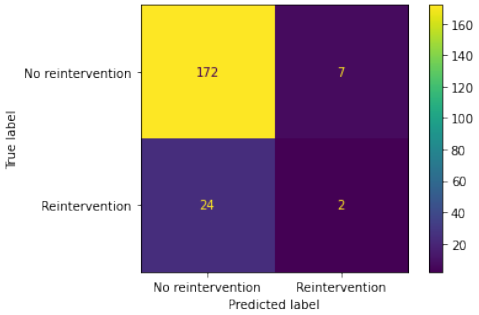
(A) Default hyperparameters (B) Tuned hyperparameters

FIGURE C.9: Confusion matrices for logistic regression on original dataset - patient characteristics and process features

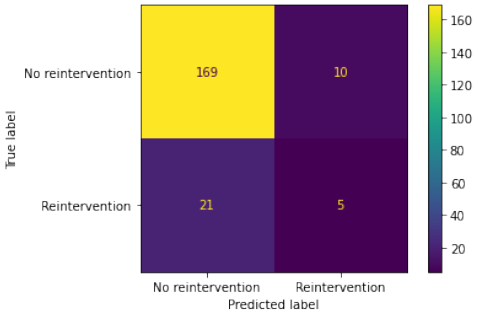


(A) Default hyperparameters (B) Tuned hyperparameters

FIGURE C.10: Confusion matrices for extreme gradient boosting on original dataset - patient characteristics and process features

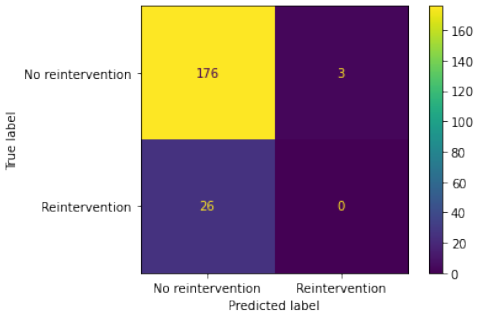


(A) Default hyperparameters

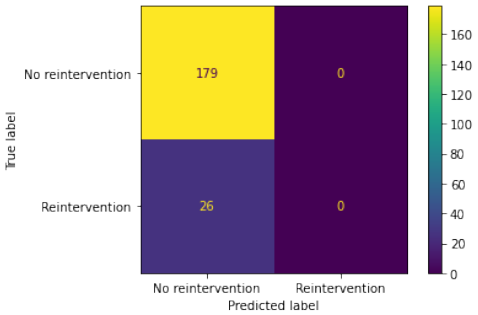


(B) Tuned hyperparameters

FIGURE C.11: Confusion matrices for neural network on original dataset - patient characteristics and process features

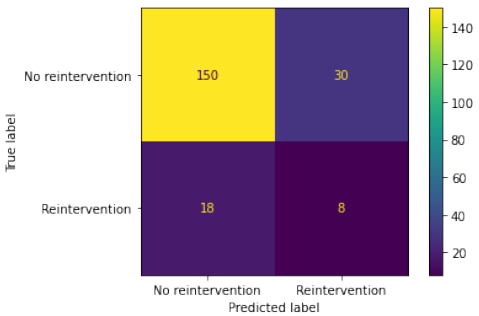


(A) Default hyperparameters

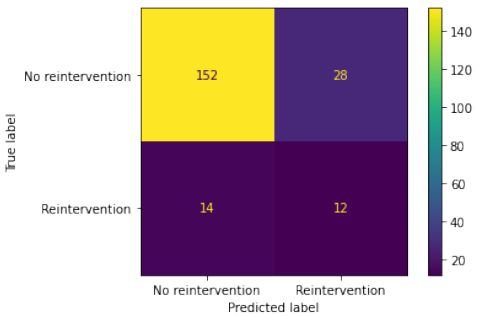


(B) Tuned hyperparameters

FIGURE C.12: Confusion matrices for k-nearest neighbour on original dataset - patient characteristics and process features

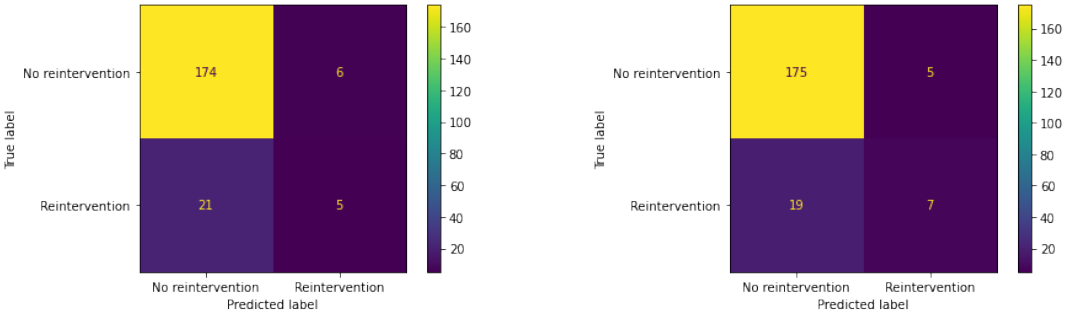


(A) Default hyperparameters



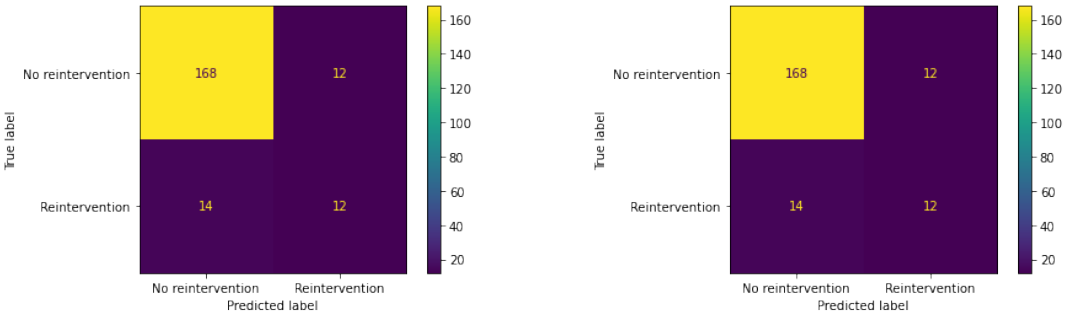
(B) Tuned hyperparameters

FIGURE C.13: Confusion matrices for decision tree on balanced dataset - patient characteristics



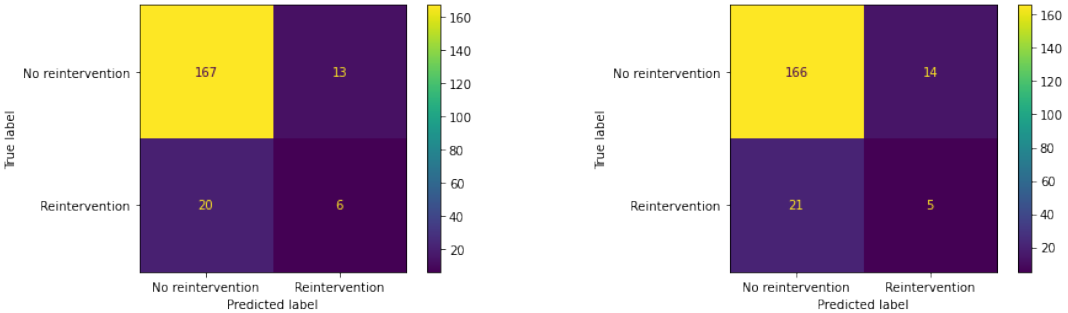
(A) Default hyperparameters (B) Tuned hyperparameters

FIGURE C.14: Confusion matrices for random forest on balanced dataset - patient characteristics



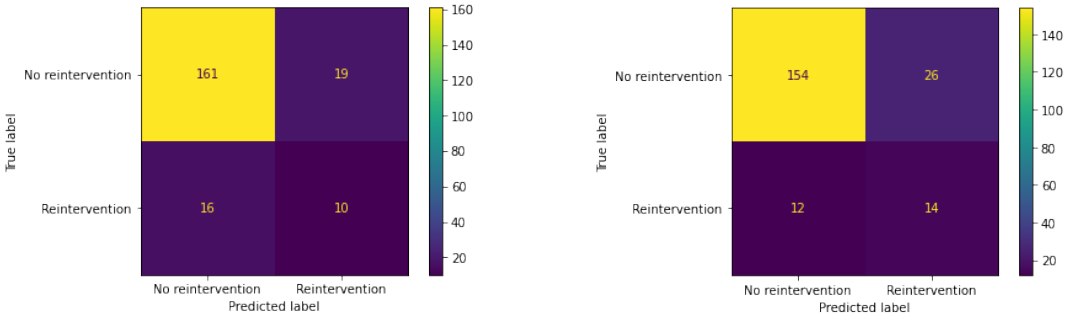
(A) Default hyperparameters (B) Tuned hyperparameters

FIGURE C.15: Confusion matrices for logistic regression on balanced dataset - patient characteristics



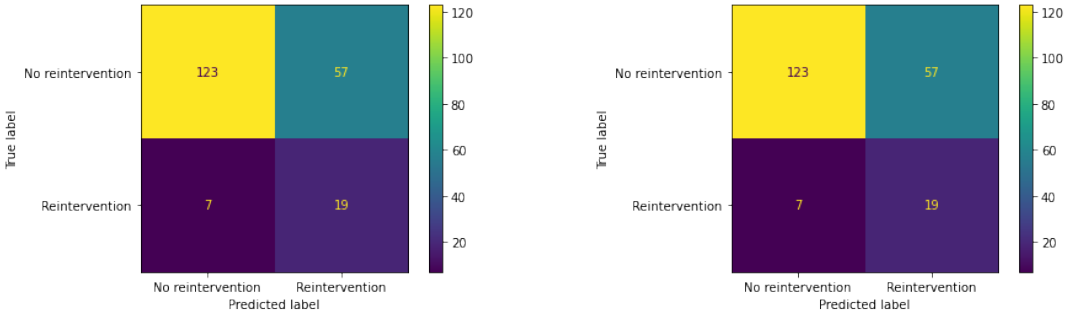
(A) Default hyperparameters (B) Tuned hyperparameters

FIGURE C.16: Confusion matrices for extreme gradient boosting on balanced dataset - patient characteristics



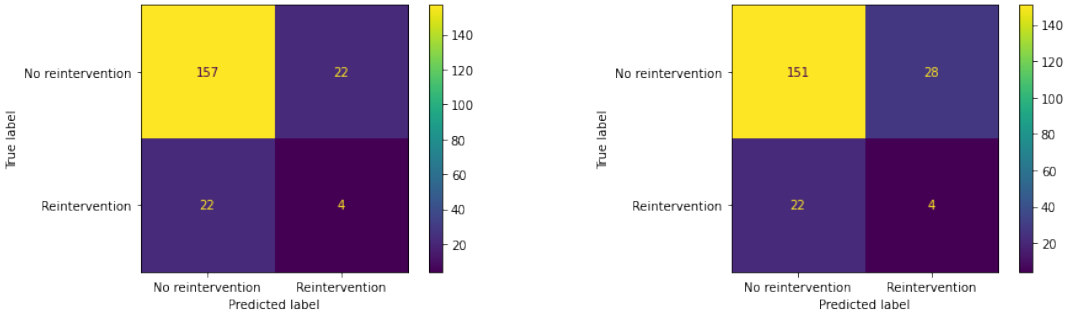
(A) Default hyperparameters (B) Tuned hyperparameters

