Machine learning-based prediction model for in vitro fertilization success

Predicting the likelihood of a live birth in subsequent in vitro fertilization cycles following an unsuccessful first cycle

Master's thesis in Biomedical Engineering

ELVIRA CARLÉN
ELSA THORESTRÖM

DEPARTMENT OF ELECTRICAL ENGINEERING

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2025
www.chalmers.se

Master's thesis 2025

Machine learning-based prediction model for in vitro fertilization success

Predicting the likelihood of a live birth in subsequent in vitro fertilization cycles following an unsuccessful first cycle

ELVIRA CARLÉN
ELSA THORESTRÖM

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Electrical Engineering
Signal Processing and Biomedical Engineering
Chalmers University of Technology
Gothenburg, Sweden 2025

Machine learning-based prediction model for in vitro fertilization success
Predicting the likelihood of a live birth in subsequent in vitro fertilization cycles following an unsuccessful first cycle
ELVIRA CARLÉN
ELSA THORESTRÖM

Supervisor: Anton Matsson, Department of Computer Science and Engineering
Co-supervisor: Dimitra Lappa, AI Competence Center, Sahlgrenska University Hospital
Examiner: Ida Häggström, Department of Electrical Engineering

Master's Thesis 2025
Department of Electrical Engineering
Signal Processing and Biomedical Engineering
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: A woman holding a newborn (AI-generated image with Microsoft Copilot, 2025).

Typeset in LaTeX
Printed by Chalmers Reproservice
Gothenburg, Sweden 2025

Machine learning-based prediction model for in vitro fertilization success
Predicting the likelihood of a live birth in subsequent in vitro fertilization cycles following an unsuccessful first cycle
ELVIRA CARLÉN
ELSA THORESTRÖM
Department of Electrical Engineering
Chalmers University of Technology

Abstract

Patients undergoing in vitro fertilization (IVF) are often informed about their average chances of achieving a live birth, which is mainly influenced by the woman's age. This project is part of a larger study and aims to predict individualized chances of success using machine learning (ML), helping patients set realistic expectations and guide their decisions. Three research questions were addressed, each focused on predicting the likelihood of a live birth in one of the following scenarios: the second cycle following an unsuccessful first cycle; the second or third cycle following an unsuccessful first cycle; and the third cycle following an unsuccessful first and second cycle. A complete IVF cycle includes ovarian stimulation, oocyte retrieval, fertilization, embryo culture and all resulting fresh and frozen embryo transfers. The dataset used to train the ML models was collected from Sahlgrenska University Hospital between January 2016 and December 2022 and included 21 features from 3,217 patients undergoing 5,170 cycles. Three ML models were evaluated: logistic regression, random forest (RF) and support vector machines (SVM). The predictive performance of logistic regression and RF was consistent with results from similar studies that used datasets comprising more than 70,000 IVF cycles. This study suggests that comparable results can be obtained with a smaller dataset and with fewer adjustments, such as accounting for non-linear relationships between the features and the live birth outcome in logistic regression, which RF handles naturally.

Keywords: IVF, live birth prediction, machine learning, logistic regression, support vector machines and random forest.

Acknowledgements

We would like to express our deepest gratitude to our supervisors, Dimitra Lappa from the AI Competence Center at Sahlgrenska University Hospital and Anton Matsson from Chalmers University of Technology. We are truly thankful for their guidance, patience and encouragement throughout the course of this project. Their support has not only helped shape the direction of our work but also given us the confidence to take on similar projects in the future.

We are also very grateful for the support given by Kersti Lundin and Zoha Saket, clinicians at the Department of Reproductive Medicine at Sahlgrenska University Hospital. Their expert medical knowledge and significant work with the patient data has been essential to making this project possible.

Special thanks to Ida Häggström, the examiner of this project from Chalmers University of Technology, for her guidance on conducting a master's thesis and writing the thesis report.

Elvira Carlén & Elsa Thoreström, Gothenburg, May 2025

List of Acronyms

Below is an alphabetical list of the acronyms used throughout this thesis:

AFC antral follicle count
AMH anti-mullerian hormone
AUROC area under the receiver operating characteristic
BMI body mass index
CLBR cumulative live birth rate
DET double embryo transfer
DT decision tree
ET embryo transfer
FET frozen embryo transfer
GQE good quality embryo
ICSI intracytoplasmic sperm injection
IVF in vitro fertilization
MCC matthews correlation coefficient
ML machine learning
RCS restricted cubic splines
RF random forest
SET single embryo transfer
SHAP shapley additive explanations
SVM support vector machines

Contents

List of Figures

List of Tables

1

Introduction

Fertility problems affect 10 to 15 percent of all heterosexual couples worldwide, with causes linked to both women and men [1]. Annually, over 25,000 in vitro fertilization (IVF) treatments are conducted in Sweden; about half in the public sector and half in the private. In the public sector, each infertile couple is typically offered up to three full IVF cycles or until the birth of a child, whichever occurs first. In Sweden, about two-thirds of all patients who start an infertility investigation and treatment receive a live birth [1]. Before starting the IVF process, patients are usually informed about the average chance of pregnancy, which is primarily influenced by the woman's age and the underlying cause of infertility [3]. However, this does not reflect the uncertainty in individual cycle outcomes, especially after a first IVF cycle that does not result in a live birth. There is a clear need for more personalized predictions to help guide decisions for patients considering undergoing additional IVF.

The goal of this project was to develop an individualized prediction model for the chance of a live birth during the second or third IVF cycle, given a failed first cycle. This chapter begins with an overview of the IVF procedures, followed by a presentation of related work, project objectives, research inventiveness, limitations and the societal and ethical aspects of the project.

1.1 In vitro fertilization

An IVF cycle consists of several steps, which can be seen in figure 1.1. In this section, these steps will be described.

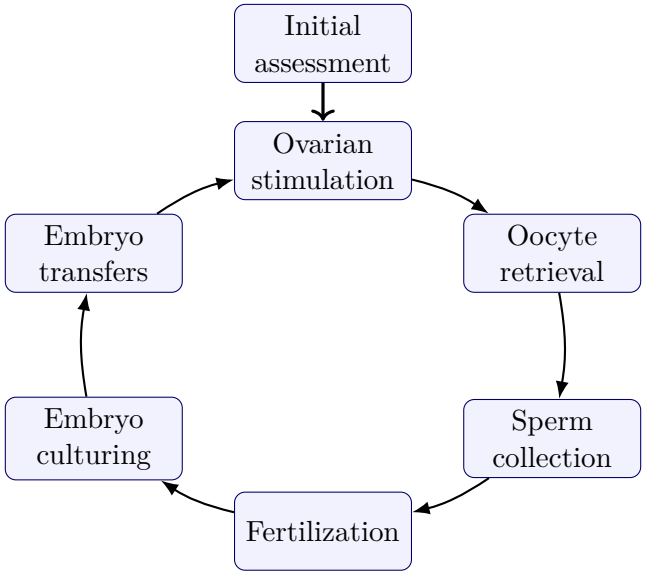


Figure 1.1: A complete cycle of IVF refers to the process of beginning with an initial assessment, where clinical parameters and fertility status are evaluated. This is followed by ovarian stimulation, oocyte retrieval, sperm collection, fertilization, embryo culturing and all ETs (fresh or frozen) resulting from that particular cycle of ovarian stimulation.

The initial phase of an IVF treatment involves an assessment of the individual patient's clinical profile. Key factors considered during this evaluation include the patients' age (both the woman's and potential partner's), body mass index (BMI), duration of infertility and potential underlying causes for infertility. As shown in figure 1.2, it is estimated that approximately one-third of infertility cases are attributable to male-related causes, another third to female-related causes and one-third to factors present in both partners [1]. Some patients are categorized as "unexplained infertility", where no definitive cause can be identified. Two of the most common causes of infertility in women are problems with ovulation which is often linked to polycystic ovary syndrome (PCOS) and damage to the fallopian tubes caused by endometriosis [4]. Female fertility declines drastically from the age of 35 [1]. In men, the primary contributing factors to infertility include reduced sperm quality and decreased sperm count.

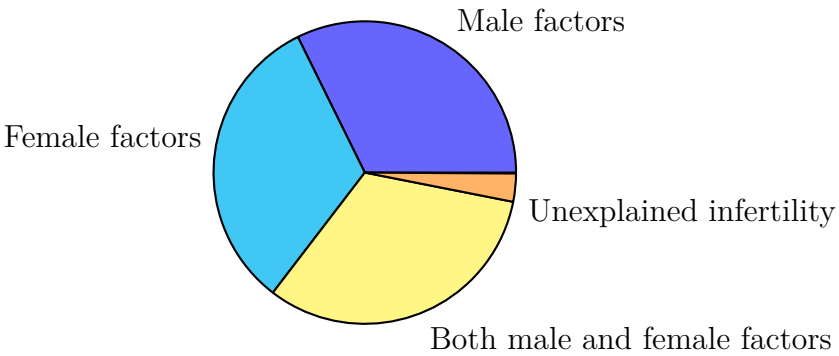


Figure 1.2: Infertility causes: approximately one-third are due to male factors, one-third to female factors and one-third to both [1]. In some cases, the infertility is unexplained.

The patient's initial assessment also include a measurement of the ovarian reserve, which refers to the number of immature follicles in the ovaries [5]. A follicle is a small, fluid-filled sac in the ovary that contains an immature oocyte (the female egg). When an oocyte matures, the follicle ruptures and releases the oocyte from the ovary into the fallopian tubes. The ovarian reserve is an indication of the number of oocytes that may be retrieved during an IVF stimulation. Ovarian reserve is commonly measured through two methods: anti-mullerian hormone (AMH), a hormone produced by the follicles which provides an indirect measure of the number of follicles; and antral follicle count (AFC) which is the number of antral follicles visible via ultrasound. Antral follicles are immature follicles that begin growing in response to hormonal signals at the start of each menstrual cycle. Both AMH and AFC decrease naturally with age as hormone levels go down. Together, these methods provide insights into a woman's fertility status.

Based on the initial assessment, a personalized treatment plan is formed which begins by selecting either so called "standard IVF" or intracytoplasmic sperm injection (ICSI) as fertilization method [1]. Standard IVF refers to the process in which sperm and oocytes are placed together in a nutrient solution, allowing the sperm to fertilize the oocytes "naturally". In ICSI, a single sperm is directly injected into the oocyte with a thin needle. ICSI is often used when the sperm has poor mobility or when a previous IVF cycle showed no fertilization or very low fertilization rates.

The process of designing the personalized treatment plan also includes determining how the ovaries will be stimulated to produce mature oocytes [6]. The stimulation is done using hormone treatment, which can follow either a long or short protocol. In the long protocol, often referred to as the agonist protocol, the process begins with down-regulation which suppresses the hormones normally released during the menstrual cycle to prevent ovulation. To stimulate the ovaries to produce multiple mature oocytes, the hormone gonadotropins are given. The entire agonist protocol typically lasts four to five weeks with the down-regulation step taking about two weeks. In contrast, the short protocol known as the antagonist protocol lasts around two weeks and skips the down-regulation step. Today, it is more common to use the short protocol, or the antagonist protocol. The reasons are its ease of use, as well as fewer side effects.

Once the patient has completed the hormone treatment for ovarian stimulation, an oocyte retrieval is performed. This is also called oocyte aspiration. On the same day as the aspiration, the male partner provides a sperm sample (unless donor sperm is being used). The sperms with the highest quality are selected for fertilization which can be performed using either standard IVF or ICSI. In Sweden, standard IVF is used for approximately half of all fresh cycles while ICSI is used for the other half [1]. If multiple embryos of good morphology develop, one or rarely two can be transferred to the woman's uterus, a process called embryo transfer (ET). The remaining embryos can be frozen for future transfers. For fresh ET, embryos are transferred two to three days after fertilization (early/cleavage stage transfer) or after five days, after which the embryo is called a blastocyst, hence a blastocyst transfer. To minimize the risk of multiple births, common practice is that only one embryo is transferred, called single embryo transfer (SET). However, for

early ET, sometimes two embryos can be transferred which is called double embryo transfer (DET), mainly when couples have poorer fertility prognosis or when previous treatments have failed. In such cases, the risk of twin pregnancy is generally considered low.

One of the largest reproductive medicine units in Sweden is the Department of Reproductive Medicine at Sahlgrenska University Hospital. At the time of the data collection for this project, the clinic initiated approximately 1,100 IVF treatments annually using the patients' own oocytes and sperm [7]. In addition, around 700 transfers with frozen embryos were performed, a process called frozen embryo transfer (FET).

In Sweden, infertility investigations for heterosexual couples in the public sector are initiated after at least one year of unsuccessful attempts to conceive [8]. To be eligible for IVF treatment at Sahlgrenska, the couple must have been in a stable, living-together relationship for at least two years. The woman must be younger than 40 years old before the start of the IVF treatment and have a BMI between 18 and 35. The partner must be younger than 56 years old. Single women and same-sexed couples are also granted infertility treatment under the same conditions. In Sweden, public health care includes up to three cycles of IVF treatment for each couple. However, the woman is only eligible to continue IVF until achieving a first live birth. This means that if she experiences a live birth during an IVF cycle, any further FETs must be paid for privately. A failed cycle is defined as the absence of a live birth following a complete IVF cycle, including all ETs (both fresh and frozen-thawed).

Frozen embryos are allowed to be stored for up to ten years in Sweden [1]. If a fresh ET does not result in a live birth, the frozen embryos can then be thawed and used, a process known as FET. For FET, embryos are typically cultured to day five or six before being frozen for storage and later transfer. It is even more common to perform SET for FET than for fresh ET. For some patients, all embryos are frozen without doing a fresh ET which can happen when an increased risk of ovarian hyperstimulation syndrome (OHSS) is observed, a common complication of IVF. It is also possible to freeze unfertilized oocytes, which allows women to preserve their fertility for future family planning. This is useful for patients who, for example, are about to undergo chemotherapy, as there is a risk of damaging the eggs during the cancer treatment.

1.2 Related work

Given the high emotional strain associated with IVF, there is a clear need for reliable prediction tools. These tools can provide patients with individualized prognoses, helping couples prepare and set realistic expectations. Over the years, several studies have sought to develop predictive models for IVF outcomes. Recent research shows a shift toward more advanced methods, with machine learning (ML) demonstrating strong potential to improve prediction accuracy. Table 1.1 provides an overview of studies that have predicted IVF outcomes in various ways, highlighting

the authors, the size of the dataset, the country of collection and the results of the best-performing ML models.

An example of an IVF study is the prediction model developed by McLernon et al. (2016), which estimated individualized probabilities of achieving a live birth over six IVF cycles[3]. The study utilized data from over 250,000 women in the United Kingdom, covering up to six IVF cycles per patient between 1999 and 2008. The analysis excluded individuals who used donated sperm or oocytes. However, the model did not incorporate measures of ovarian reserve which may be a relevant predictor of treatment success.

In 2022, McLernon et al. published a new prediction model with data from the United States which included factors like BMI and ovarian reserve markers such as AMH [9]. The population consisted of women who started their first autologous IVF cycle during 2014-2015. Women who underwent IVF treatment with oocyte or sperm donations were excluded from the study. Instead of predicting over six cycles, they predicted the pre-treatment cumulative probability of live birth over three cycles of IVF. They also created a post-treatment model with updated probabilities given a failed first cycle. Using a discrete-time logistic regression model, the post-treatment model predicted the cumulative probability of a live birth in the second and third IVF cycles, following a first unsuccessful cycle.

Another large-scale prediction model, published in 2022 by Ratna et al., used data from treatments between 1999 and 2008, but did not incorporate ovarian reserve measures such as AFC [10]. In the study, the cumulative probability of a live birth for a couple starting their second complete cycle of IVF was evaluated, regardless of whether the first cycle resulted in a live birth. A multivariate discrete-time logistic regression model was used in the study.

In the study by Barnett-Itzhaki et al. (2020), the ML algorithms support vector machines (SVM) and neural networks were shown to have an advantage over logistic regression models in predicting IVF outcomes [11]. The finding was based on a small cohort with strict inclusion criteria to ensure a homogeneous study group for comparing analysis methods. In the study, the likelihood of achieving a live birth was predicted following one IVF cycle. The study suggested that ML algorithms, such as SVM and neural networks, are better at identifying complex and non-linear relationships that may not be captured by traditional approaches like logistic regression. A larger cohort study conducted in Belgium, involving 1,052 women who underwent fresh ETs of blastocyst-stage embryos in their first IVF cycle, also supported the superior accuracy of ML algorithms in predicting the result of ongoing IVF cycles [12]. Specifically, a random forest (RF) model outperformed multivariate logistic regression models in terms of prediction accuracy.

In another study, various models, including logistic regression, naive bayesian, k-nearest neighbor, SVM, decision tree (DT) and ensemble learning models, such as RF, XGBoost, LightGBM and Voting, were compared to predict clinical pregnancy outcomes during IVF [13]. The study, which included 615 patients between 2015 and 2021, found that ensemble models outperformed the other models. In the study, the probability of achieving a live birth following a single fresh blastocyst transfer

was predicted.

Furthermore, one research project compared five different ML methods for predicting IVF success following only fresh ETs. It was found that AdaBoost in combination with feature selecting using genetic algorithm achieved the highest accuracy [14]. Feature selection using genetic algorithm was utilized because of its ability to evolve the selected set in an iterative process and to capture complex relationships between variables. By applying feature selection, the dimension of the dataset was deducted, enabling an improvement in performance and more efficient data mining. The authors found that genetic algorithm improved the performance for all models in terms of all performance metrics.

Table 1.1: Summary of existing ML models for predicting IVF outcomes, including authors, dataset sizes, countries of data collection and results of the best-performing models. In the results column, C-index refers to the concordance index while AUROC stands for the area under the receiver operating characteristic (ROC) curve. The three first studies, i.e. McLernon 2016, McLernon 2022 and Ratna 2022, predicted the cumulative probability of achieving a live birth over several IVF cycles, compared to the other studies whose objectives were more different from ours.

Author	Dataset	ML Models	Result
McLernon 2016	133,873 women with 184,269 complete cycles. (UK, 1999 -2008)	Discrete time logistic regression	C-index: 0.72 (0.71-0.73)
McLernon 2022	88,613 women with 121,561 complete cycles. (US, 2014-2016)	Discrete time logistic regression	C-index: 0.71
Ratna 2022	49,314 women with 73,053 complete cycles. (UK, 1999-2008) External validation: 39,442 women. (UK, 2010-2016)	Discrete time logistic regression	C-index: 0.65
Barnett-Itzhaki 2020	136 women. (Israel, January 2014-August 2016)	Neural networks	Accuracy = 0.874 AUROC = 0.77
Blank 2019	1052 women. (Belgium, July 1, 2015-August 3, 2017)	Random forest	AUROC = 0.74 ± 0.03
Liu 2024	615 women. (China, July 2015-June 2021)	Voting model	AUROC = 0.845
Dehghan 2024	812 women. (Iran)	AdaBoost with genetic algorithm	Accuracy = 0.899 AUROC = 0.910

1.3 Objective

This project aims to develop a model that provides individualized predictions of the likelihood of achieving a live birth across the second and third IVF cycle, following an unsuccessful first or second cycle. The term "individualized predictions" refer to a predicted probability for a unique set of feature values, as opposed to a set of feature values that apply to more than one patient. This project will address the following three research questions:

1. Is it possible to develop an ML algorithm to predict the individualized chance of live birth in a second IVF cycle following an unsuccessful first cycle, that reaches better performance than a logistic regression model?
2. Is it possible to develop an ML algorithm to predict the aggregated, individualized chance of live birth in a second and third IVF cycle following an unsuccessful first cycle, that reaches better performance than a logistic regression model?
3. Is it possible to develop an ML algorithm to predict the individualized chance of live birth in a third IVF cycle following an unsuccessful first and second cycle, that reaches better performance than a logistic regression model?

ML models will be developed and compared to logistic regression models, which will be trained on the Sahlgrenska dataset. Logistic regression was selected as a benchmark due to its widespread use in related research. The baseline models for this project will be two logistic regression models: one logistic regression trained on age as sole feature and another trained on the complete set of available features. The first model allows comparison with a well-established clinical predictor, while the later one enables comparison with findings from related studies.