FIGURE C.17: Confusion matrices for neural network on balanced dataset - patient characteristics



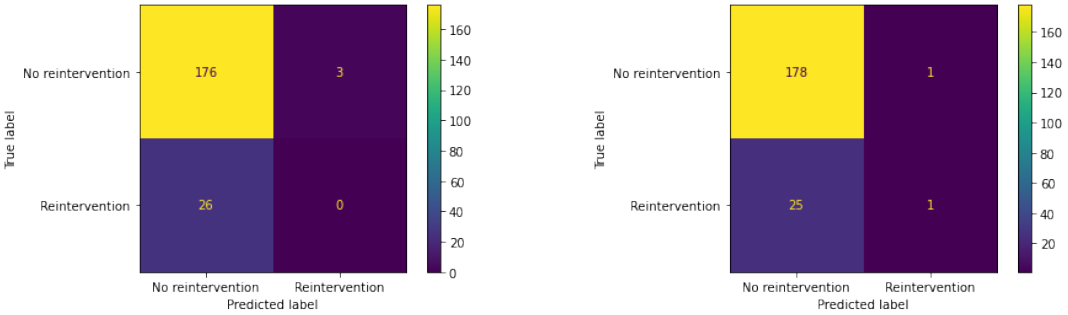
(A) Default hyperparameters (B) Tuned hyperparameters

FIGURE C.18: Confusion matrices for k-nearest neighbour on balanced dataset - patient characteristics



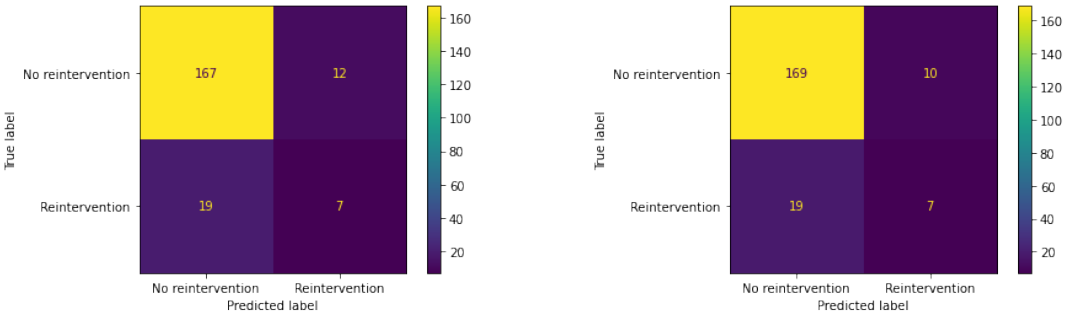
(A) Default hyperparameters (B) Tuned hyperparameters

FIGURE C.19: Confusion matrices for decision tree on balanced dataset - patient characteristics and process features



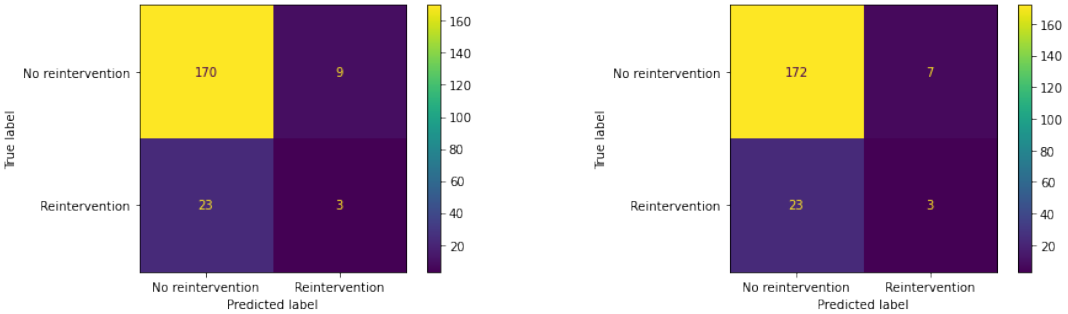
(A) Default hyperparameters (B) Tuned hyperparameters

FIGURE C.20: Confusion matrices for random forest on balanced dataset - patient characteristics and process features



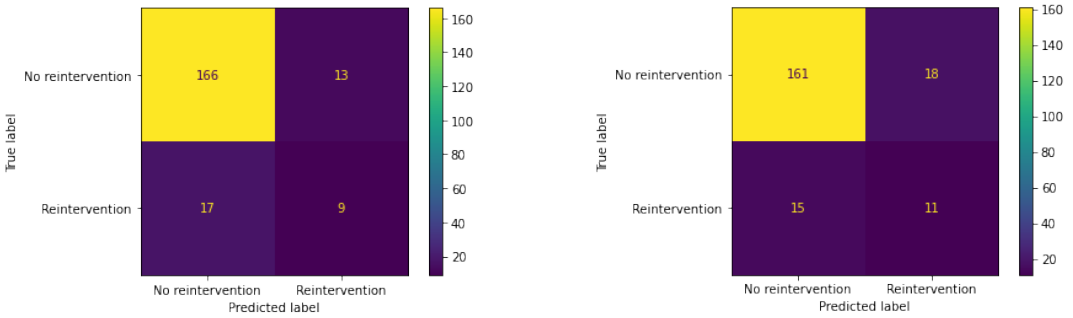
(A) Default hyperparameters (B) Tuned hyperparameters

FIGURE C.21: Confusion matrices for logistic regression on balanced dataset - patient characteristics and process features



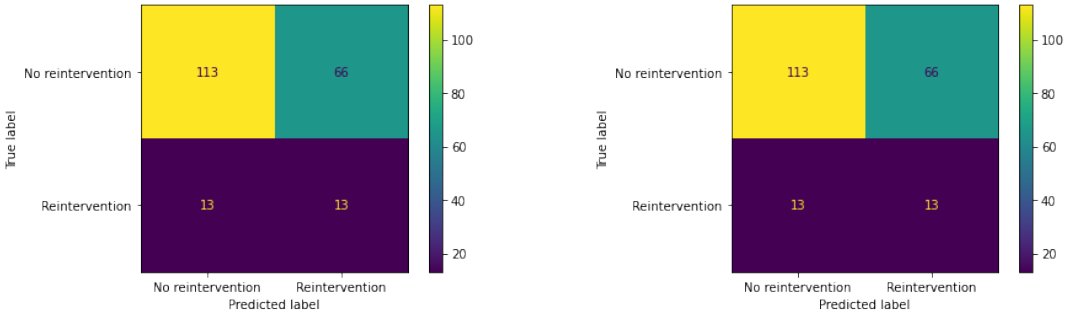
(A) Default hyperparameters (B) Tuned hyperparameters

FIGURE C.22: Confusion matrices for extreme gradient boosting on balanced dataset - patient characteristics and process features



(A) Default hyperparameters (B) Tuned hyperparameters

FIGURE C.23: Confusion matrices for neural network on balanced dataset - patient characteristics and process features



(A) Default hyperparameters (B) Tuned hyperparameters

FIGURE C.24: Confusion matrices for k-nearest neighbour on balanced dataset - patient characteristics and process features

Appendix D

Feature importances per data input per algorithm

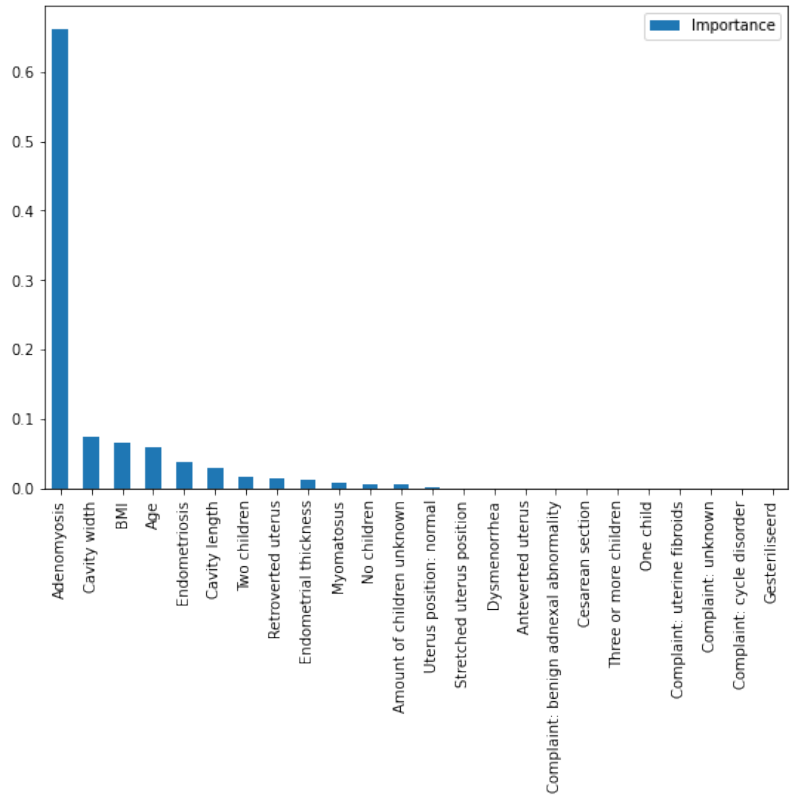


FIGURE D.1: Feature importance for decision tree on original dataset - patient characteristics

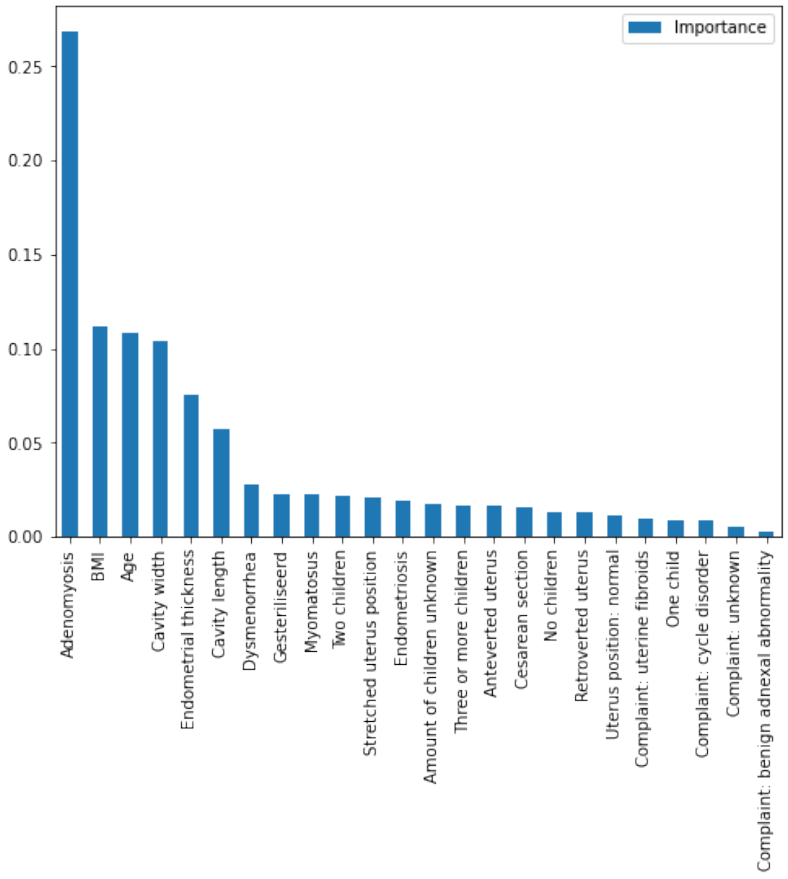


FIGURE D.2: Feature importance for random forest on original dataset - patient characteristics

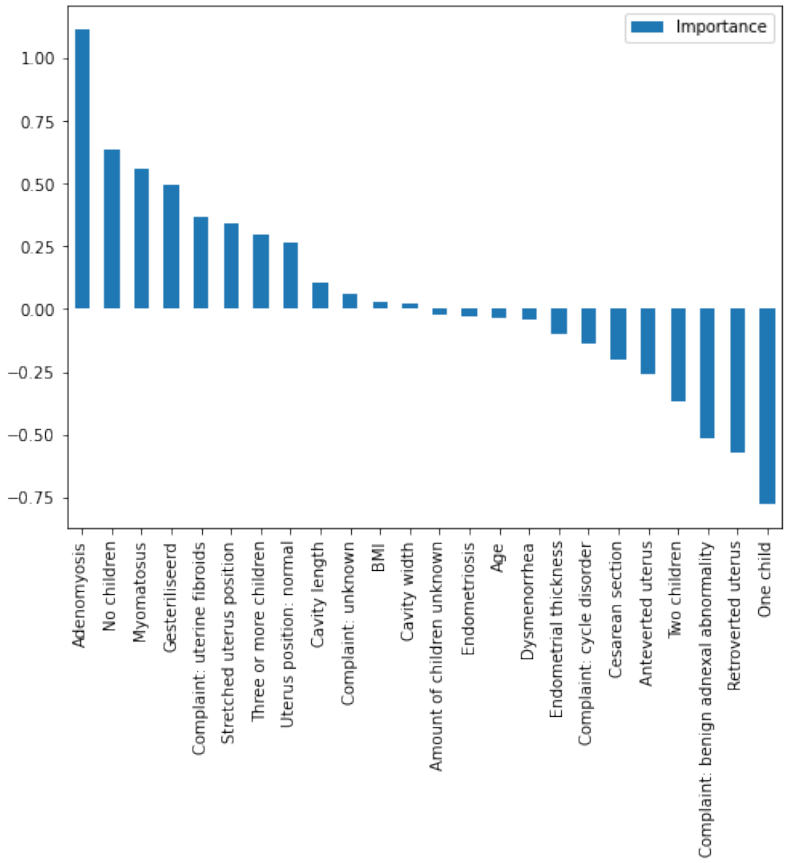


FIGURE D.3: Feature importance for logistic regression on original dataset - patient characteristics

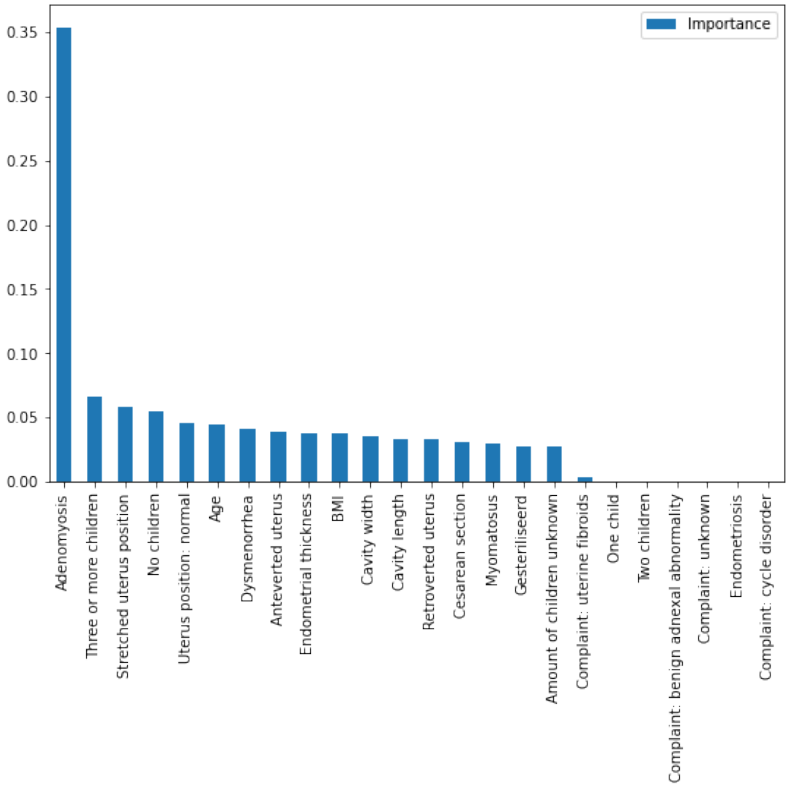


FIGURE D.4: Feature importance for extreme gradient boosting on original dataset - patient characteristics

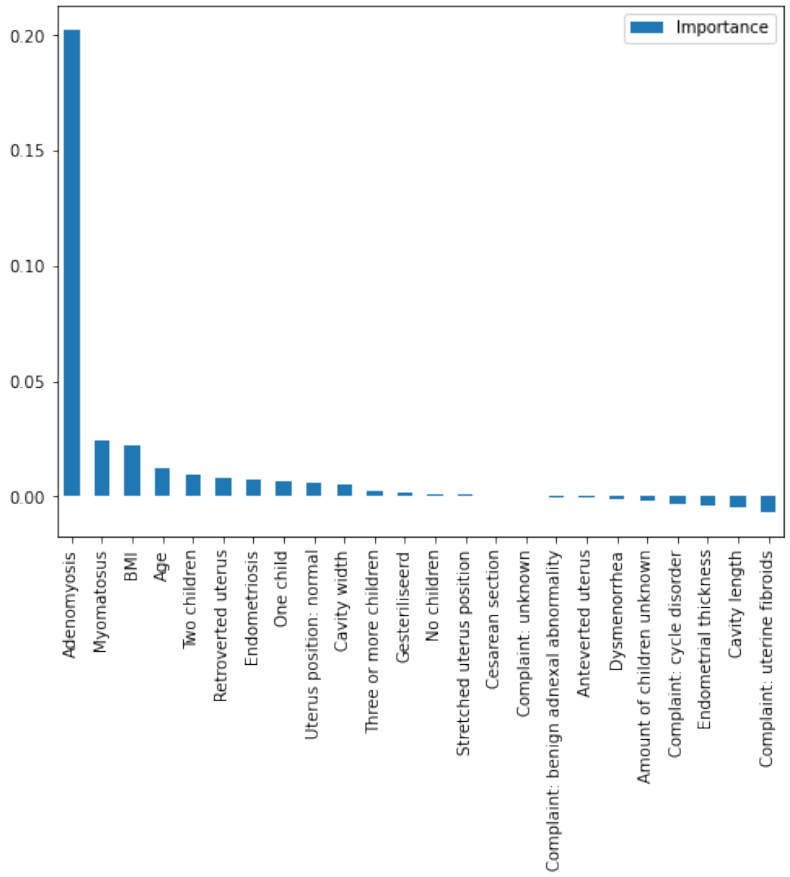


FIGURE D.5: Feature importance for neural network on original dataset - patient characteristics

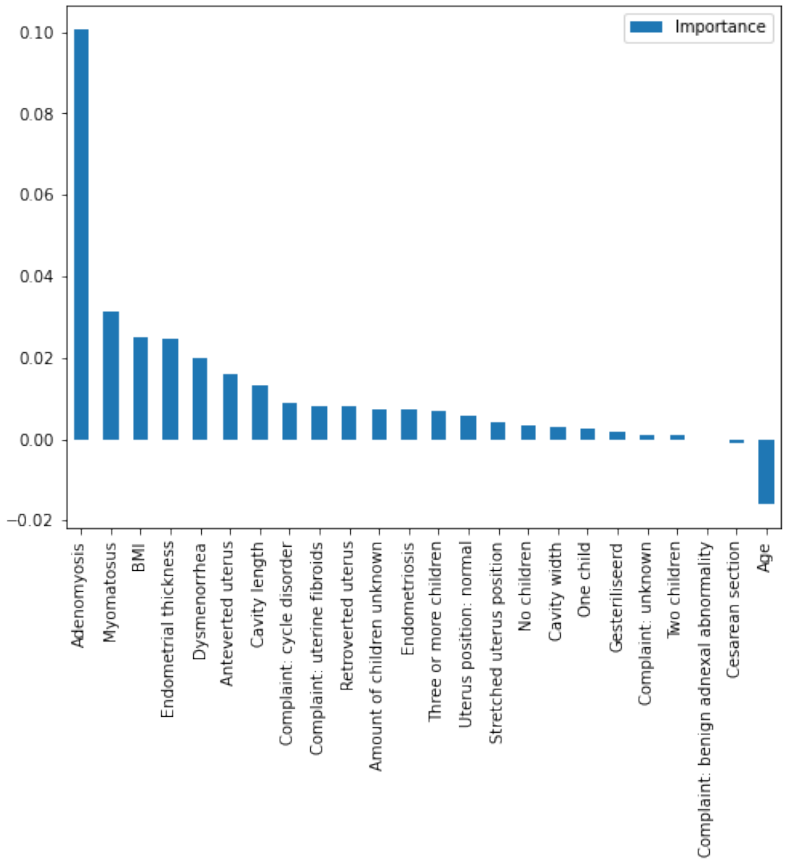


FIGURE D.6: Feature importance for k-nearest neighbour on original dataset - patient characteristics

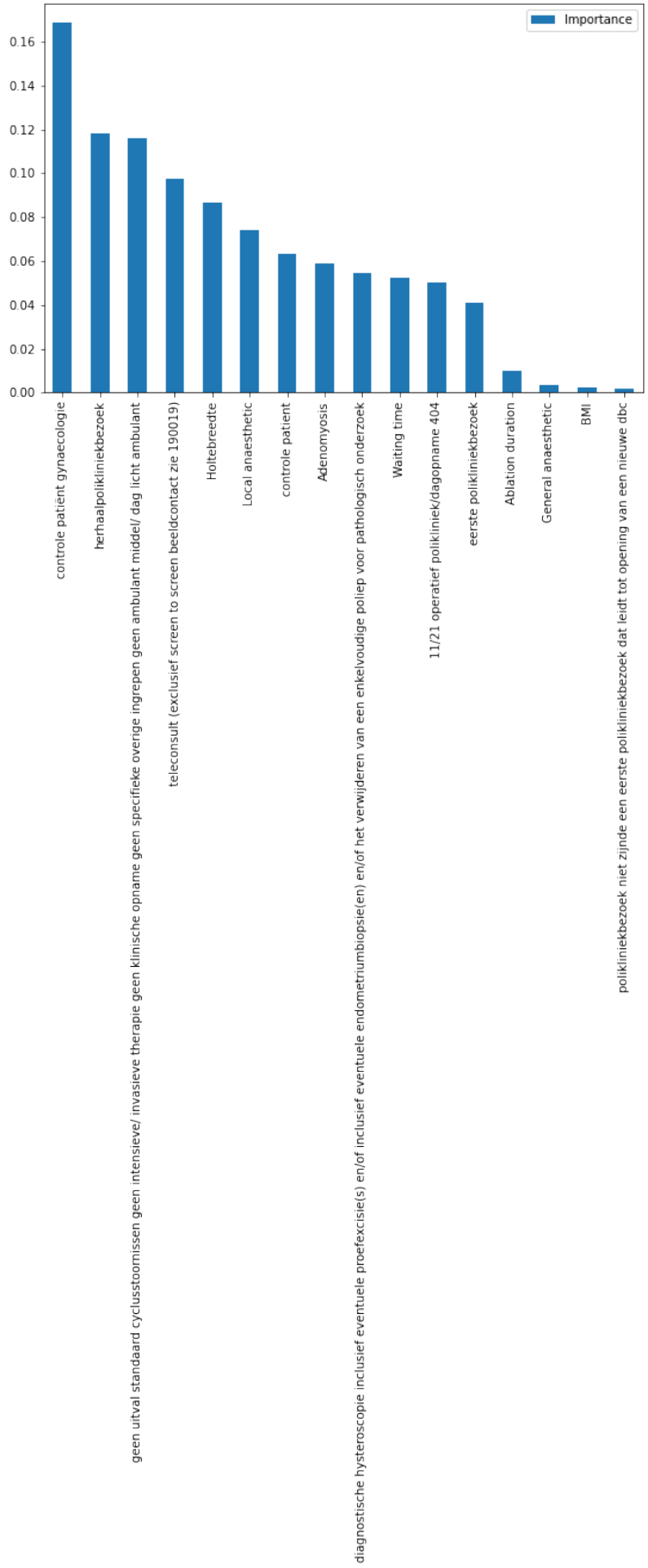


FIGURE D.7: Feature importance for decision tree on original dataset - patient characteristics