1.4 Research inventiveness

This project relates to research that has already been published, such as the articles by McLernon et al. in 2016 [3] and 2022 [9], and Ratna et al. in 2022 [10]. However, there are aspects in this project that enables a novel perspective. One of the more important factors is an up-to-date dataset with patient data from 2016-2022. In comparison, both McLernon and Ratna used data from 1999-2009 [3], [10], whereas Ratna used data from 2010 to 2016 for external validation. Saket et al. found that the cumulative live birth rate (CLBR) per oocyte aspiration has increased from 27.0% to 36.3% in Sweden during 2007–2017 and suggested emerging trends in IVF care as a reason for this change [15]. Today, an increasing number of transfers are performed at the blastocyst stage with improved results per transfer. Evidently, IVF care has experienced an improvement in recent years which outdates the McLernon and Ratna studies.

Apart from previous research being outdated, it is also geographically indifferent. To the best of our knowledge, no similar studies have used patient data from Sweden. Instead, patient data from the United Kingdom and the United States have been

utilized [3], [9], [10]. As clinical practice can vary between countries, a geographically relevant dataset is required for a future applicable model. The overall IVF success rate can also be affected by demographic factors, such as public health and genetics.

This project will include patients and features that McLernon et al. and Ratna et al. have excluded. First of all, none of them included treatment with sperm donations in their study populations [3], [9], [10]. There are several reasons for a patient or couple to use donated sperm, such as male infertility and the couple being same-sexed. The advantages of a prediction success score is highly relevant to this patient group as well which motivates the inclusion of patients using donated sperm in this project. Secondly, McLernon et al. and Ratna et al. did not include the BMI of the woman. This feature is included in the Sahlgrenska dataset which provides further information, believed to be relevant to the ML models.

1.5 Limitations

To define the scope of the project, a set of assumptions and limitations have been established, particularly regarding the models evaluated and the dataset used. Given the importance of interpretability in clinical applications, this project focuses on relatively simple but still reliable models. While recent studies have explored more complex ML models, such as neural networks and hybrid methods like AdaBoost combined with genetic algorithm for feature selection, this project employs logistic regression, SVM and RF. These models are assumed to strike a balance between predictive performance and transparency, making them well-suited to the dataset's size and complexity. A more detailed discussion can be found in section 4.2 regarding model selection.

Another limitation of this project is that the variable AMH had too much missing data and was therefore excluded. Moreover, external validation has not been performed in this project. The models were developed and evaluated using data from Sahlgrenska University Hospital which may limit the generalizability of the findings to other clinical settings or populations. Without validation on external datasets, the extent to which the model can be reliably applied elsewhere remains uncertain. However, this limitation can also be seen as a potential advantage as the model is likely well-calibrated to the specific clinical practices and patient demographics at Sahlgrenska. This may result in better performance in that local context.

1.6 Societal and ethical aspects

When using ML for healthcare purposes, one must address societal and ethical challenges. Does the project impose a patient risk? What happens if the model ends up in the wrong hands? The most adverse societal and ethical aspects will be discussed in this section. This project has ethical approval from the Swedish Ethical Review Authority, applied for by the Department of Reproductive Medicine at Sahlgrenska.

A model that provides an individualized estimate of success after a failed first cycle

could help patients prepare mentally and emotionally. Importantly, such a model is intended to support, not replace, patients own decisions about whether to proceed with further treatment. In this way, the model may contribute positively to patient autonomy and well-being.

However, an IVF prediction model could potentially have the risk of being used to deny IVF treatment to patients if the chance of a live birth is predicted to be too low. Also, such a model could be used by private IVF companies to sift through a waiting list and pick the patients with the highest possibility of success.

It is important to emphasize that this is solely a research project and the developed model will not be applied in clinical practice in the near future and if it is, not before it has been thoroughly validated. We consider these reasons enough for this research project to be ethically and socially responsible.

2

Theory

This section provides an overview of key machine learning concepts and introduces the specific models relevant to this project. Additionally, the section covers feature scaling, model performance metrics and methods for improving interpretability.

2.1 Machine learning

After introducing the fundamental concepts of machine learning, the core models central to this project are described: logistic regression, SVM and RF. A solid understanding of these models is crucial for the analysis performed throughout the project.

2.1.1 Fundamental concepts within machine learning

ML refers to a collection of methods used to identify a function $f : X \to \hat{y}$ that maps a set of input variables X to a predicted output $\hat{y}$, aiming to approximate the true output y [2]. In supervised learning, the ML algorithm uses observed data with known labels y to train the function f. Supervised learning tasks include both regression (predicting continuous values) and classification (predicting discrete labels). This thesis focuses on a classification problem—predicting whether a live birth occurs—thus framing the discussion of machine learning specifically within the context of classification.

The objective of a classification task is to estimate the conditional probability $P(y \mid x)$ that an input x belongs to class y [2]. For example, in the context of IVF outcome prediction, x may represent a set of patient features such as age and BMI, while $y \in \{0, 1\}$ could indicate whether the patient achieves a live birth. The ML algorithm assigns a predicted label $\hat{y}$ by applying a decision rule, such as predicting $\hat{y} = 1$ if $P(y = 1 \mid x) > 0.5$. During training, the model has access to input-label pairs (x, y) and learns the function f by minimizing a loss function over the training data. In this context, f is referred to as a classifier.

The dataset used when training a classifier is typically represented as a feature vector $\boldsymbol{x} \in \mathbb{R}^l$, where l is the number of features in the dataset [2]. For the Sahlgrenska dataset, $l = 20$, including variables such as patient age, BMI and duration of infertility. Each feature vector $\boldsymbol{x}$ represents a single IVF cycle for an individual patient. The corresponding dataset thus consists of N labeled examples $\{(y_n, \boldsymbol{x}_n)\}_{n=1}^{N}$, where

y_n denotes the outcome label for $\boldsymbol{x}_n$. Since the data is retrospective in this project, the true labels y (live birth or not) are known in advance.

To train a classifier, it is essential to define a criterion for evaluating how well a candidate function f fits the data [2]. In parametric models, this is typically achieved by measuring the discrepancy between the predicted label $\hat{y}$ and the true label y using a loss function $\mathcal{L}$. The cumulative loss across all data points, known as the cost, serves as the objective to minimize during training. The choice of loss function depends on both the ML model and the specific characteristics of the classification task.

The fundamental goal of a classifier is to establish a decision boundary that separates classes within the feature space. In parametric models, this involves finding a parameter vector $\boldsymbol{\theta}$ that defines a boundary where samples from different classes fall on opposite sides—yielding positive scores for one class and negative scores for the other. By contrast, nonparametric models, such as decision trees, construct the boundary by iteratively selecting thresholds on input features to partition the data into homogeneous class regions.

Beyond minimizing training error, a key challenge in ML is ensuring that the model generalizes well to unseen data [16]. Regularization techniques are employed to reduce test error without significantly affecting training error. Regularization introduces additional constraints or penalties into the learning algorithm, which can be tailored based in prior knowledge (e.g., restricting the parameter space) or applied more generally. One form of regularization is the use of ensemble methods, which improve predictive performance by combining multiple individual models.

2.1.2 Logistic regression

Logistic regression is a discriminative model used for classification tasks [2]. It is a generalized linear model, as it uses a link function to transform a linear combination of input features into a probability estimate [17]. The link function must be continuous and map the entire real line to an interval $[0, 1]$, a criterion fulfilled by the logistic function. The logistic function is defined as

$$y = \frac{1}{1 + e^{-z}},$$

and is visualized in figure 2.1 below. In the equation, y represents the predicted probability which lies between 0 and 1 and z is the input to the logistic function, typically a linear combination of the features (e.g., $\boldsymbol{\omega}^\top \boldsymbol{x} + b$). As can be seen in figure 2.1, the logistic function has an S-shaped (sigmoidal) curve. The function is centered around $z = 0$, where the output is 0.5. In the context of IVF outcome prediction, $z = 0$ means the model assigns equal likelihood to a live birth or no live birth, resulting in a logistic output of 0.5 based on the patient's features.

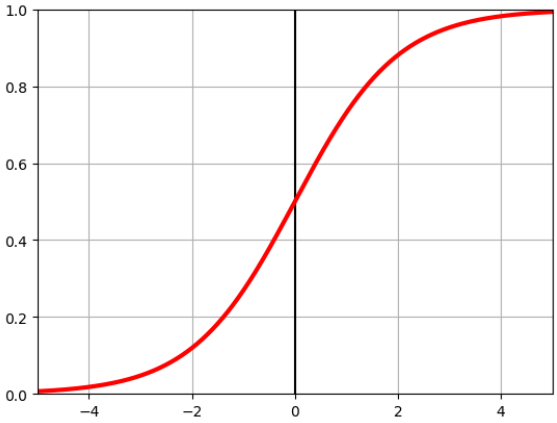


Figure 2.1: The logistic function $y = \frac{1}{1+e^{-z}}$. The variable y denotes the predicted probability, ranging from 0 to 1, while z is the input to the logistic function, typically a linear combination of the input features, such as $\boldsymbol{\omega}^\top \boldsymbol{x} + b$.

By substituting $z = \boldsymbol{\omega}^\top \boldsymbol{x} + b$ into the logistic function, the following expression is obtained:

$$p(y = 1 \mid \boldsymbol{x}) = \frac{1}{1 + e^{-(\boldsymbol{\omega}^\top \boldsymbol{x} + b)}}.$$

This equation can then be rewritten by taking the natural logarithm of both sides, leading to the following form

$$\ln \frac{p(y = 1 \mid \boldsymbol{x})}{p(y = 0 \mid \boldsymbol{x})} = \boldsymbol{\omega}^\top \boldsymbol{x} + b,$$

where $\boldsymbol{\omega}$ and b denote the parameters that define the discriminant function [17]. This equation represents the log-odds of the probability $p(y = 1 \mid \boldsymbol{x})$, such as for a live birth, relative to $p(y = 0 \mid \boldsymbol{x})$, which corresponds to no live birth. It shows that the log-odds are modeled as a linear function of the input features $\boldsymbol{x}$. The model parameters $\boldsymbol{\omega}$ and b are typically estimated using the maximum likelihood method by maximizing the probability of the observed data. The log-likelihood function is defined as

$$\ell(\boldsymbol{w}, b) = \sum_{i=1}^{m} \ln p\left(y_i \mid \boldsymbol{x}_i; \boldsymbol{w}, b\right),$$

where $p(y_i \mid \boldsymbol{x}_i; \boldsymbol{\omega}, b)$ is the probability of observing the true label y_i given the input features $\boldsymbol{x}_i$, parameterized by the coefficient vector $\boldsymbol{\omega}$ and bias term b [17]. Each coefficient ω_j represents the weight or influence of the corresponding input feature x_j and the log-odds of the outcome. If there are 20 features in the dataset, there will be 20 coefficients—one for each feature.

2.1.3 Support vector machines

SVM focus on learning a decision boundary that separates classes by maximizing the margin between them, rather than modeling class probabilities as in logistic regression [2]. SVM construct a discriminant function f, which is trained by optimizing the model parameters to minimize an empirical loss function $\mathcal{L}(y_n, f(\boldsymbol{x}_n))$, where y_n is the true label of a data point n of N data points and $f(\boldsymbol{x}_n)$ is the predicted score of the data point $\boldsymbol{x}_n$. The cost $J(f)$ of the discriminant function is computed by summarizing the losses of all individual points in the dataset. The cost function is given by:

$$J(f) = \sum_{n=1}^{N} \mathcal{L}(y_n, f(\boldsymbol{x}_n)).$$

Figure 2.2 below visualizes an example of an SVM separating the green and blue points, which represent two different classes (for example, live birth and no live birth). The discriminant function $f(\boldsymbol{x})$ is represented by the black line and is called the decision boundary. It represents a relationship between feature 1 and feature 2 that defines the classification of a data point. For a point to belong to the blue class, the point must exceed that relationship. The dashed red lines define the margin, which defines the distance between the decision boundary and the points that lie closest to it. These closest points, shown in red in the figure, are called support vectors and determine the direction of the margin and thus the decision boundary. The decision boundary is located in between the margins, meaning the distance from the boundary to either margin has the same size for both margins.

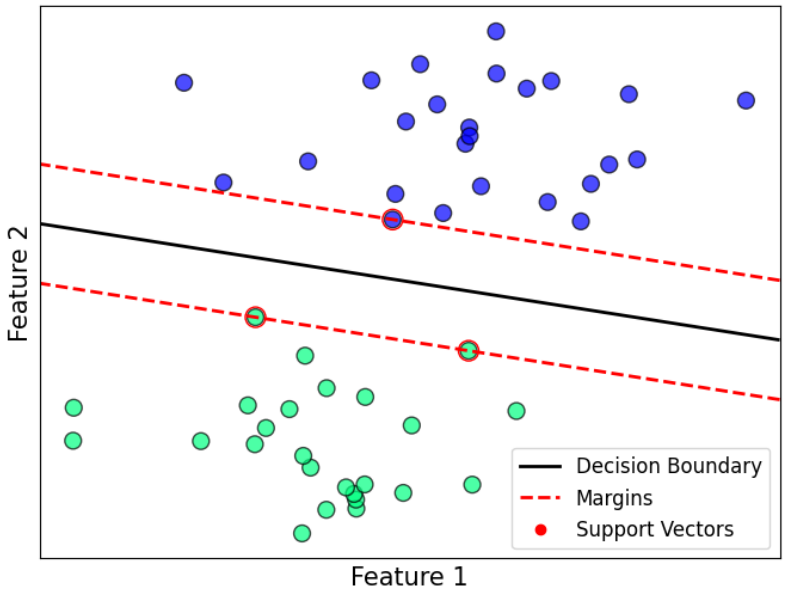


Figure 2.2: Illustration of an SVM for two features divided by a decision boundary (black line) and two margins (red dashed line). The red-circled dots indicate the support vectors. The blue and green dots represent data points from two different classes, where one blue dot and two green dots function as support vectors.

The objective of the optimization task is to identify a hyperplane

$$f(\boldsymbol{x}) = \boldsymbol{\theta}^T \boldsymbol{x} + \theta_0$$

in the feature space, that maximizes the margin between two classes [2]. The function $f(\boldsymbol{x})$ returns a continuous value, which is then used to assign a class label. If $f(\boldsymbol{x}) > 0$, the data point is classified as class 1 (e.g., live birth), and if $f(\boldsymbol{x}) < 0$, the data point is classified as class 0 (e.g., no live birth). So, even though $f(\boldsymbol{x})$ is continuous, the classification is made by checking on which side of the hyperplane each point lies. The variable $\boldsymbol{\theta}$ defines the direction of the hyperplane in the multidimensional space, with its position determined by θ_0. In figure 2.2 above, the hyperplane is the two-dimensional decision boundary that separates feature 1 and feature 2. For a dataset where the classes have no overlap, it is possible to separate the classes linearly. Although, in reality, classes in datasets often overlap, complicating the task of maximizing the margin. Points on the correct side of the margin have smaller loss than those on the wrong side.

A commonly used loss function for SVMs is the hinge loss [2]. It is defined as

$$\mathcal{L}_\rho(y, f(\boldsymbol{x})) = \max(0, \rho - y \ \cdot f(\boldsymbol{x}))$$

and does not only penalizes the model for misclassifying data points, but also imposes a penalty for points that are too close to the decision boundary. If the product between the true label y and the predicted score $f(x)$ exceeds a threshold value ρ, the loss is zero. Otherwise, the loss is greater than zero and it decreases linearly as the product $y \ \cdot \ f(\boldsymbol{x})$ approaches the threshold. The hinge loss is visualized in figure 2.3.

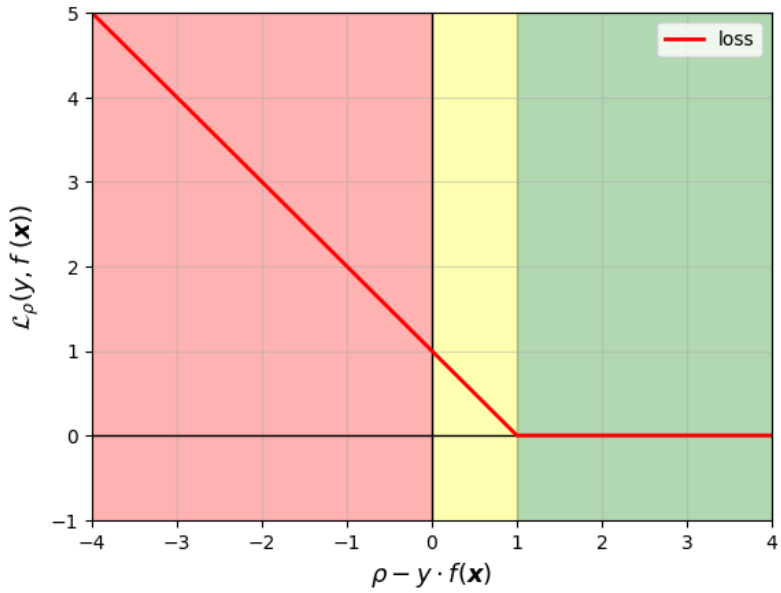


Figure 2.3: Illustration of the hinge loss (red line) as a function of the difference between the threshold value ρ and the product between the true and predicted label $y \cdot f(\boldsymbol{x})$ (x-axis). The hinge loss is a commonly used loss function for SVMs [2].

The SVM seeks to minimize the overall cost

$$J\left(\boldsymbol{\theta}, \theta_0\right) = \frac{1}{2}\|\boldsymbol{\theta}\|^2 + C\sum_{n=1}^{N} \mathcal{L}_\rho\left(y_n, \boldsymbol{\theta}^T\boldsymbol{x}_n + \theta_0\right),$$

where the first term promotes a large margin and the second term penalizes data points that were classified with the wrong label. C is a regularization parameter.

2.1.4 Random forest

As mentioned in the beginning of section 2.1, ensemble models is one of many regularization techniques. One of the most well-known ensemble methods is RF, which is based on DTs. Therefore, this section begins with an introduction to DTs.

A DT generates predictions by iteratively partitioning the input data based on a sequence of decision rules [18]. Each rule corresponds to a node in the tree, with the topmost node referred to as the root node. Nodes branching from the root are called decision nodes, while the terminal nodes, known as leaf nodes, represent the final classification outcome (e.g., live birth or no live birth). The construction of a DT follows a divide-and-conquer strategy, selecting the optimal feature and threshold at each step to split the data. The algorithm aims to achieve class homogeneity within each leaf node, such that all data points within a leaf node belong to the same class. Achieving homogeneity is more feasible with a smaller DT, as larger trees tend to increase complexity and may lead to overfitting, where each subtree is constructed from a limited number of data points. To mitigate overfitting, DTs undergo pruning, a process that removes branches with minimal importance to enhance generalizability.

Figure 2.4 shows an example of a small hypothetical DT for the outcome of an IVF treatment. The root node uses the age of the woman to separate the dataset into two groups. Two different decision nodes depend on whether the woman is younger than 35 or not. For example, a 29-year-old woman with twelve collected oocytes (female eggs) will be predicted to experience a live birth, while a 39-year-old woman who has been trying to become pregnant for two years will not experience a live birth.

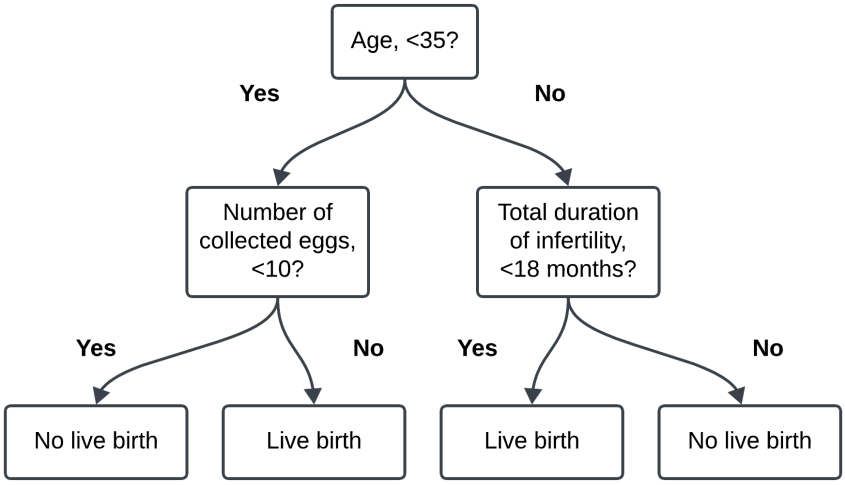


Figure 2.4: Example of a small hypothetical DT that predicts live birth or not.