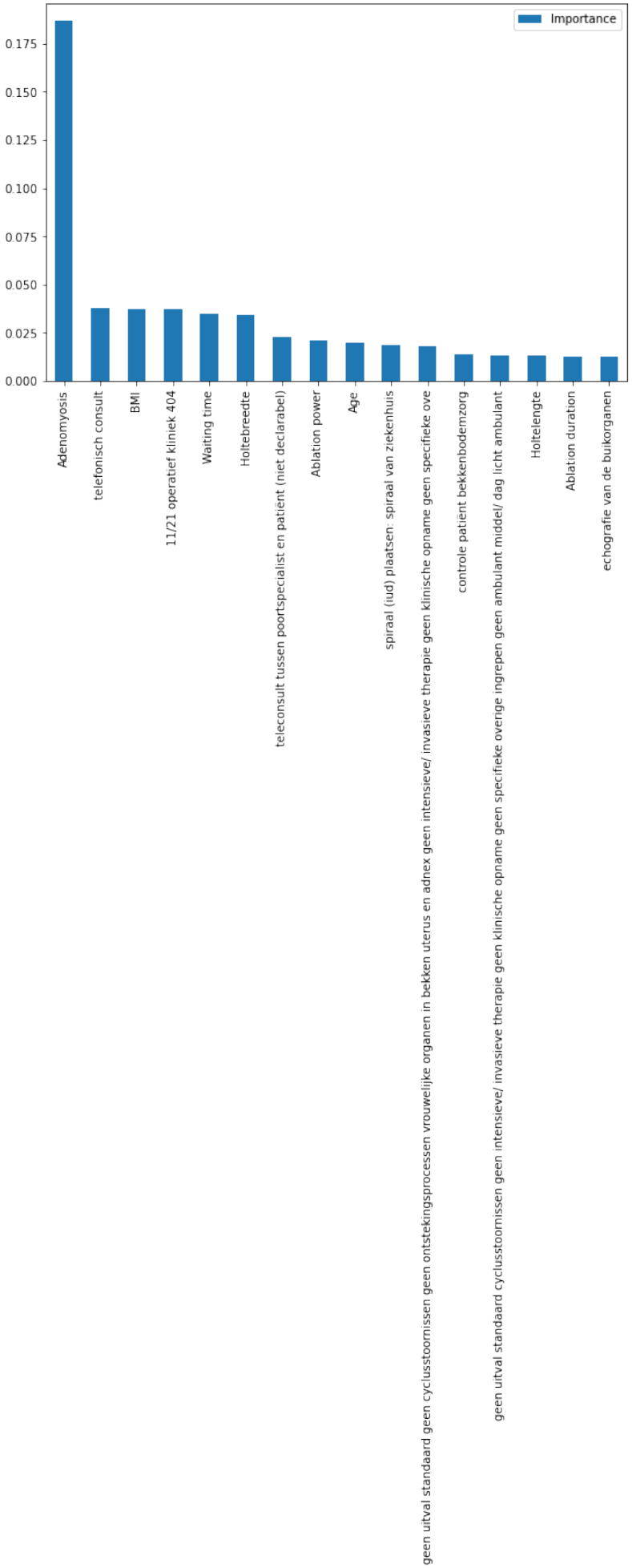
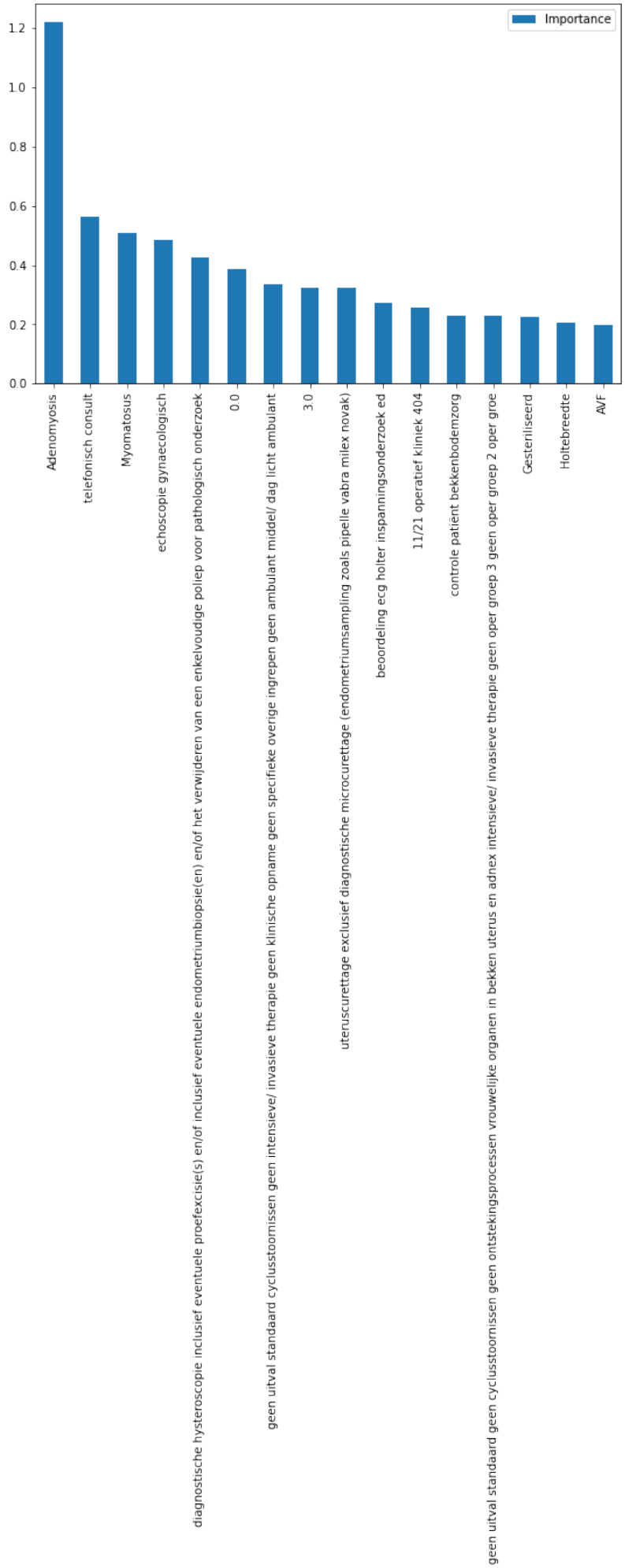


FIGURE D.8: Feature importance for random forest on original


Importance
1.2
1.0
0.8
0.6
0.4
0.2
0.0
Adenomyosis
telefonisch consult
Myomatosus
echoscopie gynaecologisch
diagnostische hysteroscopie inclusief eventuele proefexcisie(s) en/of inclusief eventuele endometriumbiopsie(en) en/of het verwijderen van een enkelvoudige poliep voor pathologisch onderzoek
0 0
geen uitval standaard cyclusstoornissen geen intensieve/ invasieve therapie geen klinische opname geen specifieke overige ingrepen geen ambulant middel/ dag licht ambulant
3 0
uterusc urettage exclusief diagnostische microcurettage (endometriumsampling zoals pipelle vabra milex novak)
beoordeling ecg holter inspanningsonderzoek ed
11/21 operatief kliniek 404
controle patient bekkenbodemzorg
geen uitval standaard geen cyclusstoornissen geen ontstekingsprocessen vrouwelijke organen in bekken uterus en adnex intensieve/ invasieve therapie geen oper groep 3 geen oper groep 2 oper groe
Gesteriliseerd
Holtebreedte
AVF

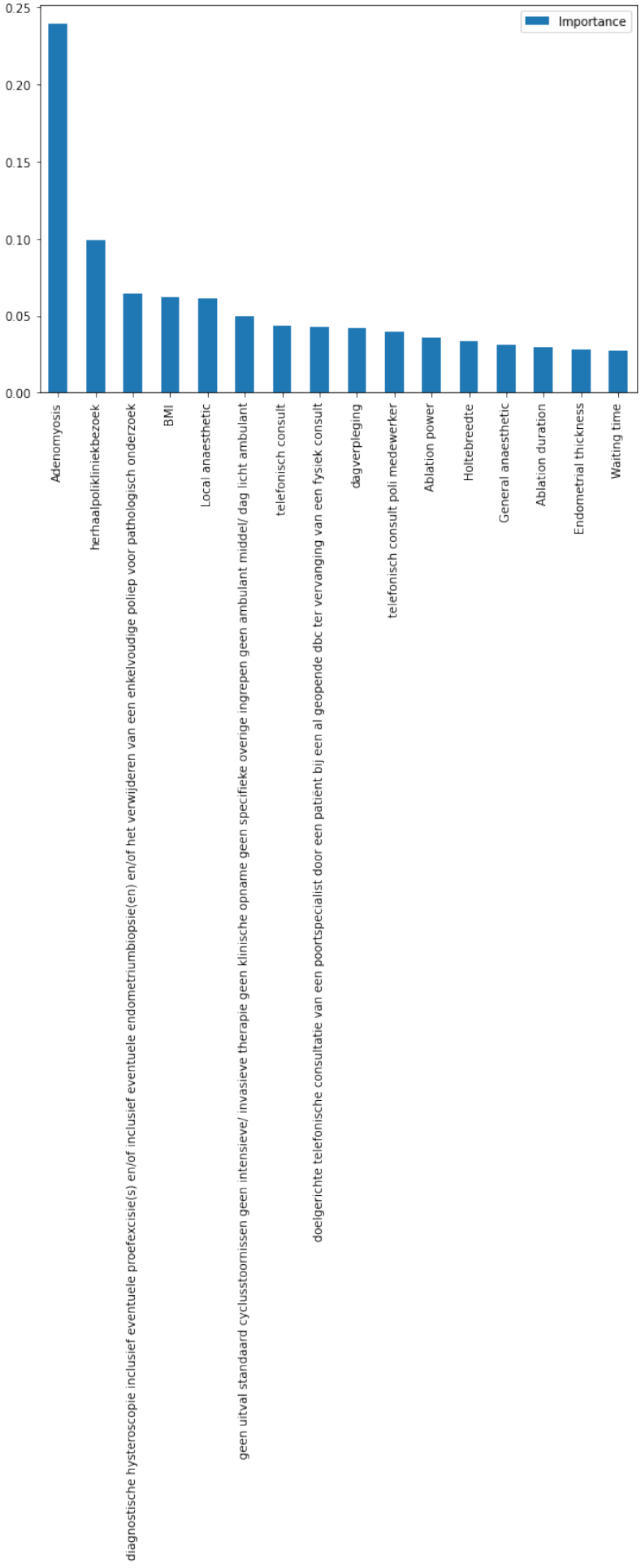


FIGURE D.10: Feature importance for extreme gradient boosting on original dataset - patient characteristics

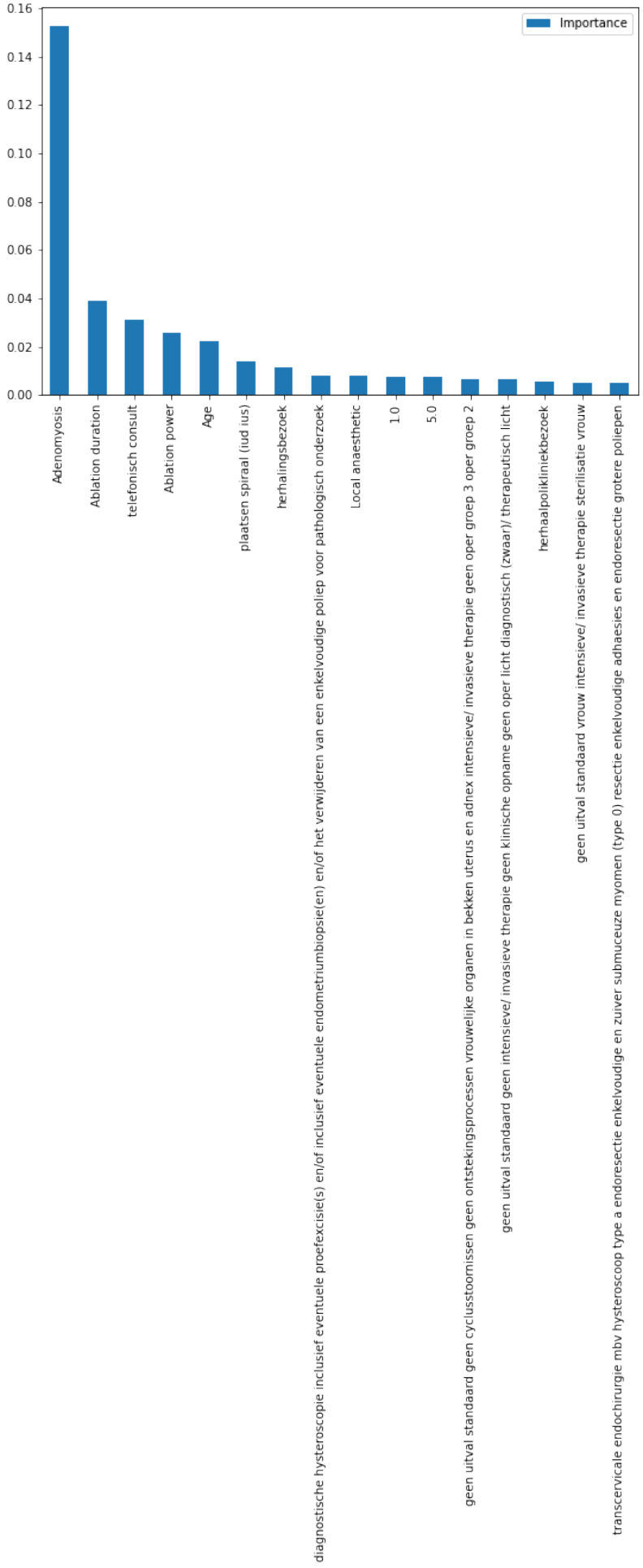


FIGURE D.11: Feature importance for neural network on original dataset - patient characteristics

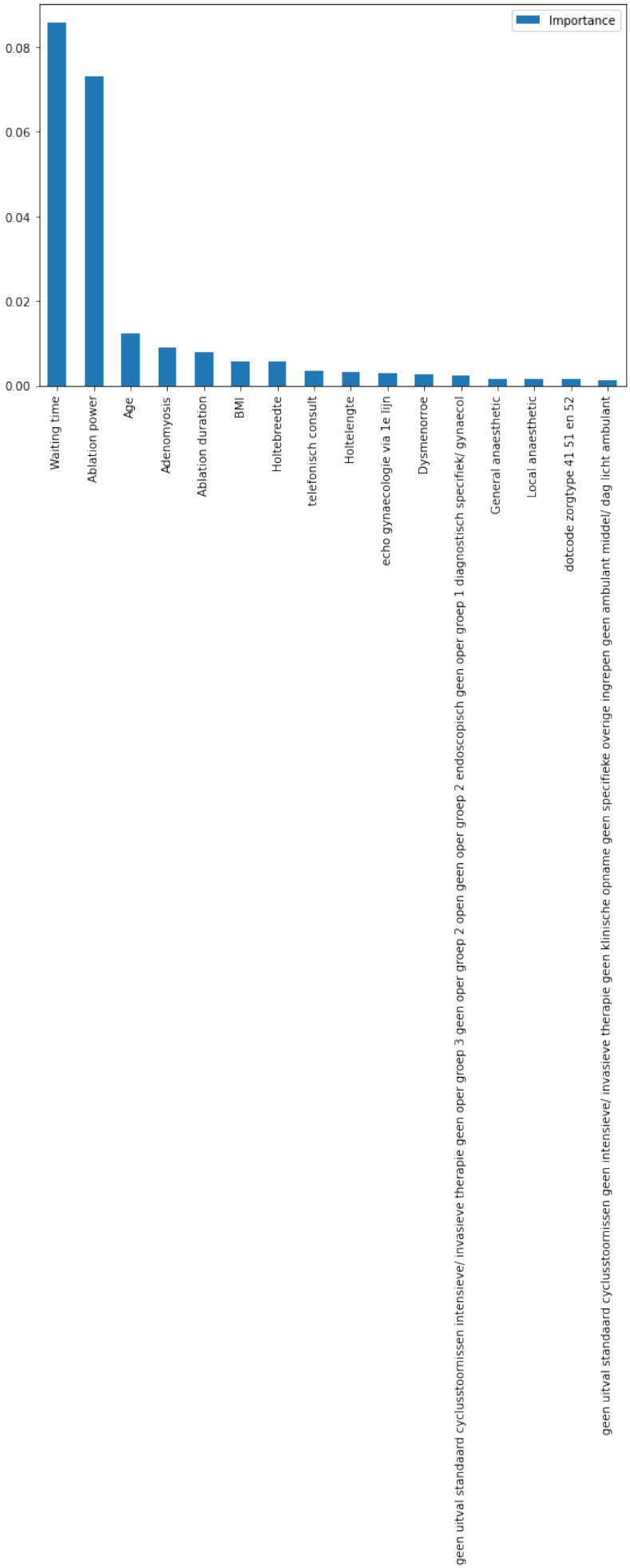


FIGURE D.12: Feature importance for k-nearest neighbour on original

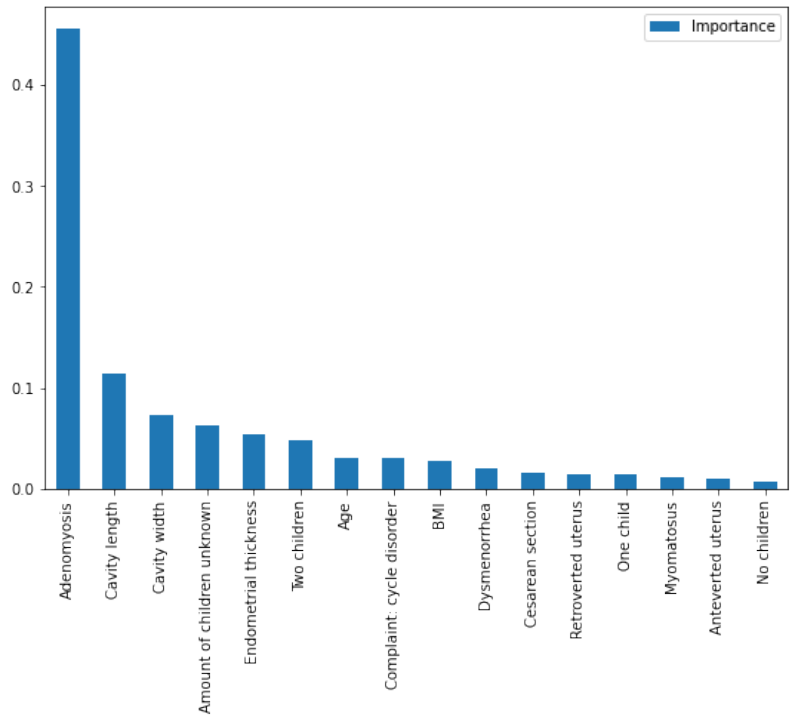


FIGURE D.13: Feature importance for decision tree on sampled dataset - patient characteristics

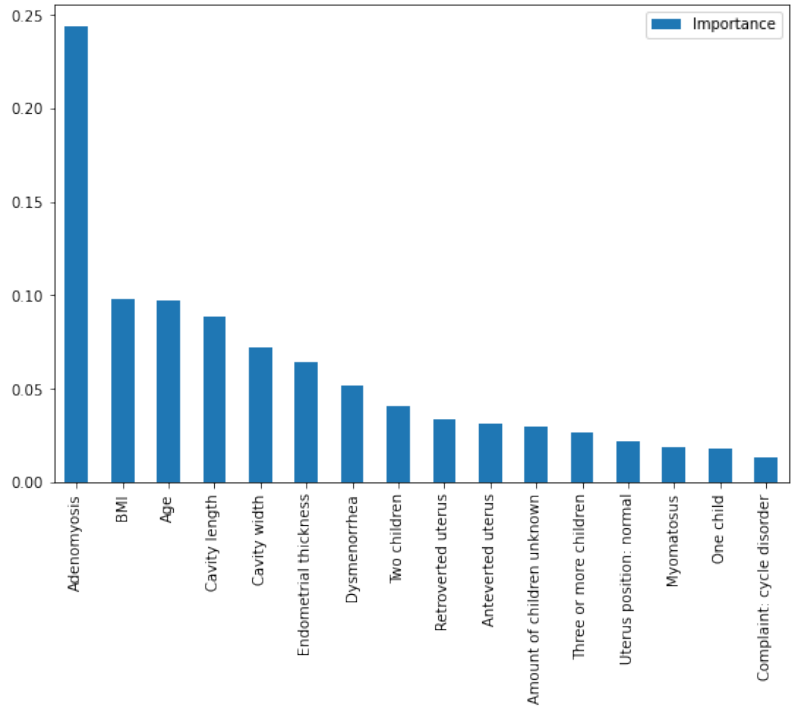


FIGURE D.14: Feature importance for random forest on sampled dataset - patient characteristics