To find the best features to use for a data split in a DT, the level of impurity is measured, which is a way of quantifying disorder [2]. One way of quantifying impurity is by calculating the entropy, which at node t is defined as

$$I(t) = -\sum_{m=1}^{M} P(\omega_m \mid t) \log_2 P(\omega_m \mid t),$$

where M is the number of classes and $P(\omega_m \mid t)$ is the probability of observing class ω_m at node t [2]. When the tree-building algorithm selects features that split a set of data points in two subsets, it seeks to maximize the difference between the impurity $I(t_1)$ and $I(t_0)$ of each new node t_1 and t_0.

Another way of measuring the impurity is the Gini index. Compared to entropy, the Gini index results in a slightly sharper maximum [2]. It is defined as:

$$I(t) = \sum_{m=1}^{M} P(\omega_m \mid t)(1 - P(\omega_m \mid t)).$$

One way to mitigate the risk of bias for DTs is to create an ensemble of DTs that together vote about the output. One example of an ensemble model is RF [18]. A RF model consists of multiple DTs and can be applied for both classification and regression problems. In the case of a classification problem, the outcome with the majority of the votes is output. The model is trained using the bagging method, which generates random samples of the training data for individual training of the models and is a common method for reducing variance in a noisy dataset. The data subsets are chosen with replacement, meaning each sample can be selected multiple times. Feature selection is utilized automatically using feature bagging, which generates random subsets of the features and ensures a low correlation between each decision tree. About one-third of each bagging subset is used for cross-validation

to test the model's performance and is called the out-of-bag (oob) sample. Bagging and feature bagging help account for variations in the data to prevent an overfitted and biased model, which individual DTs are prone to.

A RF will help mitigate bias that occurs in only one tree. The DT in figure 2.4 will probably have poor accuracy and might also introduce a bias. However, by creating more trees and letting them vote, the outcome becomes more reliable. One tree might say that a couple will not experience a live birth, while two other trees say otherwise. Such a result is much more reliable than only one tree.

2.2 Restricted cubic splines

Some statistical models, logistic regression included, assume a linear relationship between each predictor and the outcome [19]. Consequently, a one-unit change of a predictor is expected to have a constant effect on the outcome, no matter the value of the predictor. For example, a one-year increase of a patient's age is assumed to result in the same change of probability of a live birth, regardless if the patient's age is 28 or 38 years. This assumption imposes an issue when logistic regression is applied to the Sahlgrenska dataset, as there is a well-established non-linear relationship between the patient's age and the likelihood of achieving a live birth [1]. The probability remains relatively stable between the ages of 26 and 33, declines gradually between 33 and 38 and then decreases more drastically after the age of 38.

One way of addressing non-linear relationships while still assuming linear relationships between each predictor and the outcome, is to use flexible methods that are better fitted to capture such patterns. Restricted cubic splines (RCS) are a commonly used technique for this purpose [19]. Instead of fitting a single straight line through the data points, RCS divide the range of a predictor into intervals using specific points called knots. In each interval, a separate cubic curve is fitted. These curves are joined smoothly at the knots to ensure the overall function is continuous. Beyond the outermost knots, the function is constrained to be linear, which helps prevent overfitting in the tails of the variable's range.

Mathematically, when K knots are placed at values $t_1 < t_2 < \cdots < t_K$, the transformed version of the predictor X includes the original variable along with $K - 2$ new variables that capture the non-linear shape [19]. These variables can then be included in regression models to account for complex relationships between predictors and outcomes. When five knots are used, typically based on percentiles or clinically relevant cutoffs, three additional variables are created to represent the non-linear part of the relationship. These variables can be calculated using the following formulas:

$$X_1 = (X - t_1)_+^3 - \frac{t_5 - t_1}{t_5 - t_4}(X - t_4)_+^3 + \frac{t_4 - t_1}{t_5 - t_4}(X - t_5)_+^3$$
$$X_2 = (X - t_2)_+^3 - \frac{t_5 - t_2}{t_5 - t_4}(X - t_4)_+^3 + \frac{t_4 - t_2}{t_5 - t_4}(X - t_5)_+^3$$
$$X_3 = (X - t_3)_+^3 - \frac{t_5 - t_3}{t_5 - t_4}(X - t_4)_+^3 + \frac{t_4 - t_3}{t_5 - t_4}(X - t_5)_+^3$$

RCS have been used in several clinical prediction studies, including those by McLernon et al. and Ratna et al., to model the effect of the women's age on IVF success rates [3], [9], [10]. Apart from patient age, McLernon et al. 2016 used RCS to adjust for treatment year and the number of eggs retrieved, which also had non-linear relationships with the outcome. BMI and AMH levels were adjusted for in McLernon et al. 2022. In the study by Ratna et al. 2022, RCS were utilized to account for the time interval between the first and second egg retrieval, treatment year, as well as the number of eggs retrieved at the first complete cycle.

2.3 Feature scaling

When preparing a dataset for modeling, feature scaling is often an essential preprocessing step. Feature scaling ensures that all features are on a comparable scale, maintaining the shape of their distributions [20]. This can significantly improve the performance for many ML models. However, the need for feature scaling depends on the type of model being used. A study evaluating the impact of the two common feature scaling methods – standardization and normalization – on a medical dataset, as well as the performance with no scaling. The study showed that different scaling methods performed best for SVM, RF and logistic regression [20]. The study showed that normalization performed best for SVM, standardization was most effective for logistic regression and no feature scaling was best for RF.

SVM is highly sensitive to the scale of the variables since it is a distance-based algorithms [20]. In contrast, RF are generally not sensitive to feature scaling, as they make decisions based on feature thresholds and splits. On the other hand, logistic regression can benefit from feature scaling, to ensure the features contribute equally to the model.

Standardization, known as standard scaling, transforms the features of a dataset such that they have a mean of zero and a standard deviation of one [20]. This method is useful when the data is normally distributed. The formula for standardization is given by

$$x_{\text{standardized}} = \frac{x - \mu}{\sigma},$$

where x is the original value of the feature, μ is the mean of the feature and σ is the standard deviation of the feature.

Normalization, often known as min-max scaling, rescales the features so that they lie within a specified range between zero and one [20]. Min-max scaling is more robust to small standard deviations and it preserves zero entries in sparse data. It is helpful when the data distribution does not follow a normal distribution. The mathematical formula for min-max scaling is

$$x_{\text{normalized}} = \frac{x - x_{\min}}{x_{\max} - x_{\min}},$$

where x is the original value of the feature, $x_{\min}$ is the minimum value of the feature and $x_{\max}$ is the maximum value of the feature.

2.4 Model calibration

When evaluating model performance, it is important to ensure the predicted probabilities reflect the true likelihood of an event [21]. How well a model is calibrated can be evaluated using a calibration curve. As shown in figure 2.5, a calibration curve plots the predicted probabilities against the observed frequencies of the event. For instance, if the model predicts a 20% chance of an event, the event should occur about 20% of the time. The curve visualizes how well this relationship holds across different probabilities, showing whether the model over- or under-predicts the event likelihood. A perfectly calibrated model follows the diagonal line of the calibration plot. Good calibration ensures that the predicted probabilities accurately reflect the observed outcomes.

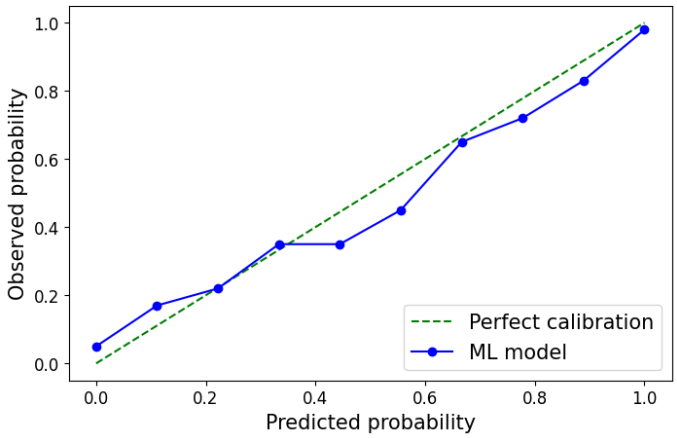


Figure 2.5: Example of a calibration curve, illustrating how well the model's predicted probabilities align with the actual outcomes. A perfectly calibrated model follows the diagonal line.

Calibrating a model can be done in different ways. Two commonly used approaches are platt scaling and isotonic regression [21]. Platt scaling fits a logistic regression model to the output of the original model using a separate dataset (i.e., a validation set). The calibration is done with the following formula

$$P(y = 1|\mathbf{f_i}) = \frac{1}{1 + \exp(A(\mathbf{f_i}) + B)},$$

where y is the true label and $\mathbf{f_i}$ is the raw output of the uncalibrated model. The parameters A and B are real numbers determined via maximum likelihood when fitting the regressor. The method is often most effective for small sample sizes or when the classifier is under-confident and has similar calibration errors across output values [21]. It works best when the calibration error is symmetrical and the

outputs for each class are normally distributed with equal variance. However, this assumption can be problematic in imbalanced datasets.

In contrast, isotonic regression is a non-parametric technique that fits a non-decreasing function to the predicted probabilities using a separate calibration dataset [21]. It minimizes

$$\sum_{i=1}^{n} \left(\mathbf{y}_i - \hat{f}_i\right)^2,$$

where $\mathbf{y_i}$ is the true label and $\hat{\mathbf{f}}_\mathbf{i}$ is the predicted probability. Compared to platt scaling, isotonic regression is more flexible because the only restriction is that the that mapping function must be monotonically increasing [21]. Isotonic regression generally performs better than platt scaling when the dataset contains more than approximately 1,1000 samples as it avoids overfitting.

2.5 Model evaluation

There are several ways to quantify the performance of a ML model and each metric bring its own perspective. An description of the metrics used in this project are presented below.

Accuracy Accuracy measures the general proportion of correct predictions made by the model [22]. It is calculated as

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN},$$

where:

- TP refers to true positives (women who had a live birth were correctly predicted to have a live birth)
- TN refers to true negatives (women who had no live birth were correctly predicted to not have a live birth)
- FP refers to false positives (women who had no live birth were incorrectly predicted to have a live birth)
- FN refers to false negatives (women who had a live birth were incorrectly predicted to have no live birth)

However, high accuracy can be misleading when the dataset is imbalanced, i.e. one outcome is much more common than the others [22]. In this project, this imbalance could occur because more individuals do not have a live birth through IVF than those who do. A high accuracy can be reached by labeling all instances as the more prevalent class, which paints a false picture of the model's ability to separate between the classes. Despite its limitation, accuracy is commonly used because it

is simple to calculate and understand, providing a quick overview of the model's overall performance.

Sensitivity Sensitivity, also known as the True Positive Rate, is defined as

$$\text{Sensitivity} = \frac{TP}{TP + FN}$$

and measures the proportion of actual positive cases (e.g. successful IVF outcomes) that are correctly identified by the model [22]. A sensitivity of 100% means that all women who had a live birth were predicted to have a live birth.

Specificity Specificity, also known as True Negative Rate, measures the proportion of actual negative cases (e.g. unsuccessful IVF outcomes) that are correctly identified by the model [22]. It is calculated as

$$\text{Specificity} = \frac{TN}{TN + FP},$$

where a specificity of 100% means that all women who did not experience a birth were predicted to not experience a live birth.

There is often a trade-off between sensitivity and specificity [22]. In many cases, improving sensitivity might reduce specificity and vice versa. A highly sensitive test is able to find almost all positive instances, but does so by also identifying lots of false positive (negative instances labeled as positive). The same reasoning lies behind a high specificity, which indicates a high chance the test is negative, given a negative label. No model is perfect, meaning perfect sensitivity and specificity can not be achieved at the same time. To accept a lower sensitivity means the model will be a bit pessimistic, meaning women will be predicted as not having a live birth, but in the end experiencing a live birth. In contrast, to accept a lower specificity will result in an optimistic model, meaning women will be predicted to experience a live birth, but end up not having a baby. Furthermore, since accuracy, sensitivity and specificity often are used in evaluation metrics in clinical trials, their usage enables clinical readers to put the metrics in context.

Area under the receiver operating characteristic (AUROC) ML models are frequently evaluated using the Receiver Operating Characteristic (ROC) curve [23]. In IVF prediction, the ROC curve shows how well a model separates couples who will have a live birth from those who will not. It reflects the probability that a randomly selected patient who achieves a live birth is ranked higher than a randomly selected patient who does not [24]. As shown in figure 2.6, this is done by plotting the true positive rate against the false positive rate. A true positive means correctly predicting a live birth, while a false positive means an incorrectly predicted live birth. The AUROC score ranges from 0.5 (random guessing) to 1.0 (perfect classification).

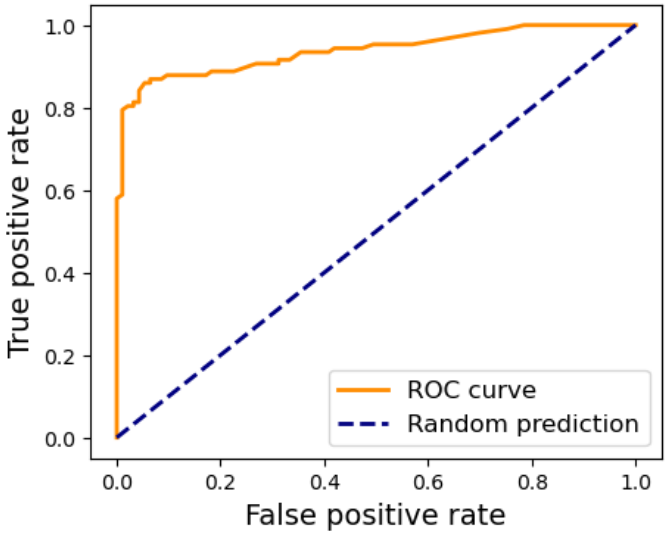


Figure 2.6: Example of a ROC curve, illustrating how well a model distinguishes between classes. The ROC curve, shown in orange in the figure, plots the true positive rate against the false positive rate at various thresholds. The area under the ROC curve (AUROC) quantifies the model's overall ability to rank a randomly selected positive sample higher than a randomly selected negative sample. An AUROC of 0.5 indicates random guessing, while a value of 1.0 reflects perfect classification.

Matthews correlation coefficient (MCC) MCC is another commonly used evaluation metric, particularly for binary and multiclass classification tasks [25]. MCC ranges from -1 to $+1$. A perfect classifier achieves an MCC score of $+1$, an average random classifier yields a score of 0 and a completely incorrect classifier results in an MCC of -1. The MCC formula is defined as

$$\text{MCC} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

and takes into account all four components of the confusion matrix. In this project, a high MCC value indicates that the model can accurately predict both women who will experience a live birth and those who will not. This ensures that any potential imbalances in the dataset are effectively captured.

Brier score Another common evaluation method is the Brier score [26]. It evaluates the accuracy of probabilistic predictions and measures the mean squared error between the predicted probability and the actual outcome. The Brier score ranges from 0 to 1, where zero refers to a perfect prediction, meaning the predicted probability exactly matches the actual outcome. A Brier score of 1 indicates that the predicted probabilities are far from the actual outcome. The Brier score is calculated as

$$\text{Brier score} = \frac{1}{N}\sum_{i=1}^{N}(p_i - y_i)^2,$$

where:

- N is the number of observations
- p_i is the predicted probability that the event occurs for observation i
- y_i is the actual outcome for observation i (1 if the event occurs, 0 if it does not occur)

Assume P represents the predicted probability that a woman will experience a live birth. If $P = 1$ and she experiences a live birth, the Brier score is 0, indicating a perfect score. On the other hand, if $P = 1$ and she does not experience a live birth, the Brier score is 1, which represents the worst possible score. If $P = 0.3$ and she experiences a live birth, the Brier score is $(0.3 - 1)^2 = 0.49$. Similarly, if $P = 0.7$ and she experiences a live birth, the Brier score is $(0.7 - 1)^2 = 0.09$.

2.6 Interpretability

Building trust in ML models depends heavily on their interpretability. This means that the reasoning behind the model's decisions can be understood, allowing insight into how the model arrives at its conclusions [27]. The reasoning behind ML models is often trying to be understood in terms of feature importance, meaning to find the features with the most significant impact on the prediction. This method can be applied on two levels: global and local explanations. Local explanations display the reasoning behind the classification of one patient, by showing the features with the greatest impact on the individual outcome. Global explanations are used to explain the reasoning for a larger group of patients. Understanding feature importance helps ensure the correct reasoning behind decisions, which is essential for trust between engineers, clinicians and patients.

In the context IVF prediction, certain features are known to affect the likelihood of a live birth. One of them is the age of the woman, for which a higher age is strongly correlated with a smaller chance of getting a live birth. By creating an interpretable system, it is possible to validate that the model's basis of decision is reasonable. If only features that have a very weak link to the chances of success are used by the model, then it is very likely the model is biased.

Logistic regression is highly interpretable, as one weight for each feature is trained. This makes it possible to observe the contribution of different features. The reasoning behind SVMs and RF models can be understood when examining only a few features. However, an increasing number of features means an increasing complexity behind the reasoning of the models. For example, a RF model can be seen as a "black box" model due to its ensemble of multiple decision trees [28]. While understanding the full decision-making process across all trees is difficult, the internal structure remains interpretable because the construction and logic of individual trees are well understood. As the Sahlgrenska dataset consists of multiple features, tools for global explanations help understand the reasoning behind the model outputs.

Feature importance in ML can be evaluated using shapley additive explanations

(SHAP), a method rooted in game theory [29]. It assigns a score to each feature, indicating its contribution to the model's prediction, either by increasing or decreasing the likelihood of the outcome. In the IVF dataset, for example, a woman's age is expected to correlate with negative outcomes (no live birth). SHAP figures are presented in the result section in this report, with an explanation of how to interpret the figures.

3

Data Preparation

To answer the three research questions described in section 1.3, the majority of the overall project timeline was dedicated to data preparation. This section describes how the data preparation was carried out, which involved selecting relevant patients based on the study criteria, creating visualizations to better understand the data, combining multiple data files into one comprehensive dataset and replacing and imputing values. During the data preparation phase, many small data issues were identified which, while seemingly minor on their own, could result in a significant combined impact. Detecting and correcting these issues was essential to create an accurate and reliable dataset that appropriately reflects the patient population.

3.1 Inclusion and exclusion criteria

Several data cleaning steps were required to meet the datasets inclusion and exclusion criteria. An overview of this process is presented in the flowchart in figure 3.1. The inclusion and exclusion decisions made on patient-level are marked in orange in the figure; pink represent decisions made regarding individual cycles; and ET-related criteria are marked in yellow.

The dataset used in this project was provided by the Department of Reproductive Medicine at Sahlgrenska University Hospital in Gothenburg and has been collected from the health record systems WinIVF and Melior. For treatment data, a script used for collection of data to the Swedish National Registry of Assisted Reproduction (Q-IVF) was used. The data comprises patients who began their first IVF cycle at Sahlgrenska between January 2016 and December 2020 and consists of approximately 5,500 initiated IVF cycles form 3,200 unique women, with 3,100 having partners.

Patients who underwent treatment using donated oocytes, preimplantation genetic testing (PGT) or autologous thawed oocytes were excluded from the study. Autologous thawed oocytes refer to a patient's own eggs that were previously frozen and later thawed during treatment, as opposed to donated oocytes.

Unlike several previous studies, this dataset includes patients who underwent treatment using donated sperm. The use of donated sperm may be necessitated by a range of factor, including the absence of a male partner or male factor infertility. Notably, patients using donated sperm were excluded in the studies by McLernon et

al. (2016), McLernon et al. (2022) and Ratna et al. (2022) [3], [9], [10]. In contrast, including these patients is in this project considered to provide a more accurate reflection of the real-world target population. Moreover, patients who achieved a spontaneous pregnancy during the course of IVF treatment resulting in a live birth were excluded from the analysis.