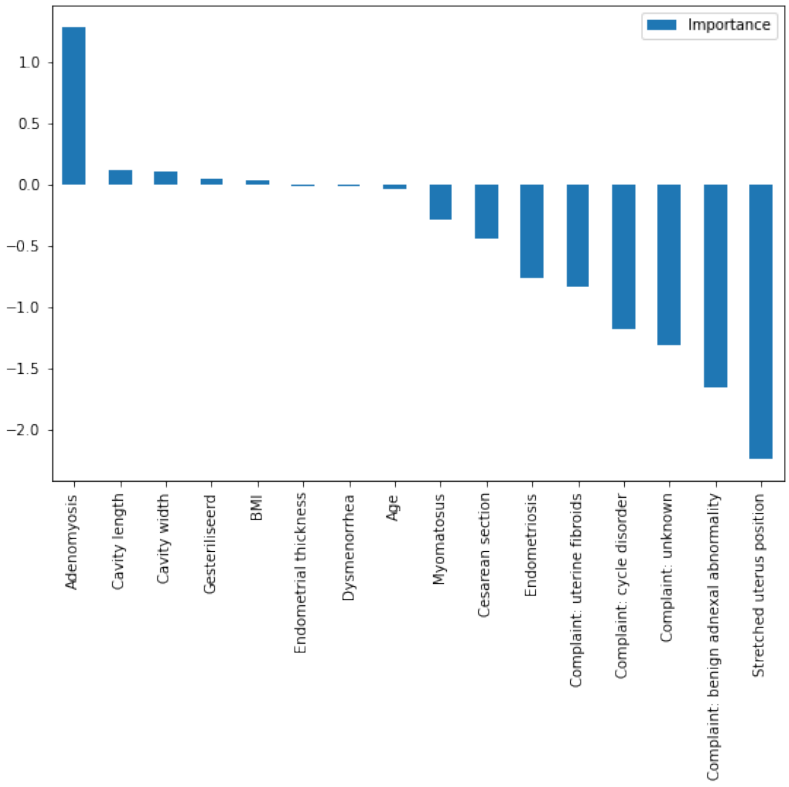


FIGURE D.15: Feature importance for logistic regression on sampled dataset - patient characteristics

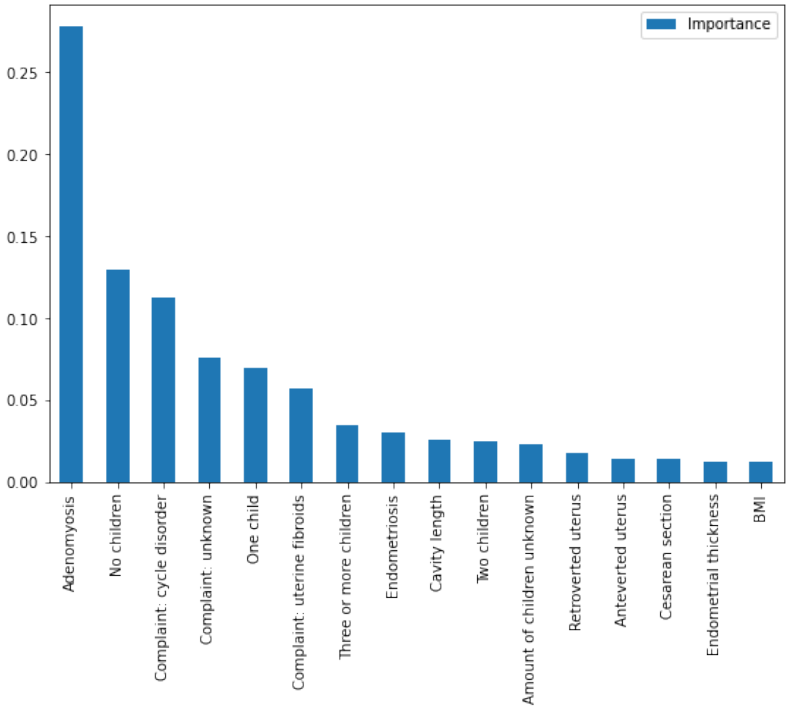


FIGURE D.16: Feature importance for extreme gradient boosting on sampled dataset - patient characteristics

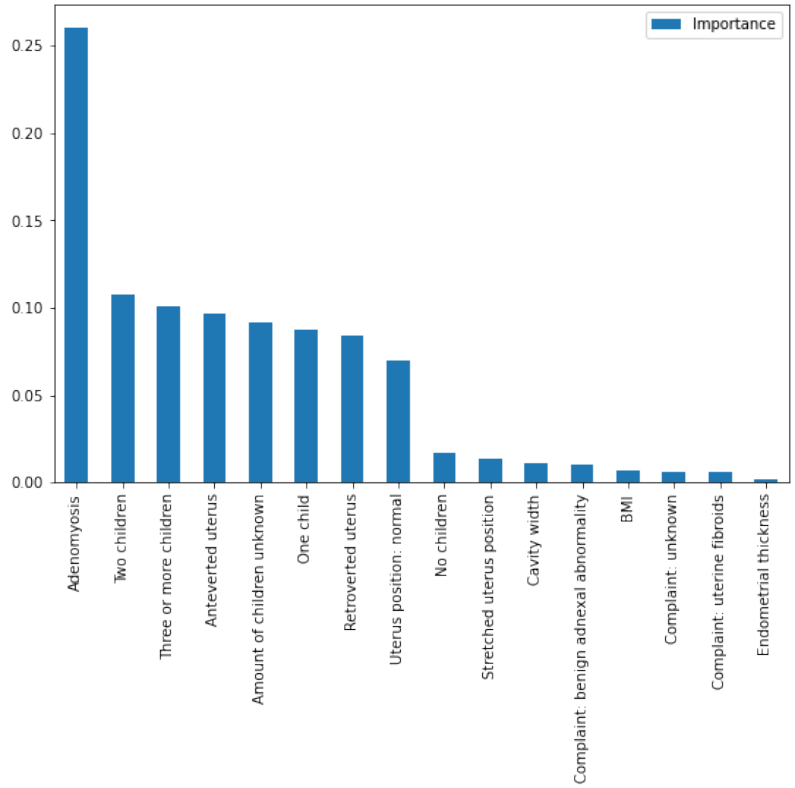


FIGURE D.17: Feature importance for neural network on sampled dataset - patient characteristics

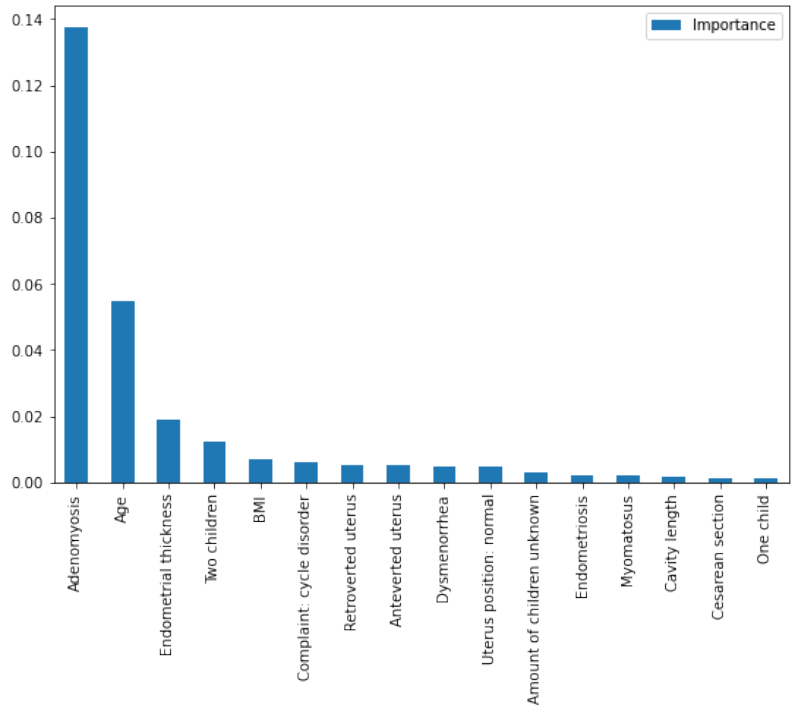


FIGURE D.18: Feature importance for k-nearest neighbour on sampled dataset - patient characteristics

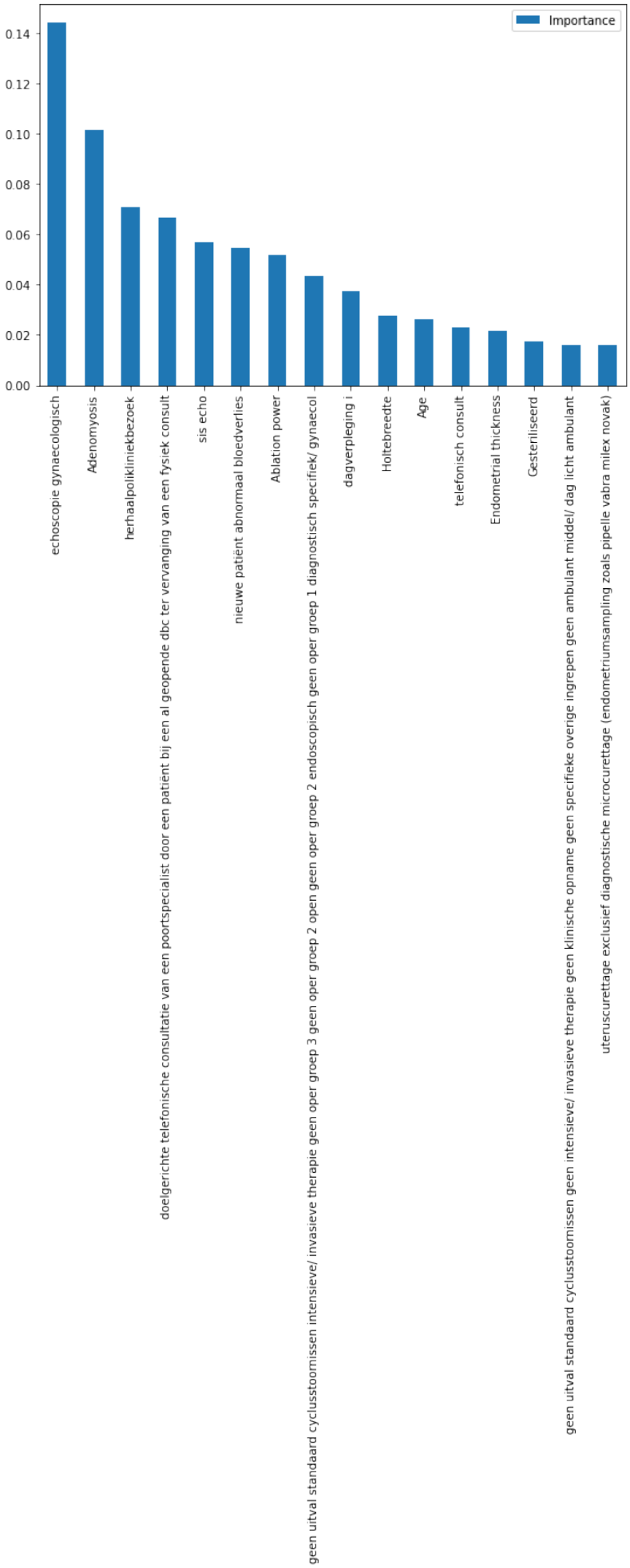
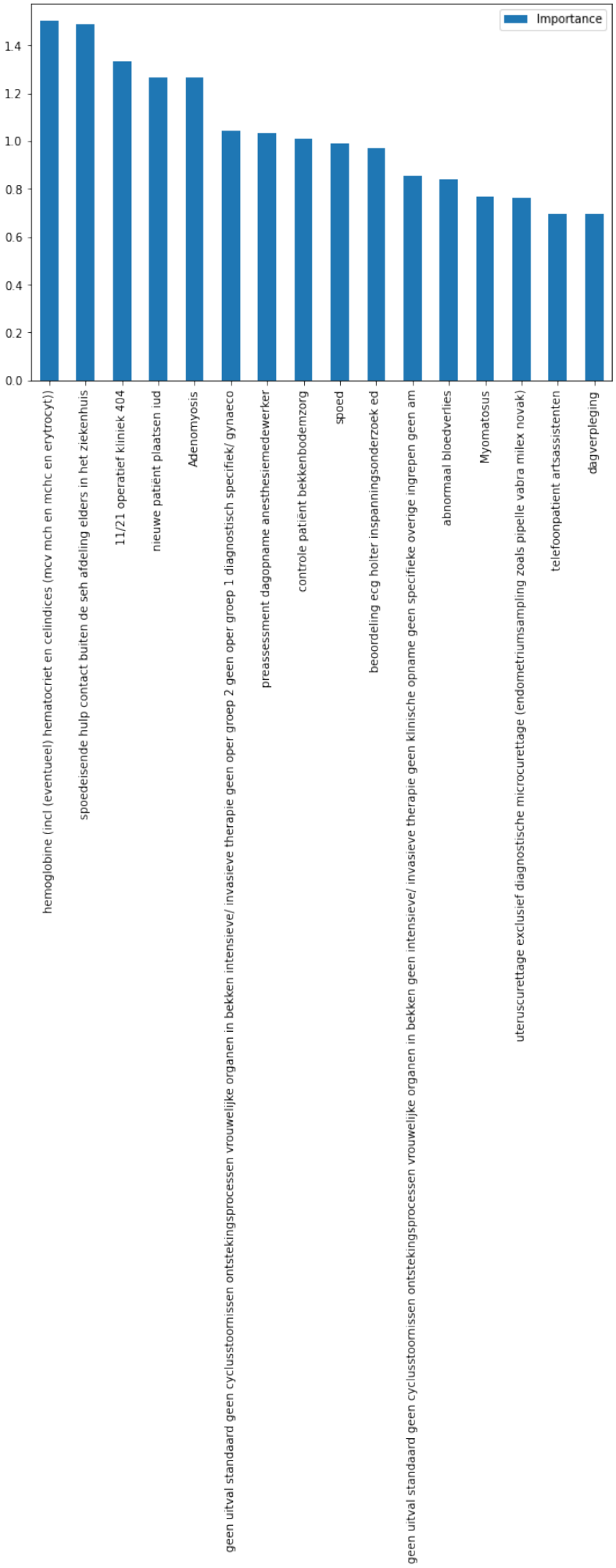


FIGURE D.19: Feature importance for decision tree on sampled

FIGURE D.20: Feature importance for random forest on sampled dataset - patient characteristics


Importance
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
hemoglobine (incl (eventueel) hematocriet en celindices (mcv mch en mchc en erytrocyt))
spoedeisende hulp contact buiten de seh afdeling elders in het ziekenhuis
11/21 operatief kliniek 404
nieuwe patient plaatsen iud
Adenomyosis
geen uitval standaard geen cyclusstoornissen ontstekingsprocessen vrouwelijke organen in bekken intensieve/ invasieve therapie geen oper groep 2 geen oper groep 1 diagnostisch specifiek/ gynaeco
preassessment dagopname anesthesiemedewerker
controle patiënt bekkenbodemzorg
spoed
beoordeling ecg holter inspanningsonderzoek ed
geen uitval standaard geen cyclusstoornissen ontstekingsprocessen vrouwelijke organen in bekken geen intensieve/ invasieve therapie geen klinische opname geen specifieke overige ingrepen geen am
abnormaal bloedverlies
Myomatosus
uterusscurettage exclusief diagnostische microcurettage (endometriumsampling zoals pipelle vabra milex novak)
telefoonpatient artsassistenten
dagverpleging

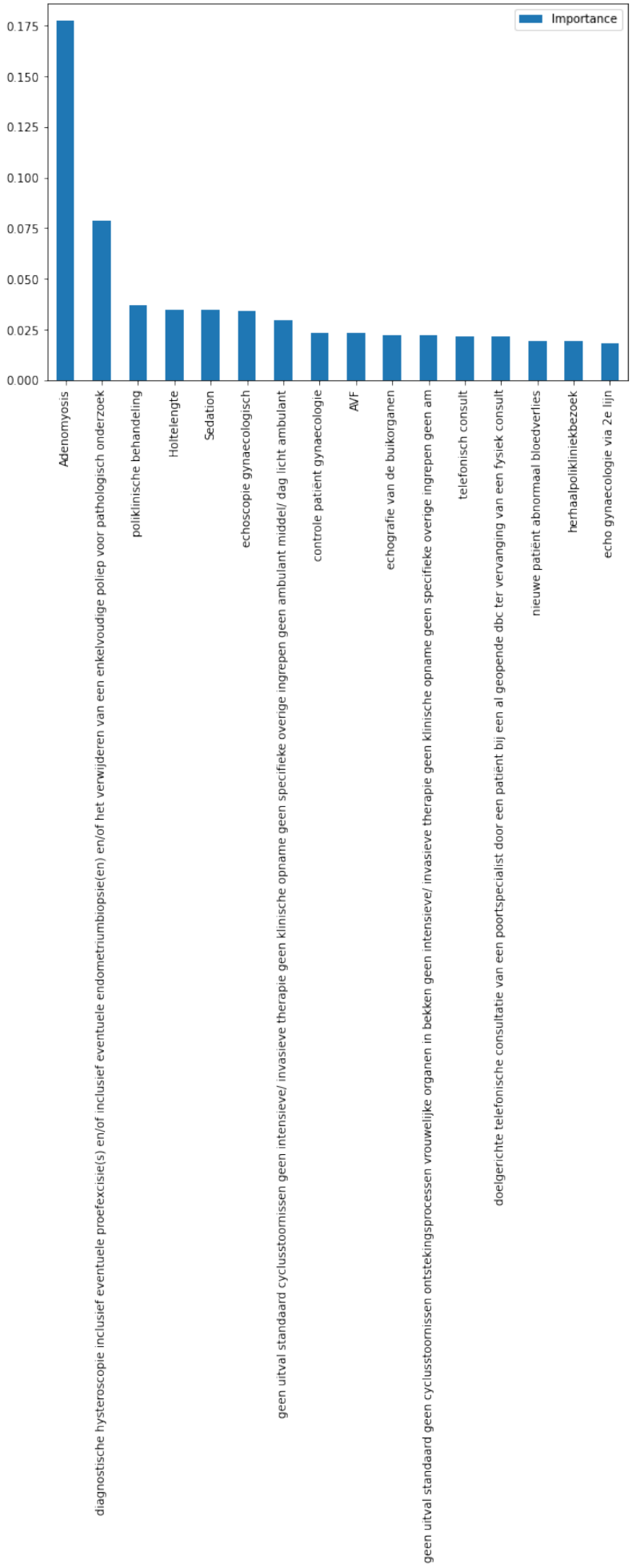


FIGURE D.22: Feature importance for extreme gradient boosting on

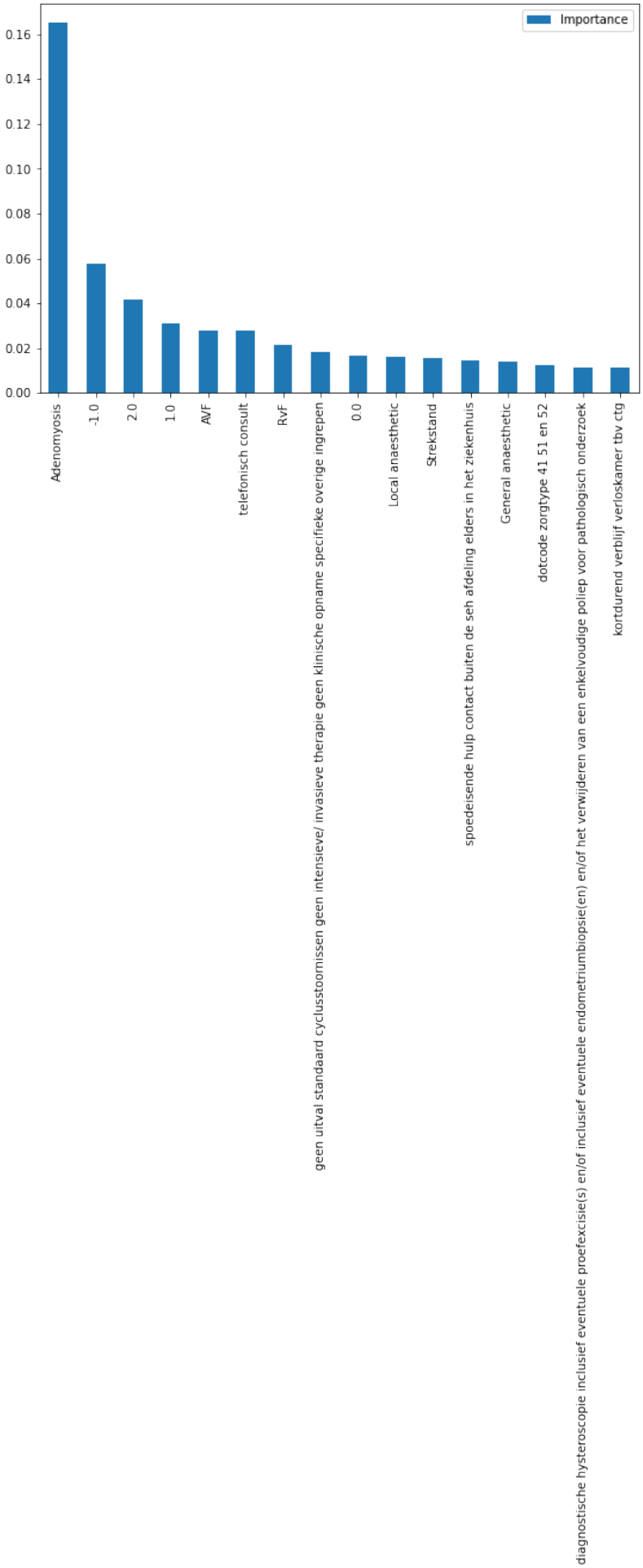


FIGURE D.23: Feature importance for neural network on sampled dataset - patient characteristics