This study is limited to include a maximum of a patient's three initiated IVF cycles, as the Swedish healthcare system subsidizes up to three IVF treatments for Swedish citizens and only until a live birth occurs. Hence, in order to capture a complete picture of a patient's journey through all available treatments, the patient's second and third IVF cycles are included in the dataset up until December 2021 and all ETs performed up until December 2022. Cycles and ETs occurring after a live birth were excluded from the analysis. In the project, only ETs within one year of the IVF cycle start were included to allow fair comparisons between cycles, as nearly all ETs are performed within that timeframe.

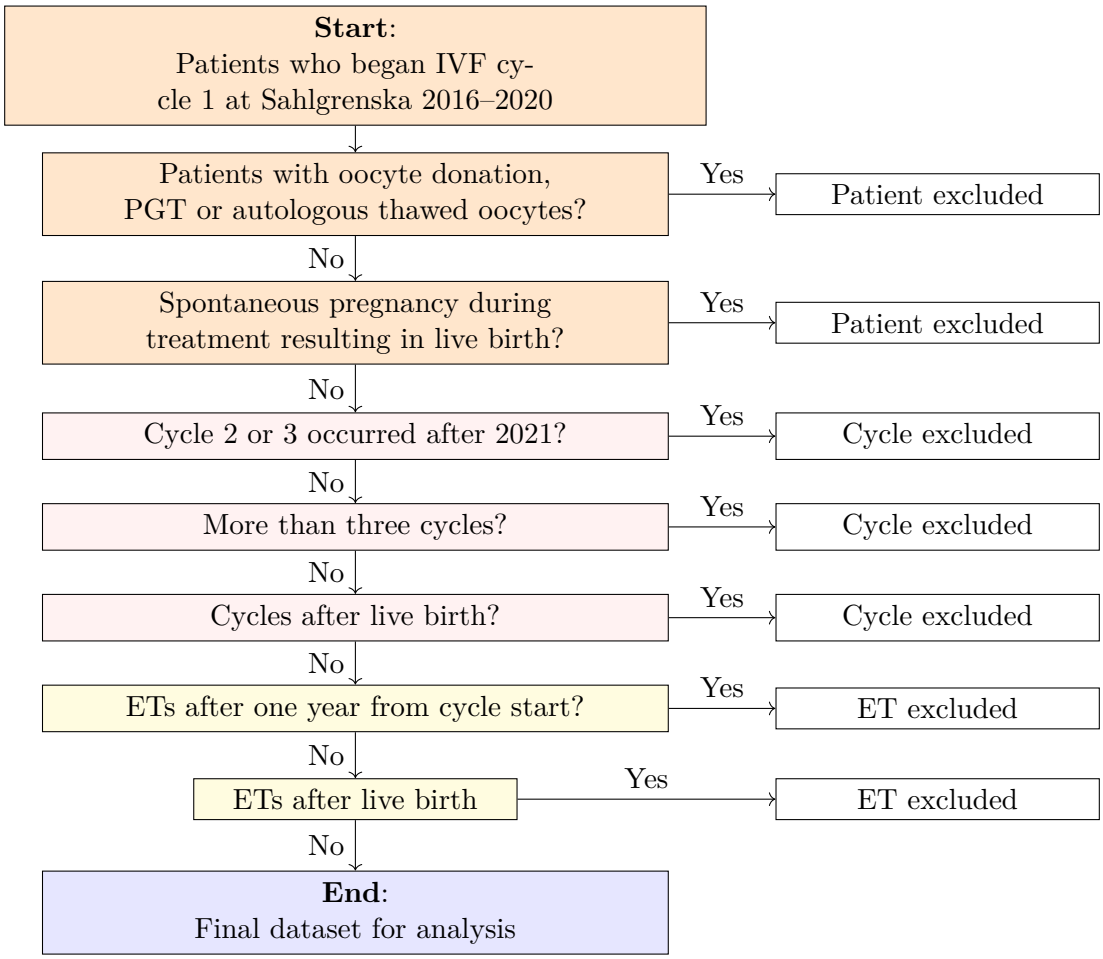


Figure 3.1: Flowchart of inclusion and exclusion criteria for the IVF dataset. Decisions made on patient-level are marked in orange; pink represent decisions made regarding individual cycles; and ET-related criteria are marked in yellow.

In addition to the study criteria described in figure 3.1, other inclusion and exclusion

aspects have been discussed throughout the project. For instance, some patients initiated their hormonal treatment but did not proceed to oocyte aspiration due to minor, self-limiting conditions such as a common cold, stomach pain or influenza. These interrupted cycles were kept in the dataset, as they reflect the real-world variability of IVF treatment.

A few patients out of those with interrupted cycles have been excluded. An interrupted cycle is defined as a cycle that ended without resulting in a live birth and before all available good quality embryo (GQE)s had been transferred, regardless of whether an ET occurred more than one year after the initiation of treatment. Among these, some never began stimulation and some completed stimulation and did a total freeze but never underwent ET. Such patients were excluded from the dataset due to more ambiguous or non-medical reasons, as identified by clinicians through a review of medical records. These reasons included, for example, relationship dissolution, misclassifications of treatments as IVF or patient transferring from Sahlgrenska to private fertility clinics.

3.2 Data overview

To provide a clearer insight of the dataset after the inclusion- and exclusion criteria, the dataset was visualized using a Sankey diagram showed in figure 3.2. It presents the number of patients who started each IVF cycle, the number of patients who had a live birth after each cycle and the number of patients who discontinued their IVF treatment before having a live birth or completing all three cycles. The later group also includes all patients who discontinued their IVF treatment prior to undergoing oocyte aspiration or before transferring all available GQEs. Patients who did not continue to additional cycles and did not achieve a live birth are marked as "Dropouts" in the figure.

As shown in the figure, 3 217 patients started their first IVF cycle. Of these, 1 508 had a live birth within one year. 1 348 patients who did not have a live birth continued to cycle 2. In cycle 2, 498 patients had a live birth and 605 patients proceeded to cycle 3, where 171 had a live birth.

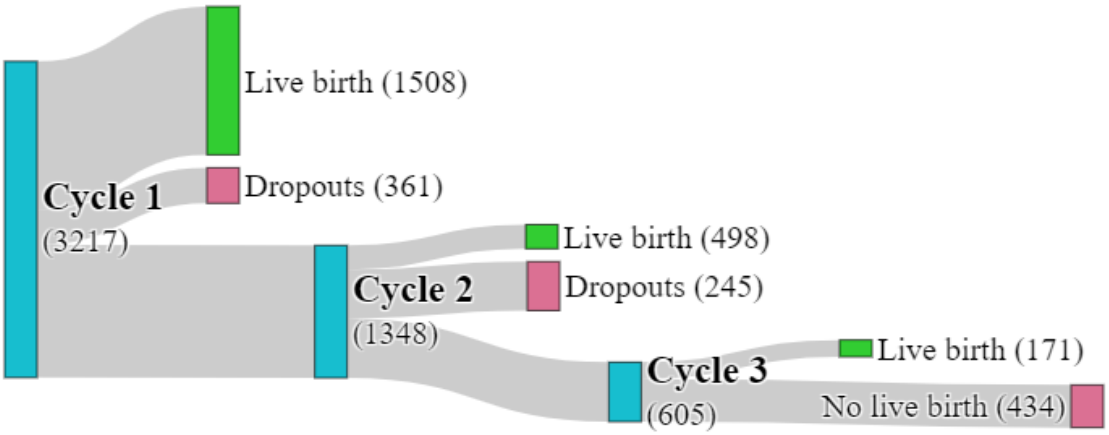


Figure 3.2: Visualization of all patients eligible for the dataset, showing the flow of patients across three IVF cycles. Of 3 127 patients who started their first IVF cycle, 1 508 had a live birth after the first cycle. Among those who did not, 1 348 continued to cycle 2, resulting in 498 live births. 605 patients proceeded to cycle 3, with 171 live births.

The amount of patients who did not continue to additional cycles and who did not achieve a live birth, marked as "Dropouts" in the figure, is relatively high. For example, the dropout rate between cycle 1 and 2 is 21.1% and between cycle 2 and 3 it is 28.8%. This issue is described in more detail in section 4.1. The dropout rates are calculated by dividing the number of patients who dropped out during a cycle by the number of patients eligible to continue, which is the number who started the cycle minus those who achieved a live birth during that cycle. As an example, the dropout rate between cycle 1 and 2 is calculated as:

$$\text{Dropout rate } (\%) = 100 \cdot \frac{361}{3217 - 1508}$$

The dataset comprises 21 features: ten categorical variables, ten numerical variables and one binary outcome variable indicating whether a live birth occurred. Table 3.1 lists the numerical variables and table 3.2 outlines the categorical variables (including the outcome variable). A statistical overview of the values is included in appendix A.

Table 3.1: All ten numerical variables present in the dataset.

Variable
Women age
Partner age
Female length, cm
Infertility duration, months
Total gonadotropin dose
Number of oocytes retrieved
Good Quality Embryos (GQE)
Number of frozen embryos

Continued on next page

Table 3.1 – continued from previous page
Variable
Number of ETs
Number of failed ETs

Table 3.2: All ten categorical variables in the dataset, including the outcome variable, with their respective subcategories.

Variable
BMI, kg/m^2
<20
20-24.9
25-29.9
30-34.9
>35
AFC right+left=total
<10
10-30
>30
Average menstrual cycle, days
<25
25-29.9
30–34.9
>35
Sperm origin
Ejaculated
Donated
Frozen ejaculated
Surgical sperm extraction
Planned fertilization method
IVF
ICSI
Infertility cause
Tubal factor
Unexplained
Male factor
PCOD
Hormonal
Endometriosis
Same sex couple/single woman
Egg factor
Other
Stimulation protocol
Antagonist
Agonist
Fertilization method for the transferred embryo
IVF

Continued on next page

Table 3.2 – continued from previous page

Variable
ICSI
Number of culture days at ET
1-3
4-8
Number of embryos transferred at ET
SET
DET
Live birth (per cycle)
Yes
No

3.3 Combining data

The dataset utilized in this project was compiled from multiple sources at different times, resulting in patient- and IVF cycle-related variables being distributed across several files. Consequently, substantial effort was devoted to data preprocessing in order to achieve a consistent dataset in a usable format. The raw data was obtained from three files: two post-treatment files and one pre-treatment file.

The first post-treatment file contained data from cycles conducted between January 2016 and December 2020, while the second covered data from cycles conducted between January 2021 and December 2022. The existence of two post-treatment files reflect the timing of data collection, which occurred on two separate occasions. Data for the 2016-2020 period was collected under the initial ethical approval, whereas data for 2021-2022 was added following an extension of the ethical approval to include one additional year of cycles and ETs. These two post-treatment files were combined to one single file, which throughout this text will be denoted as the post-treatment file.

The pre-treatment file contained variables collected prior to the patient's first IVF cycle, such as the cause of infertility, age, height and weight. This file was structured in a wide format, where each row corresponded to a single IVF cycle. In contrast, the post-treatment file included variables generated during the course of treatment, such as total gonadotropin dose, number of oocytes retrieved and number of embryos transferred. The post-treatment file was organized in a long format, where each row represented a single ET for a specific patient. As a result, data for one started IVF cycle could span multiple rows across this post-treatment file.

The goal of the data merging was to create a dataset in wide format, combining both pre- and post-treatment variables for each cycle. The intention with the data in its final format was for one row to represent one round of IVF. Table 3.3 provides an overview of the final dataset.

Table 3.3: A subset of the columns in the final dataset after merging all files together. The dataset is in wide format, with each row representing one IVF cycle. All values shown are fictional.

PRE-TREATMENT VARIABLES					POST-TREATMENT VARIABLES		OUTCOME
Personal identity number	**IVF cycle**	**Age**	**Infertility duration (months)**	**...**	**Number of oocytes retrieved**	**...**	**Live birth outcome**
YYYY-MM-DD-XXXX	1	32.1	14		15		0
YYYY-MM-DD-XXXX	2	33.4	24		17		1
YYYY-MM-DD-ZZZZ	1	28.9	18		9		0

After the data merge, it was observed that some patients had frozen embryos but no ETs recorded during their treatment cycles. In some cases, this was expected, as patients discontinued treatment during the IVF process or because ETs were performed more than one year after treatment start and had hence been excluded from the dataset. However, upon review by a clinician, it was suggested that some ETs had not been correctly recorded in the data collected from the source of the post-treatment data. To correct this, clinicians retrieved all FETs from the WinIVF database covering the years 2016–2022 and the data were then compared with the post-treatment file. Missing FETs were added to complete the dataset.

It was also found that 23 patients in the pre-treatment file were not included in the post-treatment file. It has been noted that these patients have been incorrectly assigned as declining to participate in the Q-IVF register, and therefore, their post-treatment variables were excluded in the original data collection. This was addressed by collecting the missing post-treatment data and adding it to the post-treatment file.

3.4 Data replacement and imputation

After the data files had been merged into a single dataset, decisions regarding replacement and imputation of missing values were made in close collaboration with two clinicians from the Department of Reproductive Medicine at Sahlgrenska University Hospital. An overview of the main data replacement and imputations is listed in table 3.4 and described below.

Pre-treatment variables As mentioned in the previous section, the pre-treatment file only contained data for cycles started between January 2016 and December 2020. Consequently, all second and third IVF cycles started during 2021 were not present in the pre-treatment file and hence missing from the raw data. To resolve this issue, the missing pre-treatment variables were filled in using the pre-treatment data from the patient's first IVF cycle, which was available in the pre-treatment file. Missing

partner data was filled using the same partner information from earlier cycles. It was assumed that changes in partners were rare, so the partner situation for cycles in 2021 was assumed to be the same as for 2016-2020.

Total gonadotropin dose Furthermore, conflicting information was observed in the column *Total gonadotropin dose*, which describes the hormone dose given to the patients. The issue was that some patients in the data had received oocytes after an aspiration, despite not having received a gonadotropin dose. An aspiration always follows a hormone treatment, since the gonadotropin dose is essential for stimulating oocyte growth. As the hormone doses were not correct, these values were replaced with the actual dose the patients had received.

Stimulation protocol The raw data contained patients who underwent hormone stimulation prior to aspiration but who followed neither of the two established hormone protocols: the agonist or the antagonist. Instead, the protocol were labeled as "Stimulated" or "Unstimulated", without further specification. Upon consultation with the clinicians, it was clarified that these patients had been diagnosed with a condition called "hypopituitarism", a condition characterized by reduced hormone secretion from the anterior pituitary gland [30]. Rather than introducing "stimulated" or "unstimulated" as a distinct protocol category, it was decided to classify these patients under the antagonist protocol for consistency.

Planned fertilization method During the data merge, an issue came up with the variable *Planned fertilization method*, describing whether the oocytes and sperm would be fertilized using ICSI or standard IVF. The variable appeared in both the pre-treatment and post-treatment file, sometimes with differing values due to variations in the data from WinIVF and the Q-IVF script. To resolve this, it was decided to keep the *Planned fertilization method* from the pre-treatment file. For some patients, *Planned fertilization method* was set to 'Combination', indicating that both 'IVF' and 'ICSI' were planned. In these cases, 'Combination' was replaced with 'ICSI' as it is the most invasive method.

Partner age For IVF cycles involving donated sperm, the variable *Partner age* was replaced with a randomly generated number between 23 and 44 years. This approach was adopted because the age of the sperm donor—potentially a clinically relevant factor—was not available in the dataset. The imputed age replaced the value in the first cycle and for the second and third cycles the age was updated by adding the time elapsed since the first cycle.

Other imputations Missing values in the following columns were also imputed: *Infertility duration*, *Average menstrual cycle*, *Length* and *BMI*. For *Infertility duration*, missing values due to reasons such as same-sex couples were imputed with values between 6 and 36 months. Irregular cycles in *Average menstrual cycle* were imputed randomly between 35 and 90 days. Missing values in *Length* and *BMI* were imputed with the column mean.

Fertilization method For the column *Fertilization method for the transferred embryo*, missing values were primarily replaced using another column reflecting the most common fertilization method per cycle. Although this column was excluded as a predictor in the modeling due to a high proportion of missing values, its available entries remained useful for replacement. If unavailable, the *Planned fertilization method* was used as a fallback.

Consistency across cycles A few variables in the dataset are not consistently updated across cycles, including the columns: *BMI*, *Length*, *AFC right+left=total*, *Average menstrual cycle* and *Infertility duration*. In some cases, new values are recorded in later cycles, while in others they are not. To ensure consistency, it was decided to use the values from the first cycle for these columns across all cycles.

Table 3.4: Overview of data replacement and imputations.

Columns	Description
Pre-treatment variables	Missing pre-treatment data for cycles initiated during 2021 were filled with information from the first IVF cycle (2016–2020).
Total gonadotropin dose	Incorrect values were corrected after identifying a data issue.
Stimulation protocol	Protocols labeled as "Stimulated" were classified as "antagonist" for patients with hypopituitarism.
Planned fertilization method	Conflicts between pre- and post-treatment data were resolved by prioritizing the pre-treatment values. "Combination", indicating both IVF and ICSI were planned, was replaced with "ICSI" as it is the most invasive method.
Partner age	For all 147 patients who used donated sperm, the age was randomly imputed between 23 and 44 in the first cycle and updated in later cycles by adding the time elapsed since the first.
Infertility duration	Missing values for 44 patients were imputed with values between 6 and 36 months.
Average menstrual cycle	Missing values for 209 patients were imputed randomly between 35 and 90 days.
Length and BMI	Missing values were imputed for one patient with the column mean.
Fertilization method for the transferred embryo	Missing values were primarily filled using the column describing the most common method per cycle. If missing, the *Planned fertilization method* was used.

Table 3.4 – continued from previous page

Columns	Description
Consistency across cycles	The variables *BMI*, *Length*, *AFC right+left=total*, *Average menstrual cycle* and *Infertility duration* were not consistently updated across cycles. Values from the first cycle were used for these columns across all cycles.

3.5 Feature preparation

Before the data was prepared for modeling, a few new columns were created based on existing ones. Also, numerical columns were grouped into categorical values, categorical columns were combined into broader categories and the resulting categorical columns were one-hot encoded. An overview of the newly created columns based on existing ones is provided in table 3.5.

Good Quality Embryos The column *Good Quality Embryos* was constructed by summing the number of transferred embryos at fresh ET and the number of frozen ETs for each cycle. Although this information was already available in separate columns, a new combined variable was created under the assumption that a single aggregated value might serve as a more effective predictor.

Number of ETs and number of failed ETs The columns *Number of ETs* and *Number of failed ETs* were also created based on information available in the data. *Number of ETs* describes the total number of ETs per cycle, including ETs resulting in a live birth and was calculated by summing the number of ETs performed during each cycle.

The column *Number of failed ETs* represents a cumulative count of ETs that did not result in a live birth, that a patient brings into a new cycle. For example, if a patient undergoes two failed ETs during the first cycle and three during the second, the value of *Number of failed ETs* at the start of the third cycle would be five. Thus, this variable encodes the patient's accumulated history of unsuccessful transfers across previous treatment attempts. In contrast, the variable *number of ETs* is cycle-specific and denotes the number of embryo transfers performed within that particular cycle only. It is not cumulative and provides information solely about the current cycle. Why these variables were included is explained in section 4.1.

Average menstrual cycle The column *Average menstrual cycle, days* was also generated based on information available in the dataset. Originally, the menstrual cycle length was reported with a minimum and maximum value. To summarize these two variables, the average menstrual cycle length was calculated. If only one value was available (either the minimum or maximum), that value was used as the

average menstrual cycle length. If both of the values were missing, the average menstrual cycle length was imputed as described earlier.

The column *AFC right+left = total* was created by combining the AFC measurement for the left and right ovaries separately.

Table 3.5: Overview of created columns based on existing ones.

Created columns	**Description**
Good Quality Embryos	Sum of transferred embryos in fresh ET and frozen ETs per cycle.
Number of ETs	Total number of ETs per cycle.
Number of failed ETs	Cumulative number of ETs not resulting in a live birth.
Average menstrual cycle, days	Average of reported min and max cycle length; imputed if missing.
AFC right+left = total	Sum of AFC from both ovaries.

Categorical variables To enhance interpretability when evaluating key factors in the models, several numerical variables were transformed into categorical values. The following variables were converted:

- *BMI*
- *AFC right+left = total*
- *Average menstrual cycle, days*
- *Number of culture days at ET*
- *Number of embryos transferred at ET*

The categorization of *BMI* followed the World Health Organization (WHO) guidelines. For example, defining BMI ranges such as 25-29.9 could improve clarity compared to presenting precise values (e.g., 27.4). This approach could support more effective communication of results to clinicians and patients by making key variables more accessible and easier to interpret.

Combining subcategories Certain subcategories within the categorical columns, such as *Infertility cause* and *Sperm origin*, were also grouped together. It was observed that some subcategories represented less than 1% of the total entries. Including an excessive number of subcategories could lead to overfitting, while including too few may result in the loss of important information. For the column *Sperm origin*, the subcategories 'TESA', 'PESA' and 'Frozen-thawed TESA' were grouped under the category 'Surgical sperm extraction'.

Initially, the *Infertility cause* column contained 14 different subcategories. It was decided to combine the 'Egg factor' and 'Premature ovarian failure' subcategories into a single category, 'Egg factor', as both terms suggest that infertility is due to reduced egg quality. Additionally, the subcategories 'Genetic cause', 'Uterine

factor', 'Cancer' and 'Cervix factor' were grouped under the category 'Other'. The subcategories 'Same sex couple' and 'Single women' were merged into a new category, 'Same sex couple/single women'.

One-hot encoding All ten categorical columns were one-hot encoded resulting in 84 features in the dataset and 87 when using splines (see appendix B for a list of all features). This method allows models to treat each category as a distinct feature without assuming any ordinal relationship. For instance, columns like *Infertility cause* have no inherent order; assigning numerical values could lead the model to incorrectly interpret them as ordered, potentially causing biased predictions. One-hot encoding mitigates this risk by encoding each category as a separate feature.

4

Modeling

The key steps in the modeling phase included creating datasets based on the research questions, followed by training the models and assessing their performance. This also included designing the architecture of the training and testing process, as well as deciding which hyperparameters to tune during model optimization.

4.1 Datasets per research question

The modeling phase was initiated by defining the datasets for each of the three research questions. This task was mainly influenced by the discussion of how to handle patients who dropped out of the study before having a live birth or completing all three rounds of IVF treatment—an issue that could have been undertaken in several ways.

Dropouts To address dropouts, two models were trained for each research question: one excluding dropouts and another including them, assuming no live birth in the cycles they did not complete. Assigning a negative outcome (no live birth) introduces a known bias, but was considered more transparent and manageable than assigning outcomes randomly based on an assumed distribution, which would introduce an unpredictable and potentially greater bias.

Including dropout patients may lead to a "pessimistic" model, as some might have had a live birth if they had continued with more IVF cycles. Conversely, excluding these patients could result in an "optimistic" model, potentially overestimating the likelihood of a live birth. Table 4.1 provides an overview of the number of patients used to train the models per research question, with and without dropouts.

Table 4.1: Number of patients included in each dataset for the three research questions, both with and without dropout patients.

Research question	With dropouts	Without dropouts	Number of dropouts
Research question 1	1709	1348	361
Research question 2	1709	1103	606
Research question 3	850	605	245

Studies similar to this project have addressed dropouts differently. Ratna et al. used inverse propensity score weighting (IPW) to handle dropouts [10]. IPW is utilized by first fitting a model to predict the likelihood of patients dropping out and assume such patients have a lower probability of succeeding than patients who continued the study. That probability is then used to assign weights to each sample in the dataset. In this project, it was decided to not utilize IPW due to the high risk of generating an unpredictable bias. There are numerous reasons to why patients discontinue IVF treatment, such as the couple separating or mental strain. In this project, it was assumed that the available variables were insufficient to train a model for predicting such outcomes without introducing substantial bias.

Another approach to handling dropouts could involve first training a classifier on a dataset excluding dropouts, then using that model to predict outcomes for the patients who dropped out. A new model could then be trained on a dataset that includes these patients with their predicted outcomes. However, this method was not used in this project, as it was considered that relying on model predictions for dropout outcomes could introduce significant bias or misclassification, potentially leading to misleading results.

Features The variables included in the two datasets—one containing dropout patients and one excluding them—are largely identical, with the exception of two features: *number of ETs* and *number of failed ETs*. The dataset including dropouts contained the variable *number of ETs*, while the dataset excluding dropouts used *number of failed ETs*.

This distinction between the variables was made to account for patients who discontinued treatment before all available GQEs had been transferred. For example, a patient may have had four GQEs but discontinued treatment after only two embryo transfers, without achieving a live birth. In such cases, recording two failed ETs would incorrectly estimate the patient's outcomes, as one transfer could have been successful or up to four transfers could theoretically have failed. To address this limitation, the dataset including dropouts used *number of ETs* in place of *number of failed ETs*, which is only included in the dataset without dropouts.

As predictors of treatment outcome, both *number of ETs* and *number of failed ETs* are nuanced and potentially ambiguous. A high number of transfers may serve as a proxy for the availability of multiple GQEs, suggesting a favorable prognosis. However, it may also indicate that several transfer attempts were required without success, which could reflect a poorer prognosis. This shows the importance of considering context when evaluating how well these variables predict outcomes.

4.2 Model selection

After assessing the identified models in the literature review, the following ML models were considered most suitable for the Sahlgrenska dataset and were evaluated in this project: logistic regression, SVM and RF.

Motivation for logistic regression Logistic regression was chosen for its widespread use in clinical research, simplicity and interpretability. It allows for straightforward implementation and yields coefficients that clearly indicate the direction and strength of associations between predictors and outcomes—making it particularly valuable when transparency is important.

To explore the predictive value of different feature sets, logistic regression was evaluated in three versions. One model was trained using only the woman's age, which is a known strong predictor of live birth. This age-only model served as a baseline to estimate how much predictive performance could be attributed to age alone and how much additional value was gained by including more variables. Two additional models were trained on all available features, with one of them using RCS to flexibly model the non-linear effect of age.

While models like RF and SVM can handle non-linear patterns, logistic regression may fail to detect these unless adjustments like RCS are used, described in the section 2.2. Although, these adjustments increase the risk of overfitting, where the model may perform well on training data but fail to generalize to new data. This highlights the importance of considering alternative models when non-linearity is expected. Similarly to McLernon et al. (2022) in [9], five reference points were used in this project for the RCS (also known as knots). The selected age points were: $t_1 = 25$, $t_2 = 28$, $t_3 = 31$, $t_4 = 34$, and $t_5 = 37$. As a result, three additional variables were included in the dataset alongside the original variable representing the women's age. These additional variables transform age to capture non-linear relationships with the outcome. The age points $t_1 - t_5$ allow the relationship to change direction between these points, providing flexibility in modeling, while maintaining linearity at the extremes (the youngest and oldest ages), where data is sparse.

Including both unadjusted and spline-adjusted logistic regression models, trained on all features, allowed for a more balanced comparison with non-linear models like RF. This approach ensured that logistic regression was not unfairly disadvantaged and helped clarify whether the increased complexity of tree-based models was justified. Additionally, using splines aligns with earlier research, making comparisons across studies easier.

Motivation for SVM SVM was considered for this project due to its suitability for small datasets like the Sahlgrenska dataset. The key principle of SVM, i.e. maximizing the margin between distinct classes, helps identify the most critical data points near the decision boundary [2]. By prioritizing these critical points, SVM is less sensitive to noise and outliers, which enhances the model's accuracy and robustness, even when data is limited.

Motivation for RF While advanced models like boosting algorithms (e.g. XGBoost, LightGBM and AdaBoost) have shown promise in IVF outcome prediction, RF was considered to be a better fit for the dataset in this project. Boosting algorithms build trees sequentially, meaning the trees have strong correlations [17]. In contrast, RF build trees in parallel and reduces variance through voting (for classification tasks), making it more robust to overfitting.

Moreover, boosting methods tend to perform better on large datasets with many variables. Given our smaller dataset, RF is more appropriate. It also offers better interpretability through feature importance scores, which is especially valuable in a clinical context where clinicians need to trust the model's decisions. While XGBoost may achieve higher predictive performance on larger datasets, it requires more tuning and computational resources. In contrast, RF strikes a good balance between performance and ease of use, making it a practical choice for the Sahlgrenska dataset.

Regarding DT, while they are simpler to interpret, they generalize worse than RFs [17]. The ensemble approach of the RF mitigates this issue, making it a more reliable choice.

4.3 Training and evaluation

The process of training and evaluating the models is illustrated in figure 4.1. The modeling procedure was repeated over five iterations. In each iteration, the data was split into training (70%), validation (20%) and test (10%) sets. This ratio was chosen to balance model learning and evaluation, given the relatively small dataset. If the training set is too small, the model may not capture underlying patterns effectively, resulting in poor generalization to unseen data. Conversely, a test set that is too small or unrepresentative may yield unreliable performance estimates. The 70/20/10 split was therefore used to address both concerns.

Feature scaling Within each iteration, feature scaling was applied based on the specific requirements of the ML model. As shown in prior literature, RF performs best without scaling, SVM benefits from normalization and logistic regression performs best with standardization [20]. These approaches were used in this project.

Training and calibration The models were subsequently trained using the training dataset. During training, hyperparameter tuning was conducted using exhaustive search and cross-validation with scikit-learn's GridSearchCV, which searches over a predefined hyperparameter space to identify the optimal configuration for each model. The tested hyperparameter values are listed in the table C.1. Following training, model calibration was performed using the validation set to improve the reliability of predicted probabilities. Isotonic regression performed better than platt scaling and was therefore used in this project. The calibrated models were then evaluated on the held-out test dataset, providing an unbiased estimate of the model performance.

Evaluation metrics The evaluated metrics were: Accuracy, sensitivity, specificity, brier score, MCC and AUROC with standard deviation. These were chosen to provide a broad assessment of the model's performance. AUROC was primarily chosen because it provides a robust measure of the model's ability to distinguish between couples who will have a live birth and those who will not [24]. AUROC was also selected to simplify the comparison with other studies, such as those by McLernon et al. and Ratna et al. [3], [9], [10], which used the concordance index

(C-index), a metric equivalent to AUROC in the context of this study. Additional metrics were included to provide a more comprehensive evaluation of model performance. For example, a limitation of AUROC is that it does not consider how common or rare the positive class is (prevalence) or the impacts of making different types of mistakes, such as false positives and false negatives [24]. Therefore, the metrics accuracy, sensitivity, specificity and MCC were used in this project to focus on different aspects of the model's ability to correctly classify the probability of achieving a live birth. The Brier score was selected to offer a direct measure of the model's reliability in producing accurate probability estimates [26].

To ensure robustness and reduce variability due to random splits, the entire pipeline, i.e. data splitting, scaling, training, calibration and evaluation, was repeated across all five data partitions. The average model performance across these iterations on the test set was then computed to obtain a more generalizable assessment. SHAP figures were created for the models with the highest average test AUROC, both with and without dropouts, to highlight the model's reasoning and key features.

When calculating the metrics, the threshold was set to 0.5, meaning that predicted probabilities below 0.5 were classified as no live birth, and those equal to or above 0.5 as live birth. This value is commonly used as threshold and not necessarily optimal for maximizing performance metrics such as sensitivity, specificity or MCC. Adjusting the threshold, e.g. lowering it to 0.3, could potentially increase sensitivity at the expense of specificity, as more patients would be classified as achieving a live birth, and vice versa. One common method for selecting a more balanced threshold is the Youden Index, which identifies the point that maximizes the sum of sensitivity and specificity [31]. The Youden Index was not applied in this project; however, it could be considered for future work.

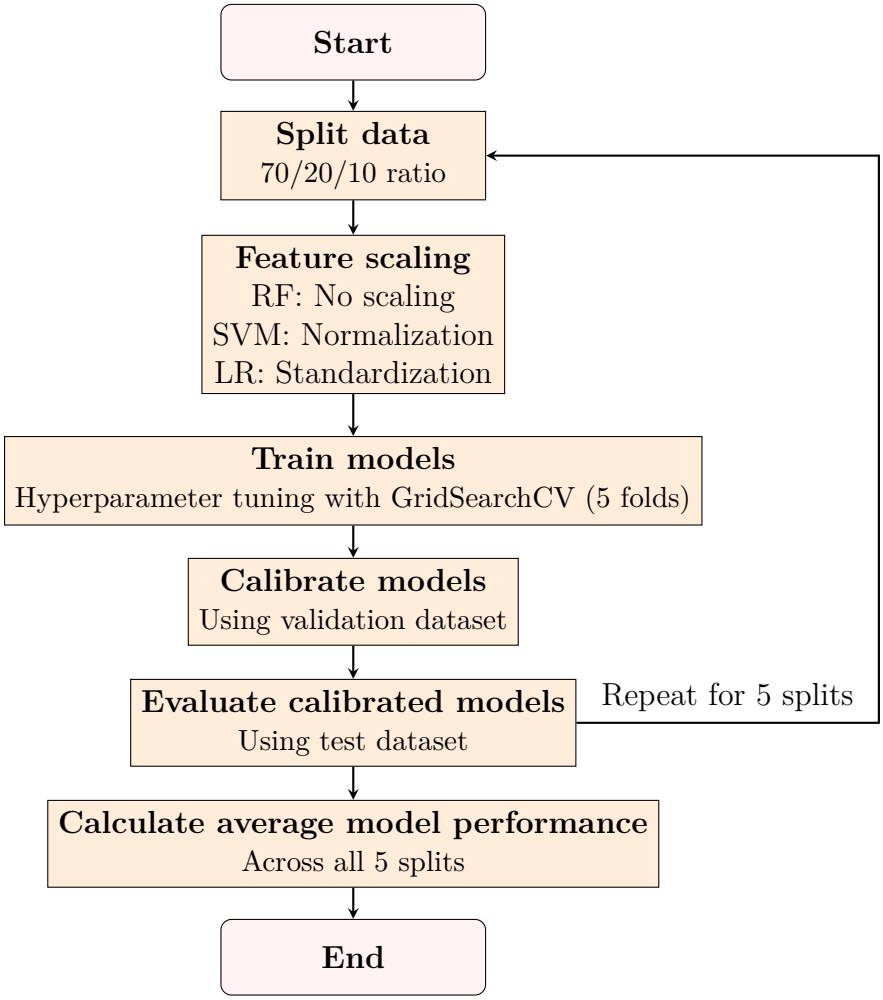


Figure 4.1: A flowchart outlining the modeling phase starts with data splitting (70/20/10 ratio) and feature scaling. Models were trained with hyperparameter tuning using exhaustive search and cross-validation via scikit-learn's GridSearchCV, then calibrated using a validation dataset. The calibrated models were evaluated on the test dataset and average performance was calculated across five splits.

Hyperparameters To optimize model performance, the most influential hyperparameters were prioritized during GridSearchCV, as shown in table 4.2. All other hyperparameters were kept at their default values to minimize computational time.

Table 4.2: List of hyperparameters and their respective values used in GridSearchCV for each ML model: RF, SVM, and logistic regression (LR).

Model	Hyperparameter	Values
RF	`max_depth`	3, 5, 10, None
	`min_samples_leaf`	1, 5, 10, 20
	`min_samples_split`	2, 4, 6
	`min_weight_fraction_leaf`	0, 0.01, 0.05
	`class_weight`	balanced, None
SVM	`C`	0.01, 0.1, 1, 10
	`class_weight`	balanced, None
LR	`penalty`	l1
	`C`	0.01, 0.1, 1, 10
	`solver`	liblinear
	`class_weight`	balanced, None

For the RF model, the hyperparameters `max_depth` (maximum distance between the root and any leaf in each DT), `min_samples_split` (minimum samples required to split a node) and `min_samples_leaf` (minimum samples required in a leaf node) were selected for tuning due to their critical role in controlling model complexity and preventing overfitting, which is particularly relevant given the relatively small dataset and the high-dimensional feature space (more than 80 variables). Deeper trees (e.g., with `max_depth=None`) can capture complex patterns but risk memorizing noise in small datasets, especially when many features are involved. Therefore, a range of values (3, 5, 10, None) were tested. In addition, increasing `min_samples_split` and `min_samples_leaf` helps regularize the model by preventing splits based on very small sample sizes, i.e. splits that could reflect random fluctuations rather than true patterns.

`class_weight` (weights given to each class in the dataset) was evaluated to determine if adjusting for class imbalance would improve model performance. Testing both `None` and `balanced` (weights are inversely proportional to the class frequencies) allowed for evaluating whether giving more weight to the minority class, i.e. patients not getting a live birth, led to better generalization. Additionally, `min_weight_fraction_leaf` was included to introduce another form of regularization when sample weights are used (e.g., from `class_weight=balanced`). While `min_samples_leaf` limits leaf size by the number of samples,
`min_weight_fraction_leaf` constrains it based on the total weight of the samples, providing more control when classes are imbalanced.

For the SVM model, the hyperparameters `C` and `class_weight` were selected for tuning. The regularization parameter `C` influences the complexity of the decision

boundary, which in this case separates patients who experience a live birth from those who do not. A smaller `C` value allows for a simpler decision boundary, helping to prevent overfitting, while a larger `C` value leads to a more complex boundary that may fit the training data more closely.

Additionally, four hyperparameters were tested for the two logistic regression models trained on all variables, referred to as LR in the table: `penalty`, `C`, `solver` and `class_weight`. The `penalty` was set to "l1" (lasso regularization), which was chosen over the default "l2" (ridge regularization) due to its ability to produce sparse models with more zero coefficients, having minimal impact on the predictions. Given the dimensionality of the dataset, `l1` regularization was assumed to be optimal, as it also enhances model interpretability. The `solver` was set to "liblinear" since it is the only solver compatible with "l1" regularization.

5

Results

This section presents the performance of the evaluated models in relation to the three research questions outlined in section 1.3. For each research question, a summary of performance metrics is reported for all models on the test set. Additionally, SHAP plots are provided to highlight feature importance for each research question. A calibration plot is included for the model that achieved the highest average test AUROC across all research questions.

The terms "pessimistic model" and "optimistic model" are used to describe models with and without dropouts, respectively. A positive instance refers to a successful IVF cycle, whereas a negative instance refers to an unsuccessful IVF cycle.

5.1 Research question 1

The models for the first research question predicted the likelihood of a live birth in the second IVF cycle after an unsuccessful first cycle. Among patients who continued to the second cycle, 36.9% had a live birth, while the rate dropped to 29.1% when including dropouts.

Table 5.1 shows the average evaluation metric values across all five cross-validation splits for each model, with and without dropouts. Arrows next to each metric indicate whether higher or lower values reflect better performance, with standard deviation denoted as std. The model LR (age only) refers to a logistic regression model trained solely on the woman's age. The other models were trained on all available features. LR denotes the standard logistic regression model, while LR (age adjusted) refers to a logistic regression model that used RCS to account for the non-linear relationship between the woman's age and the outcome variable. The column "Dropouts" indicates whether they were included or excluded.

Table 5.1: Evaluation summary of models for research question 1. LR (age only) refers to logistic regression using only age as a feature. All other models were trained on all available features. LR denotes standard logistic regression, while LR (age adjusted) used RCS. Metrics represent average values across five splits, with standard deviation denoted as std. Arrows indicate whether higher or lower values are preferred. The "Dropouts" column shows whether dropouts were included.

Model	Dropouts	Accuracy (↑)	Sensitivity (↑)	Specificity (↑)	Brier score (↓)	MCC (↑)	AUROC (std) ((↑)
LR (age only)	Excluded	0.65	0.07	0.95	0.22	0.01	0.601 (0.056)
LR (age only)	Included	**0.69**	0.00	**1.00**	0.21	0.00	0.618 (0.042)
LR	Excluded	0.65	0.18	0.89	0.22	0.03	0.635 (0.042)
LR	Included	**0.69**	0.11	0.94	**0.20**	0.05	0.666 (0.029)
LR (age adjusted)	Excluded	0.64	0.15	0.91	0.22	0.06	0.63 (0.056)
LR (age adjusted)	Included	**0.69**	0.14	0.93	**0.20**	0.11	**0.667 (0.039)**
RF	Excluded	0.60	**0.24**	0.82	0.22	0.11	0.617 (0.039)
RF	Included	**0.69**	0.04	0.97	**0.20**	0.09	0.636 (0.032)
SVM	Excluded	0.64	0.20	0.88	0.22	**0.13**	0.629 (0.025)
SVM	Included	0.68	0.03	0.98	0.21	0.02	0.584 (0.042)

Overall, the performance differences between models with and without dropouts were small in terms of AUROC. This is reflected in overlapping standard deviations of AUROC scores, suggesting comparable model behavior. Logistic regression adjusted for the women's age using RCS, i.e. LR (age adjusted), achieved the highest AUROC of 0.667 (std = 0.039) when dropouts were included. Without dropouts, the best-performing model was logistic regression without age adjustment, marked as LR in the table, with an AUROC of 0.635 (std = 0.042).