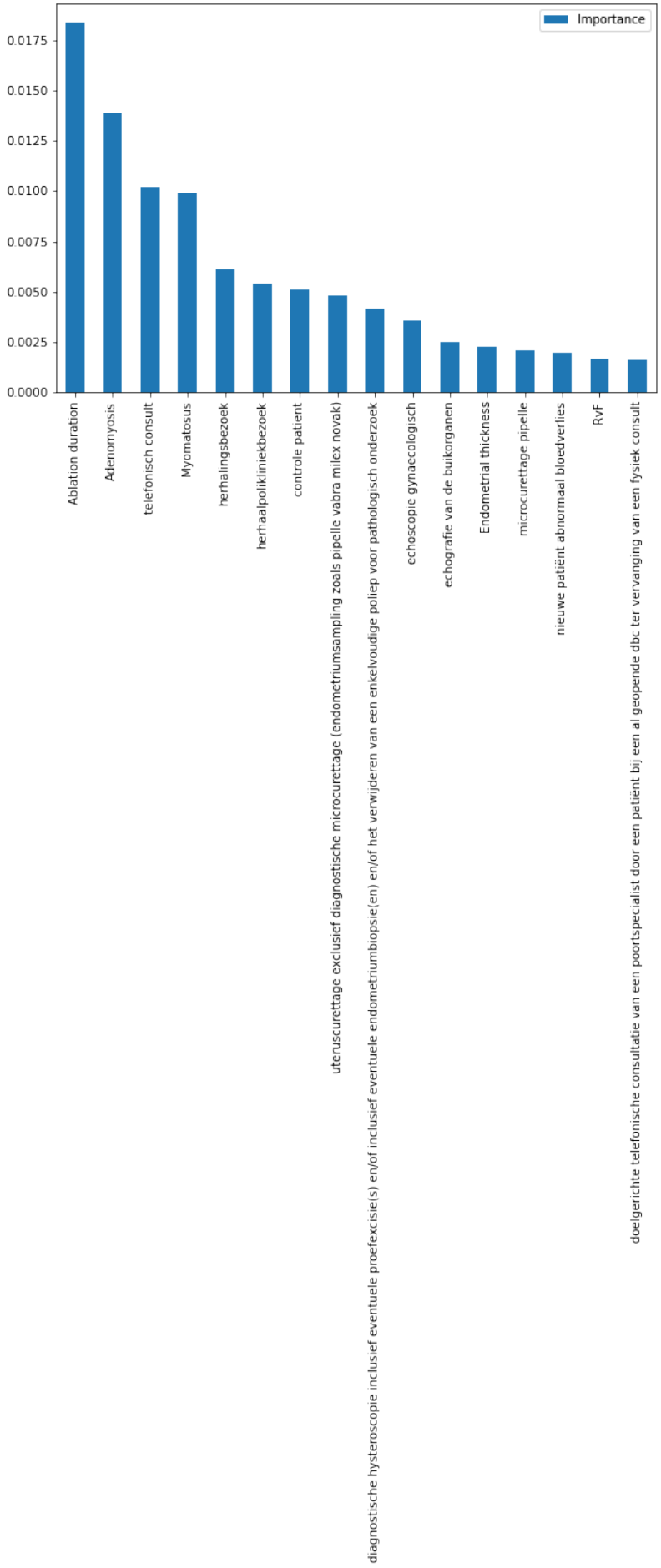


FIGURE D.24: Feature importance for k-nearest neighbour on sampled dataset - patient characteristics

Appendix E

F1 score results

Optimal hyperparameters for neural network on original dataset - patient characteristics:

- Hidden layer sizes: 10
- Batch sizes: 1
- Max iterations: 700

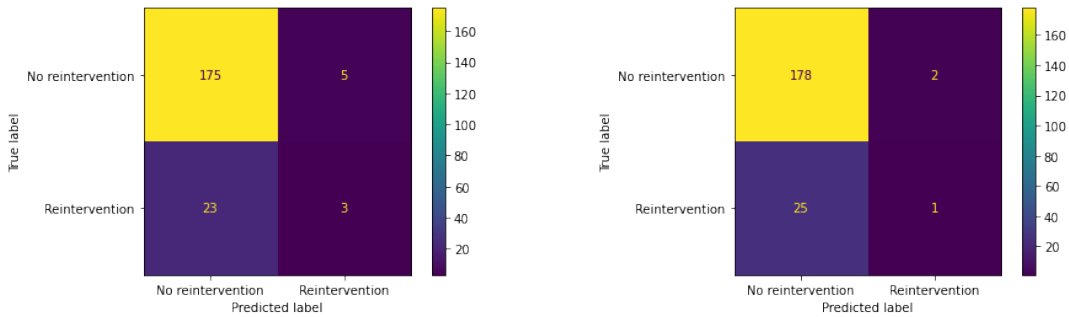
TABLE E.1: Algorithm performance for neural network on original dataset - patient characteristics

	Using default parameters		**Using tuned parameters**	
accuracy		0.86		0.87
AUC		0.77		0.77
	0	1	0	1
F1 score	0.93	0.18	0.97	0.07
precision	0.88	0.38	0.88	0.33
recall	0.97	0.12	0.99	0.04

Optimal hyperparameter for k-nearest neighbour on original dataset - patient characteristics:

- #neighbours: 2

Optimal hyperparameters for decision tree on original dataset - patient characteristics and process features:



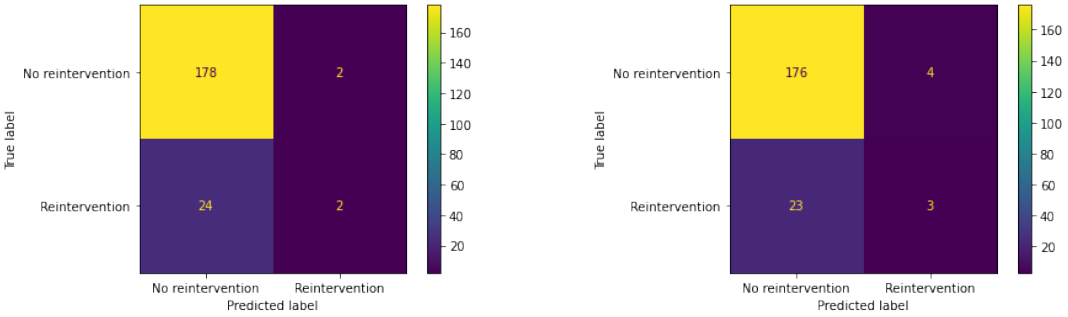
(A) Default hyperparameters

(B) Tuned hyperparameters

FIGURE E.1: Confusion matrices for neural network on original dataset - patient characteristics

TABLE E.2: Algorithm performance for k-nearest neighbour on original dataset - patient characteristics

Using default parameters			**Using tuned parameters**	
accuracy		0.87		0.87
AUC		0.65		0.57
	0	1	0	1
F1 score	0.93	0.13	0.93	0.18
precision	0.88	0.50	0.88	0.43
recall	0.99	0.08	0.98	0.12



(A) Default hyperparameters (B) Tuned hyperparameters

FIGURE E.2: Confusion matrices for k-nearest neighbour on original dataset - patient characteristics

- Max features: $\sqrt{n}$
- Max tree depth: 25
- Min samples split: 10
- Min samples leaf: 1

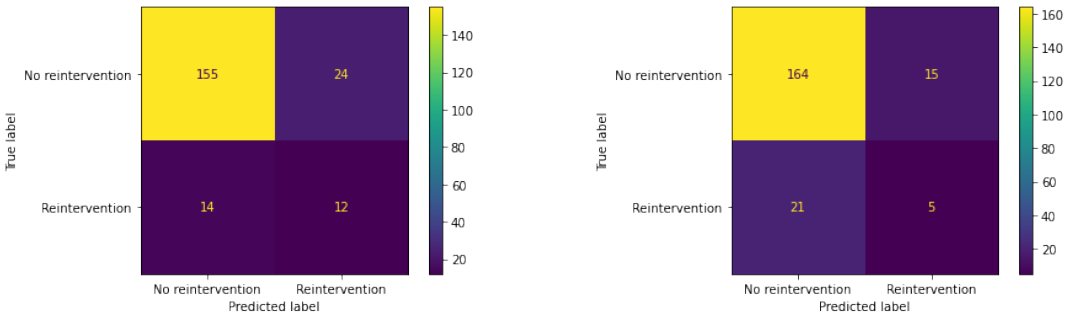
TABLE E.3: Algorithm performance for decision tree on original dataset - patient characteristics and process features

Using default parameters			**Using tuned parameters**	
accuracy		0.81		0.82
AUC		0.66		0.58
	0	1	0	1
F1 score	0.89	0.39	0.90	0.22
precision	0.92	0.33	0.89	0.25
recall	0.87	0.46	0.92	0.19

Optimal hyperparameters for random forest on original dataset - patient characteristics and process features:

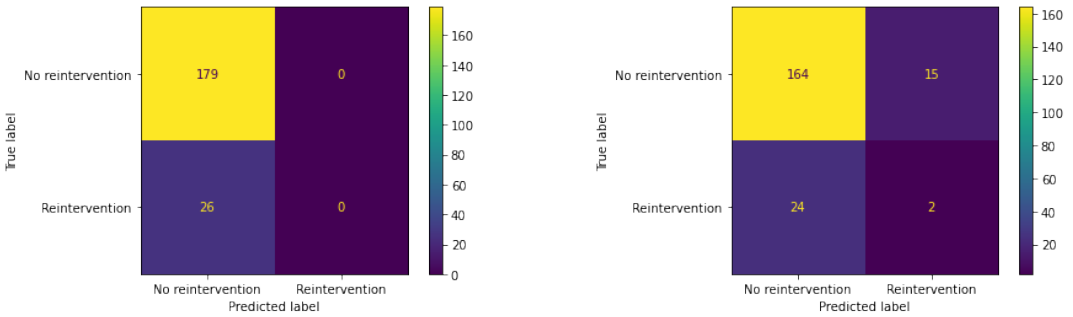
- Max tree depth: None
- #estimators: 1

Optimal hyperparameter for k-nearest neighbour on original dataset - patient characteristics and process features



(A) Default hyperparameters (B) Tuned hyperparameters

FIGURE E.3: Confusion matrices for decision tree on original dataset - patient characteristics and process features

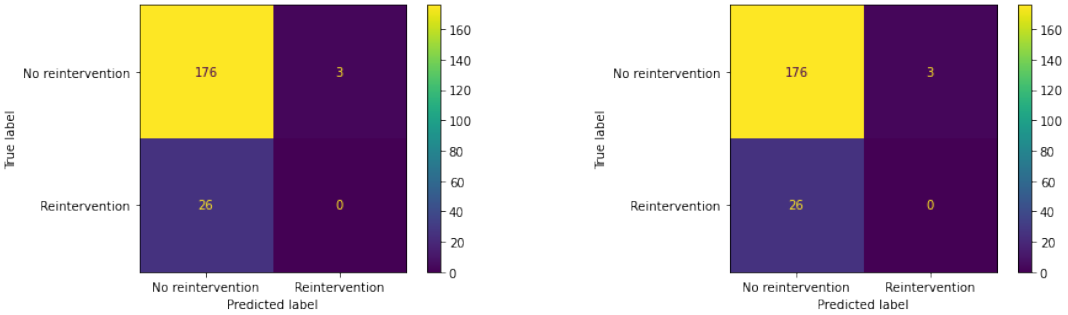


(A) Default hyperparameters (B) Tuned hyperparameters

FIGURE E.4: Confusion matrices for random forest on original dataset - patient characteristics and process features

TABLE E.4: Algorithm performance for random forest on original dataset - patient characteristics and process features

	Using default parameters		**Using tuned parameters**	
accuracy		0.87		0.81
AUC		0.80		0.50
	0	1	0	1
F1 score	0.93	0.00	0.89	0.09
precision	0.87	0.00	0.87	0.12
recall	1.00	0.00	0.92	0.08



(A) Default hyperparameters (B) Tuned hyperparameters

FIGURE E.5: Confusion matrices for k-nearest neighbour on original dataset - patient characteristics and process features

- #neighbours: 5

TABLE E.5: Algorithm performance for k-nearest neighbour on original dataset - patient characteristics and process features

	Using default parameters		**Using tuned parameters**	
accuracy		0.86		0.86
AUC		0.58		0.58
	0	1	0	1
F1 score	0.92	0.00	0.92	0.00
precision	0.87	0.00	0.87	0.00
recall	0.98	0.00	1.00	0.00