Both when including and excluding dropouts, the logistic regression model trained only on the woman's age performed slightly worse than the logistic regression models trained on all available features. When dropouts were included and age was the only feature, the logistic regression model achieved a sensitivity of 0.0 and a specificity of 1.0. A sensitivity of 0.0 means that no patient who experienced a live birth was predicted to experience a live birth.

The accuracy ranged from 0.60 (without dropouts) to 0.69 (with dropouts), while the sensitivity varied from 0.00 to 0.24. Including dropouts resulted in lower sensitivity, reflecting the study design of a pessimistic and optimistic model. The study design was also evident in the specificity difference, which ranged from 0.82 to 1.00. A specificity of 1.0 means all patients who did not have a live birth were correctly identified.

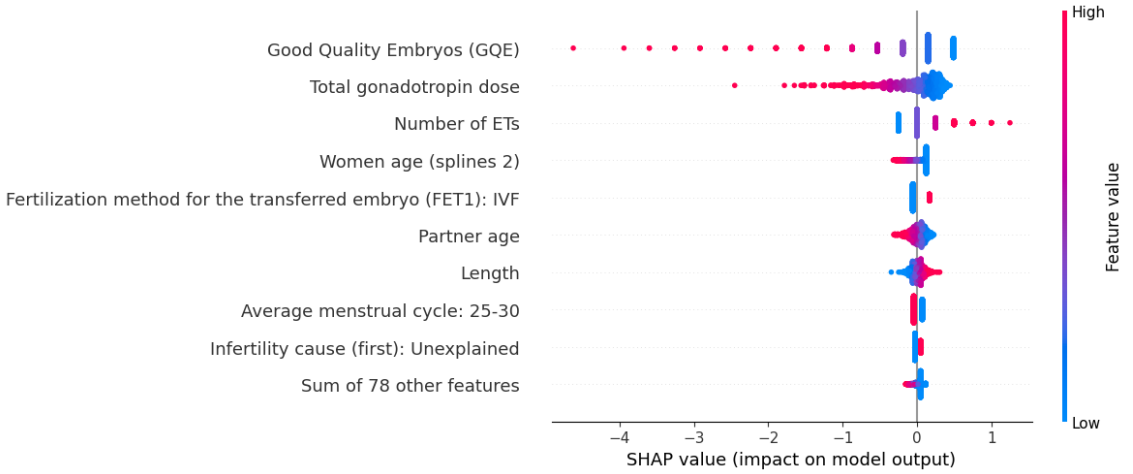
Brier scores ranged from 0.20 to 0.22, with a lower score when dropouts are included, reflecting the higher proportion of negative cases in the pessimistic model. A Brier score of 0 indicates perfect alignment between predicted probabilities and actual outcomes.

As seen in the table, MCC scores were low (0.00 to 0.13), despite AUROC ranging from 0.60 to 0.67. In theory, MCC scores range from -1 to 1, with 0 indicating random predictions and negative values showing complete incorrect results. While AUROC measures how well the model ranks positives over negatives, MCC is based

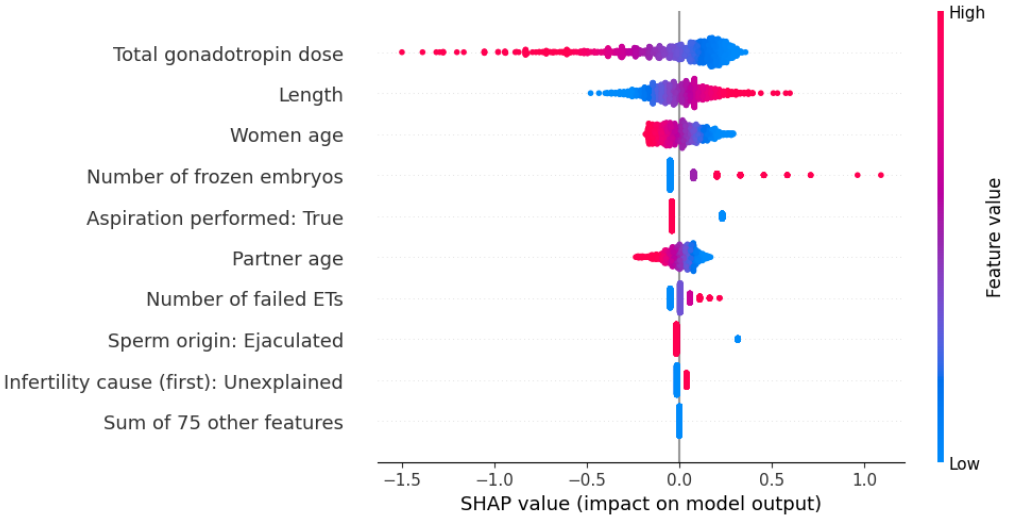
on actual predictions at a 0.5 threshold. This means that even models with high AUROC can yield low MCC scores if they misclassify many cases at that threshold.

Figure 5.1 shows SHAP plots for the models with the highest average test AUROC score for the optimistic and pessimistic models, respectively. The SHAP values for the top nine features with the largest impact on the predicted probability are displayed, together with the SHAP values for the remaining features. Figure 5.1a shows SHAP values for the model that achieved the best result when dropouts were included, meaning LR (age adjusted). In figure 5.1b, SHAP values are visualized for LR, which performed best when dropouts were excluded.

Each dot in the SHAP figure represents the SHAP value for one patient. The position of the dot relative to the vertical line (SHAP value = 0) indicates its contribution to the prediction: dots to the left reduce the likelihood of a live birth, while those to the right increase it. If many patients have the same SHAP value, the dots will stack vertically. Dot color represents the value of the feature for that patient: red indicates a high value and blue indicates a low value. By examining the color distribution on each side of the vertical line, it is possible to see how high or low feature values influence the prediction. For instance, in figure 5.1a, the feature *Good Quality Embryos* shows many red dots on the left, suggesting that patients achieving more GQE are associated with a lower probability of live birth. Ideally, red and blue dots should be clearly separated, indicating that low and high feature values consistently influence the prediction in opposite directions.



(a) Logistic regression (adjusted for women's age) with dropouts included.



(b) Logistic regression with dropouts excluded.

Figure 5.1: SHAP plots for the top nine features in the models with the highest average AUROC for research question 1. Dropouts are included in subplot 5.1a and excluded in subplot 5.1b. Each dot represents a patient's contribution, with its position showing whether the feature increases or decreases the likelihood of a live birth. Dots to the left reduce and dots to the right increase the probability. Red dots indicate high values, and blue dots indicate low values.

As shown in figure 5.1a, the top three features for logistic regression (adjusted for women's age) with dropouts included were *Good Quality Embryos*, *Total gonadotropin dose* and *Number of ETs*. A high number of GQE and a high total gonadotropin dose (indicated by red points) were associated with a lower chance of live birth. In contrast, a higher number of ETs was associated with an increased chance of live birth.

For logistic regression trained on all available features with dropouts excluded, the

top three most important features were *Total gonadotropin dose*, *Length* and *Women age*. A high total gonadotropin dose and older age were associated with a lower chance of achieving a live birth. In contrast, greater length was associated with a higher chance of live birth.

5.2 Research question 2

The models for the second research question predicted the aggregated chance of a live birth across a patient's second and third IVF cycle, given an unsuccessful first cycle. When excluding dropouts, 60.7% had a child. When including those who dropped out after the first and second cycles, the live birth rate decreased to 39.1%.

Table 5.2 presents the average evaluation metric values across all five cross-validation splits for each model, with and without dropouts. Arrows indicate whether higher or lower metric values reflect better performance. Standard deviation is noted as std. LR (age only) used only the woman's age as input. All other models used all available features. LR denotes standard logistic regression and LR (age adjusted) applied RCS to model non-linear age effects. The "Dropouts" column specifies whether dropouts were included or excluded.

Table 5.2: Evaluation summary of models for research question 2. LR (age only) refers to logistic regression using only age as a feature. All other models were trained on all available features. LR denotes standard logistic regression, while LR (age adjusted) used RCS. Metrics represent average values across five splits, with standard deviation denoted as std. Arrows indicate whether higher or lower values are preferred. The "Dropouts" column shows whether dropouts were included.

Model	Dropouts	Accuracy (↑)	Sensitivity (↑)	Specificity (↑)	Brier score (↓)	MCC (↑)	AUROC (std) ((↑)
LR (age only)	Excluded	0.66	**0.87**	0.35	0.23	0.26	0.654 (0.052)
LR (age only)	Included	0.61	0.17	**0.91**	0.23	0.11	0.632 (0.04)
LR	Excluded	**0.67**	0.84	0.40	0.23	0.28	0.668 (0.048)
LR	Included	0.62	0.35	0.81	0.23	0.19	0.664 (0.023)
LR (age adjusted)	Excluded	0.65	**0.87**	0.36	0.23	0.27	0.672 (0.045)
LR (age adjusted)	Included	0.64	0.35	0.82	**0.22**	0.24	0.667 (0.024)
RF	Excluded	**0.67**	0.84	0.37	0.23	**0.29**	**0.675 (0.055)**
RF	Included	0.64	0.38	0.81	0.23	0.23	**0.675 (0.025)**
SVM	Excluded	0.64	0.83	0.34	0.24	0.21	0.627 (0.043)
SVM	Included	0.62	0.21	0.87	0.23	0.14	0.624 (0.018)

The RF model achieved the highest AUROC of 0.675 (std = 0.025) when dropouts were included. Without dropouts, RF remained the best-performing model with an AUROC of 0.675 (std = 0.055). However, the AUROC scores overlapped for most models, indicating similar performance across models.

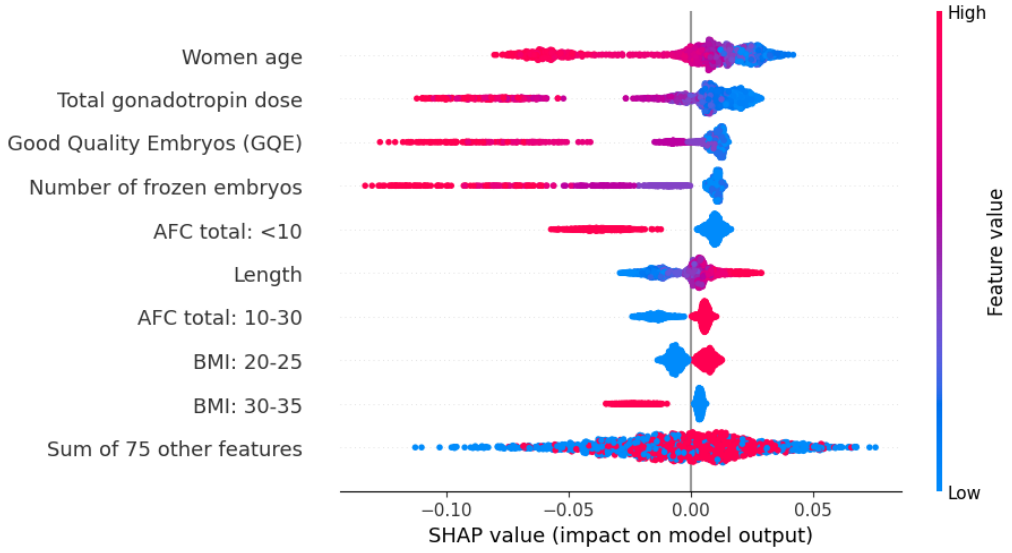
Logistic regression using age as the only feature achieved a comparable AUROC to the other models. However, it showed lower sensitivity when dropouts were included, meaning it correctly identified fewer patients who would experience a live birth compared to models trained on all available features.

As in the first research question, performance improved when dropouts were excluded. Sensitivity across all models was around 0.8 without dropouts, compared to 0.17–0.38 when dropouts were included. A sensitivity of 0.8 means that the models correctly identified 80% of the patients who actually had a live birth.

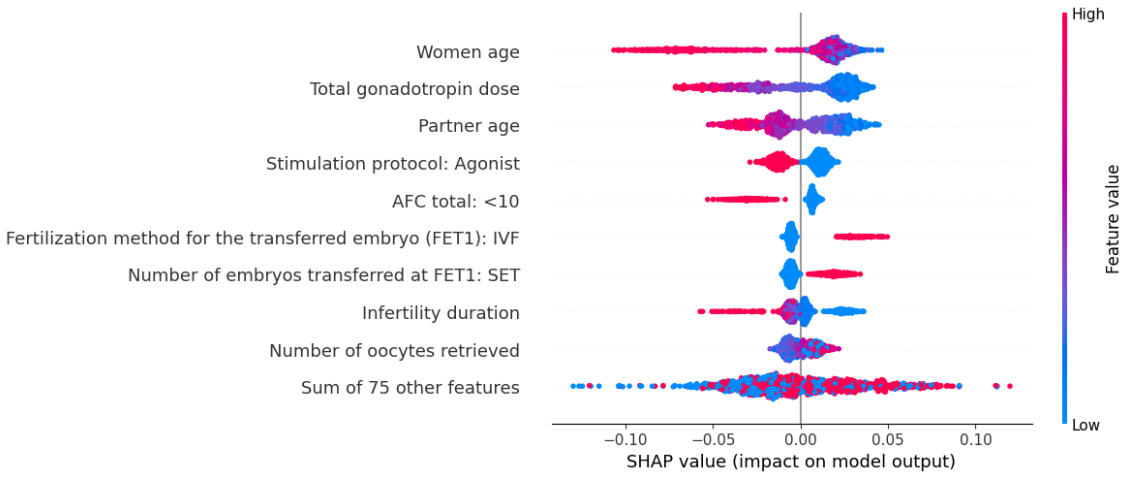
The accuracy ranged between 0.67 (without dropouts) and 0.61 (with dropouts). The Brier score remained nearly the same across all models, with values around 0.2. MCC was also slightly higher without dropouts, with values ranging between 0.21-0.29, compared to 0.11-0.24 when dropouts were included.

SHAP plots for the models with the highest average test AUROC, both with and without dropouts, are shown in figure 5.2. Subplot 5.2a displays the SHAP values for the RF model with dropouts included, while subplot 5.2b presents the SHAP values for RF with dropouts excluded.

As shown in figure 5.2a, the top three features for RF when dropouts were included were *Woman's age*, *Total gonadotropin dose* and *Good Quality Embryos*. High values of these features (indicated by red points) were associated with a lower chance of a live birth. For RF with dropouts excluded, the top three most important features were *Woman's age*, *Total gonadotropin dose* and *Partner's age*. High values of these features corresponded to a lower chance of a live birth. Notably, as in the first research question, *Good Quality Embryos* appeared as an important predictor only when dropouts were included.



(a) RF with dropouts included.



(b) RF with dropouts excluded.

Figure 5.2: SHAP plots for the nine most important features in the models with the highest average AUROC for research question 2. Dropouts are included in subplot 5.2a and excluded in subplot 5.2b. Each dot represents a patient's contribution, with its position indicating whether the feature increases or decreases the likelihood of a live birth. Dots to the left reduce and dots to the right increase the probability. Red dots represent high values, while blue dots indicate low values.

5.3 Research question 3

The models for the third research question predicted the likelihood of a live birth in a patient's third IVF cycle, given an unsuccessful first and second cycle. When excluding dropouts, 28.3% of patients who started the third cycle experienced a live birth. When including dropouts from the second cycle, the rate was 20.1%.

Table 5.3 presents the average evaluation metric values across all five cross-validation

splits for each model, with and without dropouts. Arrows indicate whether higher or lower metric values reflect better performance. Standard deviation is noted as std. LR (age only) used only the woman's age as input. All other models used all available features. LR denotes standard logistic regression and LR (age adjusted) used RCS to model non-linear age effects. The "Dropouts" column specifies whether dropouts were included or excluded.

Table 5.3: Evaluation summary of models for research question 3. LR (age only) refers to logistic regression using only age as a feature. All other models were trained on all available features. LR denotes standard logistic regression, while LR (age adjusted) used RCS. Metrics represent average values across five splits, with standard deviation denoted as std. Arrows indicate whether higher or lower values are preferred. The "Dropouts" column shows whether dropouts were included.

Model	Dropouts	Accuracy (↑)	Sensitivity (↑)	Specificity (↑)	Brier score (↓)	MCC (↑)	AUROC (std) ((↑)
LR (age only)	Excluded	0.70	0.01	0.99	0.20	-0.00	0.600 (0.108)
LR (age only)	Included	**0.84**	0.00	**1.00**	**0.13**	0.00	**0.619 (0.061)**
LR	Excluded	0.69	0.05	0.96	0.22	0.02	0.534 (0.046)
LR	Included	0.83	0.00	0.99	0.14	-0.04	0.614 (0.09)
LR (age adjusted)	Excluded	0.67	**0.06**	0.92	0.23	-0.04	0.519 (0.05)
LR (age adjusted)	Included	0.83	0.00	0.98	0.14	-0.04	0.612 (0.093)
RF	Excluded	0.70	0.05	0.95	0.21	**0.07**	0.590 (0.059)
RF	Included	0.83	0.02	0.99	0.14	-0.02	0.617 (0.136)
SVM	Excluded	0.70	0.03	0.97	0.22	-0.01	0.586 (0.052)
SVM	Included	0.82	0.04	0.96	0.15	-0.01	0.551 (0.073)

All models achieved similar performance metrics. The highest mean AUROC was achieved by the logistic regression model using age as sole feature, both with and without dropouts. When dropouts were included, the AUROC was 0.619 (std = 0.061) and without dropouts, the AUROC was 0.600 (std = 0.108). However, as in the previous research questions, several models showed overlapping AUROC values, indicating comparable performance.

In the table, sensitivity, reflecting the model's ability to correctly identify positive cases, was very low across all models ranging from 0.00 to 0.06. In contrast, specificity, indicating the model's ability to correctly identify negative cases, was much higher, ranging from 0.92 to 1.00. This shows that the models struggled at separating between the two classes.

For example, the logistic regression model trained only on the woman's age achieved a sensitivity of 0.0 and a specificity of 1.0 when dropouts were included. In this case, the model predicted that no women would achieve a live birth. Since the actual live birth rate in this group was 20.1%, this entire portion of the dataset was misclassified as negative. The model's accuracy of 0.84 (averaged across all five splits) can be explained by its correct identification of the 79.9% who did not have a live birth, despite failing to predict any of the positive outcomes. This illustrates how the model can appear to perform well in terms of accuracy while completely failing to detect the positive class.

The MCC scores ranged from -0.04 to 0.07, showing that the models performed no better than random guessing. This occurred because the models misclassified many

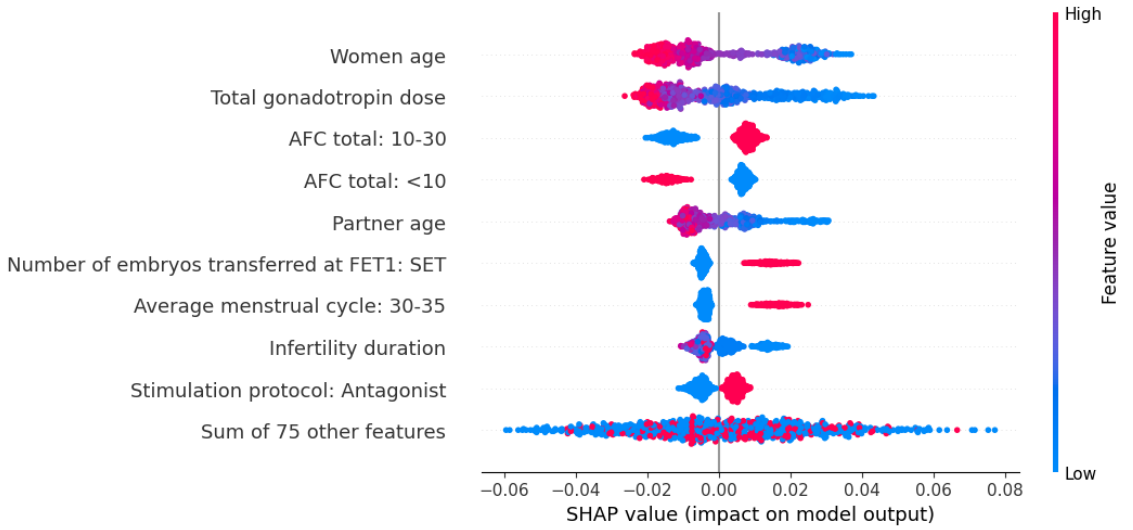
positive cases (patients who achieved a live birth) as negative, leading to a low MCC. When a model misses most positive cases, it reduces true positives (correctly predicted positive cases) and increases false negatives (positive cases incorrectly predicted as negative). Since the MCC is sensitive to these imbalances, the result is a score near zero, indicating that the models were ineffective in detecting the positive class, despite correctly classifying some negative cases.

The AUROC of approximately 0.6 across all models indicates that they could somewhat differentiate between positive and negative cases, despite low sensitivity. This suggests the model assigned higher prediction scores to positive cases than negative ones, resulting in a score above random guessing (0.5). This highlights the importance of considering AUROC alongside other metrics like sensitivity, specificity and MCC for a complete evaluation of model performance.

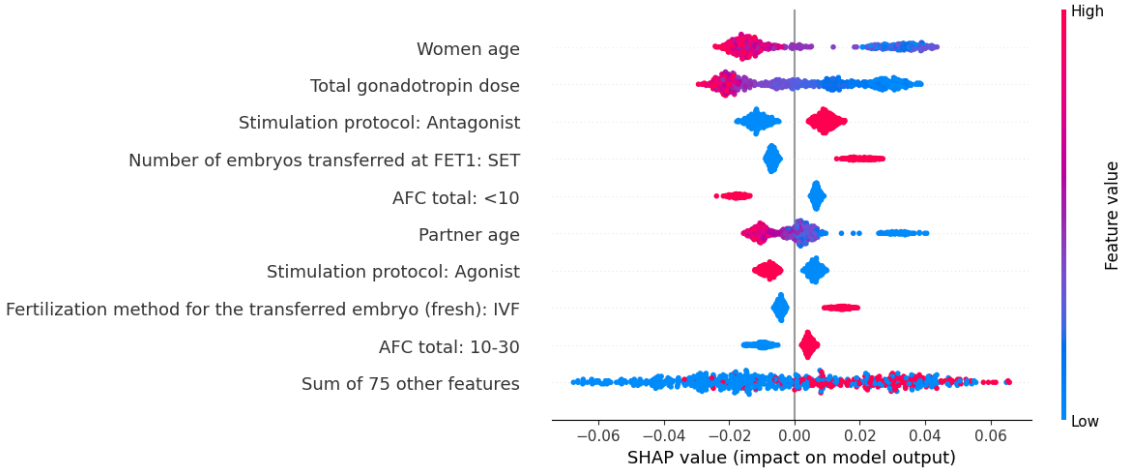
Although the logistic regression model trained on women's age achieved the highest average test AUROC, the focus is on understanding how the other variables influenced the outcome. Therefore, SHAP values for the next highest-performing model, which achieved the second-highest average test AUROC, are shown in figure 5.3. Subplot 5.3a presents SHAP values for RF with dropouts included, while subplot 5.3b displays SHAP values for RF with dropouts excluded.

As shown in figure 5.3a, the top three features were *Woman's age*, *Total gonadotropin dose* and *AFC total: 10-30*. A high value of woman's age and total gonadotropin dose (indicated by red points) were associated with a lower chance of live birth. However, patients with a total AFC of 10-30 were associated with a higher chance of live birth.

For RF with dropouts excluded, the top three most important features were *Woman's age*, *Total gonadotropin dose* and *Simulation protocol: Antagonist*. A high value of woman's age and total gonadotropin dose were associated with a lower chance of live birth. In contrast, for patients who received the antagonist stimulation protocol, it was associated with a higher chance of live birth.



(a) RF with dropouts included.



(b) RF with dropouts excluded.

Figure 5.3: SHAP plots for the top nine features of the RF models, with and without dropouts included, addressing research question 3. The plot for RF with dropouts included is shown in subplot 5.3a, while the plot for RF with dropouts excluded is shown in subplot 5.3b. Each dot represents a patient's contribution, with its position showing if the feature increases or decreases the likelihood of a live birth. Dots to the left decrease and dots to the right increase the probability. Red dots indicate high values, and blue dots indicate low values.

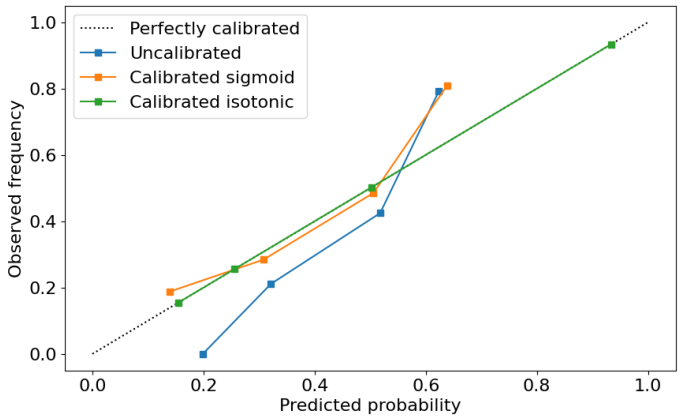
5.4 Calibration

This section presents calibration plots for the model that achieved the highest average test AUROC across all research questions: the RF model with dropouts included for research question 2. Subplots 5.4a and 5.4b show calibration plots for two different data splits. In both plots, the uncalibrated model is represented by the blue

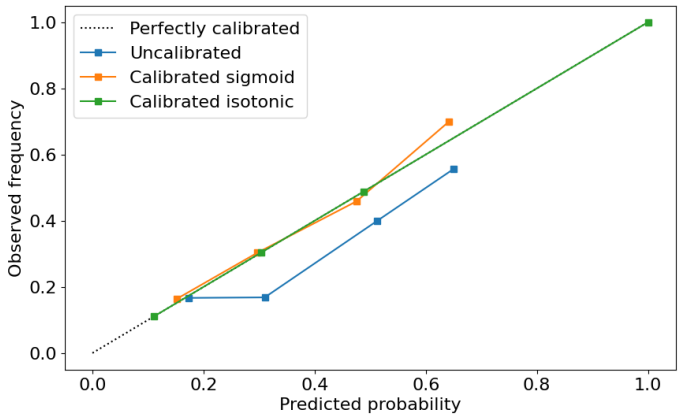
line, the sigmoid-calibrated model by the orange curve and the isotonic regression-calibrated model by the green curve.

These calibration plots highlight the effectiveness of isotonic regression, as the green line closely follows the diagonal (indicating perfect calibration). The inclusion of calibration plots for two data splits demonstrates slight variations in the ratio between predicted and observed probabilities across the splits. Nevertheless, isotonic regression consistently performs best, with the green curve closely aligning with the diagonal in all cases. This motivates the choice of isotonic regression as the used calibration method in this project.

The decision was made to display only two calibration plots instead of all five from the different data splits. This choice was based on the observation that the trend remained consistent across all models, regardless of the ML model used or whether dropouts were included. Of the five calibration plots for the pessimistic RF model, the two presented here showed the most variation.



(a) Calibration plot for data split 1 of the RF model with dropouts included.



(b) Calibration plot for data split 2 of the RF model with dropouts included.

Figure 5.4: Calibration plots for two different data splits, comparing the model with the highest average test AUROC (i.e., RF with dropouts included) for research question 2. In the plots, the uncalibrated model is represented by the blue line, the calibrated model using sigmoid is shown by the orange curve, and the calibrated model using isotonic regression is represented by the green curve.

6

Discussion

The ML models developed in this project achieved AUROC scores comparable to existing studies, which highlights the impact of a small and complete dataset. Investigations of feature importance revealed a potential distribution shift between patients who dropped out after their first cycle and patients who continued to a second cycle. However, the performance metrics points to IVF success as a difficult classification problem.

This section begins with a reflection of the model performance, followed by an analysis of feature importance. Considerations related to longitudinal data and distribution shifts are then discussed, along with potential challenges and opportunities for clinical implementation. Finally, suggestions for future improvements are presented.

6.1 Evaluation of model performance

Comparison with similar studies Overall, the predictive performance of the models developed in this project is comparable to existing studies. Similar to research question 2 in this project, both Ratna et al. (2022) [10] and McLernon et al. (2022) [9] tried to predict the probability of achieving a live birth across the second and third IVF cycle. In this project, the RF models reached AUROC scores of 0.675. Compared to the model developed by Ratna, which achieved a C-index of 0.65, this project demonstrates a similar level of performance. An AUROC ranging from approximately 0.60 to 0.67 was obtained across all three research questions, aligning closely with the C-index reported by Ratna. The model developed by McLernon et al. in 2022 achieved a slightly higher C-index of 0.71 [9].

The datasets used by McLernon and Ratna are much larger compared to ours, with data from 80 000+ and 40 000+ patients, respectively [9], [10]. The fact that our model achieved a similar score as a model trained on a much larger dataset is encouraging and gives credit to the validity of our model. This comparison highlights that a non-parametric model, such as RF, can achieve similar results without the need for methods like restricted cubic splines to address non-linear relationships between the variables and the outcome. This is an advantage, as it helps avoid overfitting and reduces the need for excessive model tuning.

However, when comparing model performance across different studies, it is impor-

tant to consider the key differences in study populations. Women who previously had a live birth were included in all the studies by McLernon and Ratna, whereas they were excluded in ours [3], [9], [10]. It is possible such patients carry different characteristics than patients who have not yet achieved a live birth, regardless of whether the birth was a result of an IVF treatment or not. Hence, the models developed by Ratna and McLernon might have found patterns in the data that does not exist in the Sahlgrenska dataset.

The model developed by Barnett-Itzhaki et al. reached an AUROC of 0.77 [11]. However, their dataset was much smaller than ours, with 136 women who underwent their first to fifth cycle of IVF. Blank et al. did also reach higher model performance compared to us, with an AUROC of 0.74 [12]. The size of their dataset was comparable to ours, with 1,052 unique women compared to our 3,217. However, the research goal of their study was to predict the outcome after one single ET where the embryo had been cultivated for five days. Similar to Blank, Dehghan et al. predicted outcomes only after a patient's first ET, meaning all FETs were excluded [14]. Liu et al. reached an AUROC of 0.845 by predicting IVF outcomes after the patient's first round of IVF [13].

Although the studies discussed in the previous section are not directly comparable to our own due to methodological and contextual differences, they nonetheless offer valuable insights into how ML has been applied to predict IVF outcomes. These studies focused on specific subsets of patients, such as only the first IVF cycle or a single ET, which may limit generalizability to broader patient populations. In contrast, our model was designed to account for multiple cycles and include both fresh and frozen embryo transfers, capturing a wider range of clinical scenarios. This broader inclusion may partly explain the slightly lower AUROC, as it introduces greater variability.

Metrics of performance The AUROC scores for all models and all research questions showed similar performance with small variations. Some models performed worse than others, such as the pessimistic version of SVM for research question 3 (which reached an AUROC score of 0.551), compared to the RF models for research question 2 (both of which reached an AUROC score of 0.675). However, no model can be said to successfully enough separate between patients who experienced a live birth and patients who did not. One possible explanation is that the available information (i.e., the predictors in the dataset) did not contain strong enough signals or patterns complex for logistic regression, RF or SVM to detect.

Are the models better than logistic regression? This project aimed to develop models that were better at predicting IVF outcomes than logistic regression models, as outlined in section 1.3. The results of this project indicate that the performance of the non-parametric models RF and SVM was comparable to the baseline logistic regression trained on all available features. Despite the fact that RF is a model more flexible than logistic regression, able to capture non-linear relationships and interactions between variables, it did not achieve a significantly higher AUROC compared to logistic regression for the Sahlgrenska dataset. A possible explanation

is lack of signal in the data, as discussed above.

In general, using the women's age alone as a predictor produced AUROC scores comparable to the multivariate models. However, the sensitivity and specificity of the univariate model in first and third research questions revealed that few or no patients were predicted to experience a live birth. Such a model offers little value for clinical decision-making. The multivariate models generally achieved higher sensitivity, both with and without including dropouts. This supports their greater clinical potential, as additional variables beyond age contribute meaningfully to predicting the likelihood of achieving a live birth.

In this project, none of the three research questions yielded a significantly better result than the others. Given that all patients in the dataset had experienced an unsuccessful first IVF cycle, the patient characteristics in our dataset may be relatively homogeneous. In turn, this could possibly make it more difficult for the model to distinguish between individuals who will achieve a live birth and those who will not.

6.2 Feature importance

The variables shown in the SHAP figures were discussed with clinicians at Sahlgrenska, who provided a medical perspective on the feature importance. In all six of the best-performing models—one for each research question, both with and without dropout included—the variable *Total dose of gonadotropin* consistently ranked among the top two most important features.

All six SHAP plots revealed that higher doses of gonadotropin were associated with a reduced likelihood of live birth. The relationship between higher doses of gonadotropin and the chance of a successful IVF cycle is likely influenced by the strong correlation between gonadotropin dosage and patient age. Older women tend to receive higher doses (even if some patients receive a high dose despite a low age). Given that higher patient age is itself negatively associated with live birth rates, the total dose of gonadotropin may be acting as a proxy for age in the models.

The variable *Good Quality Embryos* was also ranked among the most important predictors. In this project, the models for research question one and two (where dropouts were included) showed that higher values for the *Good Quality Embryos* feature were associated with a lower probability of a live birth. However, this finding contrasts with previous research, where a high number of GQEs has been identified as a strong predictor of live birth likelihood in the first IVF cycle [32]. While this appears to contradict previous findings, it is important to interpret this result in context. In our models, a higher number of GQEs may reflect patients who underwent multiple ETs during their first cycle without achieving a live birth. In this sense, the GQE variable might serve as a proxy for *Number of failed ETs*, reflecting repeated unsuccessful transfers rather than directly indicating biological potential.

The *Good Quality Embryos* feature ranked among the top nine predictors only in

models for research questions one and two when dropouts were included. This suggests that patients with a high number of GQE (and consequently more failed ETs) during their first IVF cycle and who still did not have a live birth in that cycle may be less likely to achieve a live birth in subsequent cycles. In contrast, this feature was not as influential when dropouts were excluded.

This pattern aligns with clinical practice. For many patients, the first round of IVF functions as a "trial cycle" to adjust the gonadotropin dosage. If the initial dose is too low, it can result in an insufficient number of oocytes, sometimes preventing oocyte aspiration altogether. These patients often receive a revised, higher dose in a subsequent cycle, which may improve their outcomes. However, these patients may not produce a large number of GQEs in the first cycle. This distinction highlights differing trajectories: patients who discontinue treatment after an unsuccessful first cycle may have had multiple failed ETs despite a high number of GQEs, whereas those who continue often had suboptimal first cycles that later improved with adjusted stimulation protocols.

The hypothesis of varying characteristics in different patient groups is strengthened by the fact that the number of GQE was not a significant feature in the third research question. It is possible that most patients with a high number of GQE in their second IVF cycle had a live birth, after an suboptimal first cycle with a low number of GQE. Most patients who continue to the third cycle may therefore receive fewer GQEs in their first and second cycle, which makes the feature values similar among the patients who undergo a third round of IVF. Hence, it could be difficult to distinguish between patients who will get a live birth from those who will not based on this feature for the third research question. It is important to note that our thoughts regarding number of GQE are only hypotheses and require further investigations.

Additionally, an observed trend in the SHAP figures is that the top predictors are continuous variables. One possible reason for this could be that categorical variables, such as BMI, are divided into multiple groups based on value ranges. This splitting could obscure the variable's overall impact on the prediction, compared to treating BMI as a single continuous feature that captures its full range. It would be interesting to investigate whether the sum of SHAP values across all BMI categories equals the SHAP value obtained if BMI were considered as a continuous variable.

6.3 Longitudinal data and distribution shifts

The dataset covers a seven-year period, from 2016 to 2022, during which longitudinal distribution shifts may have occurred. Notably, there was a marked improvement in cumulative live birth rates following IVF between 2007 and 2017 [15]. Saket et al. identified two primary factors contributing to this improvement. First, there has been a shift in clinical practice from transferring cleavage-stage embryos (2-3 days post-fertilization) to transferring blastocyst-stage embryos (5-6 days post-fertilization). Second, embryo cryopreservation techniques have evolved, with a transition from slow-freezing of cleavage-stage embryos to vitrification of blastocysts.

Vitrification, a rapid-freezing method, has led to higher post-thaw survival rates, contributing to improved IVF outcomes.

The findings reported by Saket et al. do not provide direct evidence of a shift in birth rates between 2016 and 2022; however, they suggest that clinical practices in IVF treatment had not reached a point of stabilization in 2016. Further analysis of the Sahlgrenska dataset reveals a similar longitudinal trend, with an apparent increase in live birth rates over time. Specifically, patients who underwent treatment in more recent years exhibited higher chances of achieving a live birth compared to those treated in 2016, as well as a higher chance of experiencing a live birth during the first cycle.

The issue of distributional shift has been discussed throughout the project. If this shift is not accounted for in the modeling, there is a risk that the models will perform poorly on new, unseen data. To evaluate the model's ability to generalize, the model can be trained using data from 2016-2018 and evaluated using data from 2019-2022. This would make it possible to assess the model's ability to adapt to newer data, which is something worth investigating in future studies. Splitting the data randomly across years has been shown to result in overly optimistic model performance, as the model is trained on data that overlaps with the data it will encounter during testing [33]. This does not reflect the real-world deployment scenario, where future data is unavailable during training. By training the model on past data and testing it on future data, the evaluation more accurately simulates deployment conditions and helps assess the model's generalization ability.

Accounting for distributional shifts in the data was not addressed in this project, as it was considered outside the project's scope. Instead, this has been identified as an important next step, where future work should focus on validating the model using newer data to better assess its generalization under realistic deployment conditions. The limited dataset in this project was another reason this was not explored.

Another suggestion for addressing the distributional shift is to introduce a dummy-coded variable representing the year each patient began their IVF treatment (e.g., 2016, 2017, etc.). This would allow the model to account for year-specific changes in treatment protocols or patient characteristics. However, a potential risk of including such a variable is the possibility of overfitting, where the model may become overly sensitive to year-specific trends that do not generalize well to future data. While this approach was not used in this project, it could be considered for future testing and evaluation.

6.4 Implementation in clinical practice

Focusing on the potential benefits of deploying an IVF prediction model in clinical practice, one of the primary goals would be to help patients set realistic expectations regarding their chances of success. The aim is for such a model to serve as a decision-support tool, rather than the sole basis for the couple's decision of whether to undergo IVF treatment or not based on their expectations. Additionally, an IVF prediction model could assist clinicians by being integrated into electronic health

record systems. This integration would provide fertility specialists with an additional estimate of a patient's likelihood of conception, based on their medical data.

Choosing between a more optimistic or more pessimistic model in a clinical setting involves trade-offs. An optimistic model may encourage patients to continue with additional IVF cycles even if the first one fails, which can help them stay motivated. The overall chance of achieving a live birth may be more accurately represented by the optimistic model if the patient completes all three cycles, as dropouts are not included in the model training. At the same time, a pessimistic model may discourage patients from continuing treatment, even when they still have a reasonable chance of success. However, one potential advantage of the pessimistic model is that it may lead to less disappointment, as expectations are set lower. Overall, selecting between an optimistic or pessimistic IVF prediction model is not straightforward, as each presents advantages and drawbacks that can influence patient motivation and treatment decisions.

6.5 Future improvements

There are several possible improvements for the project, involving both the choice of variables and the modeling approach. One area of improvement is feature engineering. For instance, in this project, several continuous variables were grouped into categorical values, such as BMI. It could be useful to test whether the results change when variables like BMI are used as continuous instead, or vice versa, like *Good Quality Embryos* as categorical.

Furthermore, regarding feature importance, it would be valuable to investigate the relationships between variables, such as GQEs and the number of failed ETs, as previously discussed. A deeper analysis could involve comparing how these variables interact and contribute to the prediction in datasets with and without dropouts included. This could help clarify whether the observed association—where a higher number of GQEs corresponds to a lower chance of live birth when dropouts are included—is a meaningful finding or an artifact of the dataset.

One potential strategy to handle the limited size of the dataset in this project might be to explore transfer learning. Transfer learning is a technique used in the field of ML, where a neural network trained for one task, is applied for a different task. For example, a neural network trained to find dogs in images might also be able to find cats in images. Recent studies show that transfer learning can improve prediction performance, particularly when training data is relatively small [34]. While transfer learning is widely used in medical imaging, its application to non-image clinical data, like tabular data, has only become more common in recent years. For example, models trained on large genetic datasets have been fine-tuned to predict outcomes for specific diseases using much smaller, disease-specific datasets.

For our project, a promising direction could involve pretraining a deep learning model on a large, publicly available reproductive health dataset, such as fertility treatment registries or broader perinatal health data. After pretraining on this general dataset, the model could be fine-tuned on our smaller, more focused dataset.

This approach would allow the model to generalize better from the limited training examples and potentially enhance predictive performance. However, one potential drawback of transfer learning is the risk that the data used for pretraining might be too different from the data in this project [34]. If the datasets differ too much, for example in patient groups, types of variables or data format, the model might learn patterns that do not apply well to our case.

Although more advanced models, such as deep neural networks, were excluded from this project due to the limited dataset size and the need for large amounts of data to effectively learn patterns, they remain a promising option for future research. Their potential could be further explored using a broader dataset, such as a national cohort.

7

Conclusion

In this project, ML models were developed to provide individualized predictions of the likelihood of achieving a live birth after one or two unsuccessful IVF cycles. The goal was to help patients set realistic expectations and guide their decisions. Three research questions were addressed, each focusing on different scenarios: predicting the chance of a live birth in the second cycle after an unsuccessful first cycle; in the second or third cycle after an unsuccessful first cycle; and in the third cycle following two unsuccessful cycles.

This project employed logistic regression, RF and SVM models. Among these, logistic regression and RF showed slightly better performance in the first two research questions. The results are consistent with findings from studies using much larger datasets, which is encouraging and suggests that reasonable predictive performance can be achieved even with smaller datasets. The use of non-parametric models like RF proved advantageous, as they can naturally model non-linear relationships between input features and the outcome without requiring extensive feature transformations, unlike logistic regression.

However, none of the models achieved an AUROC score higher than 0.68, indicating they were moderate at predicting live births. The modest results observed in both this project and in previous studies suggest that predicting the likelihood of a live birth after one or two unsuccessful IVF cycles remains a generally challenging problem.

Feature importance for the models was reviewed and discussed with clinicians at Sahlgrenska University Hospital. Three variables emerged as particularly influential: total gonadotropin dose, number of GQE and woman's age. Among these, age is a well-established predictor of IVF outcomes. However, the feature evaluation showed interesting contributions of the variable *Good Quality Embryos*, which may be worth further investigation.

In the future, it would be interesting to explore how feature engineering impacts model performance, such as prioritizing either continuous or categorical features. Additionally, investigating more advanced techniques such as transfer learning or deep neural networks, particularly within a national cohort with a larger dataset, could help improve predictive accuracy and offer deeper insights into IVF treatment outcomes.

Bibliography

[1] "Fertilitetsbehandlingar i Sverige - Årsrapport 2024." (www.qivf.se).

[2] S. Theodoridis, ed., *Machine learning: a Bayesian and optimization perspective*. NET Developers Series, London San Diego: Elsevier, Academic Press, 2015.

[3] D. J. McLernon, E. W. Steyerberg, E. R. Te Velde, A. J. Lee, and S. Bhattacharya, "Predicting the chances of a live birth after one or more complete cycles of in vitro fertilisation: population based study of linked cycle data from 113873 women," *BMJ (Clinical research ed.)*, vol. 355, p. i5735, Nov. 2016.

[4] M. H. Walker and K. J. Tobler, "Female Infertility," in *StatPearls*, Treasure Island (FL): StatPearls Publishing, 2025.

[5] S. Emady, "Antralfolliklar." Link, Apr. 2024. Accessed: 2025-05-14.

[6] Q. Lai, H. Zhang, G. Zhu, Y. Li, L. Jin, L. He, Z. Zhang, P. Yang, Q. Yu, S. Zhang, J.-F. Xu, and C.-Y. Wang, "Comparison of the GnRH agonist and antagonist protocol on the same patients in assisted reproduction during controlled ovarian stimulation cycles," *International Journal of Clinical and Experimental Pathology*, vol. 6, pp. 1903–1910, Aug. 2013.

[7] "Q-IVF: Årsrapport 2018 - Gäller behandlingar startade 2016." (www.qivf.se).

[8] Sahlgrenska Universitetssjukhuset, "Infertilitetsutredning." Link, Apr. 2018. Accessed: 2025-05-14.

[9] D. J. McLernon, E.-A. Raja, J. P. Toner, V. L. Baker, K. J. Doody, D. B. Seifer, A. E. Sparks, E. Wantman, P. C. Lin, S. Bhattacharya, and B. J. Van Voorhis, "Predicting personalized cumulative live birth following in vitro fertilization," *Fertility and Sterility*, vol. 117, pp. 326–338, Feb. 2022.

[10] M. B. Ratna, S. Bhattacharya, N. van Geloven, and D. J. McLernon, "Predicting cumulative live birth for couples beginning their second complete cycle of in vitro fertilization treatment," *Human Reproduction (Oxford, England)*, vol. 37, pp. 2075–2086, Aug. 2022.

[11] Z. Barnett-Itzhaki, M. Elbaz, R. Butterman, D. Amar, M. Amitay, C. Racowsky, R. Orvieto, R. Hauser, A. A. Baccarelli, and R. Machtinger, "Machine learning vs. classic statistics for the prediction of IVF outcomes.," *Journal of assisted reproduction and genetics*, vol. 37, pp. 2405–2412, Oct. 2020.

[12] C. Blank, R. R. Wildeboer, I. DeCroo, K. Tilleman, B. Weyers, P. de Sutter, M. Mischi, and B. C. Schoot, "Prediction of implantation after blastocyst transfer in in vitro fertilization: a machine-learning perspective.," *Fertility and sterility*, vol. 111, pp. 318–326, Feb. 2019.

[13] Z. Liu, H. Zhang, F. Xiong, X. Huang, S. Yu, Q. Sun, L. Diao, Z. Li, Y. Wu, Y. Zeng, and C. Huang, "Prediction of clinical pregnancy outcome after single fresh blastocyst transfer during in vitro fertilization: an ensemble learning perspective.," *Human fertility (Cambridge, England)*, vol. 27, p. 2422918, Dec. 2024.

[14] S. Dehghan, R. Rabiei, H. Choobineh, K. Maghooli, M. Nazari, and M. Vahidi-Asl, "Comparative study of machine learning approaches integrated with genetic algorithm for IVF success prediction," *PLOS ONE*, vol. 19, no. 10, p. e0310829, 2024.

[15] Z. Saket, K. Källén, K. Lundin, Magnusson, and C. Bergh, "Cumulative live birth rate after IVF: trend over time and the impact of blastocyst culture and vitrification," *Human Reproduction Open*, vol. 2021, no. 3, p. hoab021, 2021.

[16] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[17] Z.-H. Zhou, *Machine learning*. Singapore: Springer, 2021.

[18] H. A. Salman, A. Kalakech, and A. Steiti, "Random Forest Algorithm Overview," *Babylonian Journal of Machine Learning*, vol. 2024, pp. 69–79, June 2024.

[19] "Appendix C: Restricted Cubic Regression Splines," in *Analysis of Health Surveys*, pp. 345–346, July 1999.

[20] D. U. Ozsahin, M. Taiwo Mustapha, A. S. Mubarak, Z. Said Ameen, and B. Uzun, "Impact of feature scaling on machine learning models for the diagnosis of diabetes," in *2022 International Conference on Artificial Intelligence in Everything (AIE)*, pp. 87–94, Aug. 2022.

[21] T. Silva Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, and P. Flach, "Classifier calibration: a survey on how to assess and improve predicted class probabilities," *Machine Learning*, vol. 112, pp. 3211–3260, Sept. 2023.

[22] A. Baratloo, M. Hosseini, A. Negida, and G. El Ashal, "Part 1: Simple Definition and Calculation of Accuracy, Sensitivity and Specificity," *Emergency*, vol. 3, no. 2, pp. 48–49, 2015.

[23] T. C. F. Polo and H. A. Miot, "Use of ROC curves in clinical and experimental studies," *Jornal Vascular Brasileiro*, vol. 19, p. e20200186.

[24] S. Halligan, D. G. Altman, and S. Mallett, "Disadvantages of using the area under the receiver operating characteristic curve to assess imaging tests: A discussion and proposal for an alternative approach," *European Radiology*, vol. 25, no. 4, pp. 932–939, 2015.

[25] D. Chicco, N. Tötsch, and G. Jurman, "The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation," *BioData Mining*, vol. 14, p. 13, Feb. 2021.

[26] W. Yang, J. Jiang, E. M. Schnellinger, S. E. Kimmel, and W. Guo, "Modified Brier score for evaluating prediction accuracy for binary outcomes," *Statistical methods in medical research*, vol. 31, pp. 2287–2296, Dec. 2022.

[27] S. Bassan, G. Amir, and G. Katz, "Local vs. Global Interpretability: A Computational Complexity Perspective," 2024.

[28] M. Haddouchi and A. Berrado, "A survey and taxonomy of methods interpreting random forest models," *arXiv.org*, July 2024.

[29] A. V. Ponce-Bobadilla, V. Schmitt, C. S. Maier, S. Mensing, and S. Stodtmann, "Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development," *Clinical and Translational Science*, vol. 17, p. e70056, Oct. 2024.

[30] O. Ragnarsson, "Hypofyssvikt (hypofysär underproduktion, hypopituitarism)." Link, May 2024. Accessed: 2025-05-14.

[31] G. Shan, "Improved Confidence Intervals for the Youden Index," *PLoS ONE*, vol. 10, p. e0127272, July 2015.

[32] T. Xu, A. de Figueiredo Veiga, K. C. Hammer, I. C. Paschalidis, and S. Mahalingaiah, "Informative predictors of pregnancy after first IVF cycle using eIVF practice highway electronic health records," *Scientific Reports*, vol. 12, p. 839, Jan. 2022.

[33] H. Zhou, Y. Chen, and Z. Lipton, "Evaluating Model Performance in Medical Datasets Over Time," in *Proceedings of the Conference on Health, Inference, and Learning* (B. J. Mortazavi, T. Sarker, A. Beam, and J. C. Ho, eds.), vol. 209 of *Proceedings of Machine Learning Research*, pp. 498–508, PMLR, June 2023.

[34] A. Ebbehoj, M. Thunbo, O. E. Andersen, M. V. Glindtvad, and A. Hulman, "Transfer learning for non-image data in clinical research: A scoping review," *PLOS Digital Health*, vol. 1, p. e0000014, Feb. 2022.

A

Appendix: Data statistics

Table A.1 summarizes the categorical variables, including the outcome variable *Live birth*, with counts and percentages for each subcategory across all cycles and separately by cycle. Missing values are also included. For instance, in the BMI column, 531 of the 5,170 cycles are classified as '<20' across all cycles. In the first cycle, 338 of 3,217 patients fall into this category, with one missing value reported under 'Missing values'.

The original dataset includes two separate columns for infertility causes (cause 1 and cause 2), on which the models were trained using one-hot encoding. For presentation purposes, these columns were combined into a single summary column (Infertility cause). As each patient can have up to two recorded causes, the total value shown (e.g., 6006) reflects the number of individual cause entries, not the number of unique patients or cycles.

The statistics for the columns *Fertilization method for the transferred embryo*, *Number of culture days at ET* and *Number of embryos transferred at ET* are summarized by aggregating subcategory counts across all ETs. For example, IVF was used in 3,347 of the 6,317 ETs, including all cycles and both fresh and frozen transfers.

Table A.1: Statistical overview of the categorical columns in the dataset, with the percentage and number of values of each subcategory for all cycles in total and the three cycles separately. The number and percentage of missing values are also listed.

	All Cycles		Cycle 1		Cycle 2		Cycle 3	
Categorical columns	%	N	%	N	%	N	%	N
BMI:		5170		3217		1348		605
<20	10.3	531	10.5	338	9.6	130	10.4	63
20-24.9	49.0	2534	50.1	1613	47.0	634	47.4	287
25-29.9	26.3	1362	25.4	817	28.2	380	27.3	165
30-34.9	12.6	654	12.1	390	13.4	181	13.7	83
>35	1.7	88	1.8	58	1.7	23	1.2	7
Missing values	<0.1	1	<0.1	1	-	-	-	-
AFC total:		5170		3217		1348		605
<10	18.6	962	15.8	507	22.4	302	25.3	153
10-30	72.0	3721	74.4	2393	68.3	921	67.3	407
>30	9.4	487	9.9	317	9.3	125	7.4	45

Categorical columns	All Cycles %	All Cycles N	Cycle 1 %	Cycle 1 N	Cycle 2 %	Cycle 2 N	Cycle 3 %	Cycle 3 N
Missing values	-	-	-	-	-	-	-	-
Average menstrual cycle:		5170		3217		1348		605
<25	3.6	187	2.9	94	4.4	59	5.6	34
25-29.9	58.7	3034	58.4	1880	58.1	783	61.3	371
30-34.9	21.0	1085	21.9	703	20.0	269	18.7	113
>35	8.5	442	8.7	280	8.9	120	6.9	42
Missing values	8.2	422	8.1	260	8.7	117	7.4	45
Infertility cause:		6006		3712		1582		712
Tubal factor	9.1	470	9.1	292	9.3	125	8.8	53
Unexplained	33.6	1740	34.2	1101	32.4	437	33.4	202
Same sex/single women	3.0	153	3.6	117	2.1	29	1.2	7
Male factor	37.8	1955	36.4	1169	39.5	532	42.0	254
Egg factor	3.1	162	2.6	82	4.1	55	4.2	25
PCOD	10.5	540	10.7	343	10.6	142	9.1	55
Hormonal	11.0	567	10.5	338	11.8	159	11.5	70
Endometriosis	6.0	308	6.2	200	5.5	74	5.6	34
Other	2.2	111	2.2	70	2.1	29	2.0	12
Missing values	-	-	-	-	-	-	-	-
Sperm origin:		5170		3217		1348		605
Ejaculated	93.1	4815	92.3	2970	93.9	1266	95.7	579
Donated	3.5	182	4.4	143	2.4	33	1.0	6
Surgical sperm extraction	2.3	118	2.1	68	2.5	34	2.6	16
Frozen ejaculated	1.1	55	1.1	36	1.1	15	0.7	4
Missing values	-	-	-	-	-	-	-	-
Planned fertilization method:		5170		3217		1348		605
IVF	44.7	2309	52.5	1689	34.8	469	25.0	151
ICSI	54.9	2838	47.0	1513	64.8	873	74.7	452
Missing values	0.4	23	0.5	15	0.4	6	0.3	2
Stimulation protocol:		5170		3217		1348		605
Agonist	38.2	1973	35.8	1152	40.8	550	44.8	271
Antagonist	61.8	3197	64.2	2065	59.2	798	55.2	334
Missing values	-	-	-	-	-	-	-	-
Aspiration performed:		5170		3217		1348		605
True	93.9	4855	92.8	2984	95.7	1290	96.0	581
False	6.1	315	7.2	233	4.3	58	4.0	24
Missing values	-	-	-	-	-	-	-	-
Live birth:		5170		3217		1348		605
Yes	57.9	2993	53.1	1709	63.1	850	71.7	434
No	42.1	2177	46.9	1508	36.9	498	28.3	171
Missing values	-	-	-	-	-	-	-	-

Categorical columns	All Cycles %	All Cycles N	Cycle 1 %	Cycle 1 N	Cycle 2 %	Cycle 2 N	Cycle 3 %	Cycle 3 N
Fertilization method for the transferred embryo:		6317		4103		1570		644
IVF	53.0	3347	60.5	2484	41.7	655	32.3	208
ICSI	46.9	2963	39.4	1616	58.1	912	67.5	435
Missing values	0.1	7	<0.1	3	0.2	3	0.1	1
Number of culture days at ET:		6317		4103		1570		644
1-3	40.8	2575	35.8	1470	45.4	713	60.9	392
4-8	59.1	3732	64.0	2627	54.4	854	39.0	251
Missing values	0.3	10	0.1	6	0.2	4	0.1	1
Number of embryos transferred at ET:		6317		4103		1570		644
SET	92.7	5857	98.1	4024	91.8	1442	60.7	391
DET	7.1	447	1.75	72	7.8	123	39.1	252
Missing values	0.2	13	0.2	7	0.3	5	0.2	1

Table A.2 presents the continuous variables in the dataset, reporting median, standard deviation (std) and the number of observed values (N) prior to data replacement or imputation.

Table A.2: All ten continuous variables present in the dataset are listed. N represents the total number of values before data replacement and imputation. Reported values for the median and standard deviation (std) are calculated across all cycles, before any data replacement or imputation.

Variable	**Median (std)**	**N**
Women age	32.8 (4.3)	5170
Partner age	34.7 (5.7)	5120
Female length, cm	166.0 (6.9)	5169
Infertility duration, months	30.0 (21.2)	5107
Total gonadotropin dose	1750.0 (1262.5)	5170
Number of oocytes retrieved	9.0 (6.5)	4855
Good Quality Embryos (GQE)	2.0 (2.2)	5170
Number of frozen embryos	1.0 (2.3)	5170
Number of ETs	1.0 (0.9)	5170
Number of failed ETs	1.0 (1.3)	5170

B

Appendix: Model features

Table B.1 shows the 87 variables used to train the models. Three of these were only used in the logistic regression model adjusted for women's age using RCS: Women's age (splines 1), (splines 2) and (splines 3).

Table B.1: All 87 used for the modeling (three variables are only used for logistic regression where age had been adjusted for).

Variable
Women age
Women age (splines 1)
Women age (splines 2)
Women age (splines 3)
Female length, cm
Infertility duration, months
Partner age
BMI: <20
BMI: 20-25
BMI: 25-30
BMI: 30-35
BMI: >35
AFC total: <10
AFC total: 10-30
AFC total: >30
Average menstrual cycle: <25
Average menstrual cycle: 25-30
Average menstrual cycle: 30-35
Average menstrual cycle: >35
Infertility cause (first): Egg factor
Infertility cause (first): Endometriosis
Infertility cause (first): Hormonal
Infertility cause (first): Male factor

Continued on next page

Table B.1 – continued from previous page

Variable
Infertility cause (first): Other
Infertility cause (first): PCOD
Infertility cause (first): Same sex/single women
Infertility cause (first): Tubal factor
Infertility cause (first): Unexplained
Infertility cause (second): Egg factor
Infertility cause (second): Endometriosis
Infertility cause (second): Hormonal
Infertility cause (second): Male factor
Infertility cause (second): Other
Infertility cause (second): PCOD
Infertility cause (second): Same sex/single women
Infertility cause (second): Tubal factor
Infertility cause (second): Unexplained
Sperm origin: Donated
Sperm origin: Ejaculated
Sperm origin: Frozen ejaculated
Sperm origin: Surgical sperm extraction
Planned fertilization method: ICSI
Planned fertilization method: IVF
Total gonadotropin dose
Aspiration performed: False
Aspiration performed: True
Number of oocytes retrieved
Number of frozen embryos
Good Quality Embryos (GQE)
Stimulation protocol: Agonist
Stimulation protocol: Antagonist
Fertilization method for the transferred embryo (fresh): ICSI
Fertilization method for the transferred embryo (fresh): IVF
Fertilization method for the transferred embryo (FET1): ICSI
Fertilization method for the transferred embryo (FET1): IVF
Fertilization method for the transferred embryo (FET2): ICSI
Fertilization method for the transferred embryo (FET2): IVF
Fertilization method for the transferred embryo (FET3): ICSI
Fertilization method for the transferred embryo (FET3): IVF

Continued on next page

Table B.1 – continued from previous page

Variable
Fertilization method for the transferred embryo (FET4): ICSI
Fertilization method for the transferred embryo (FET4): IVF
Fertilization method for the transferred embryo (FET5): ICSI
Fertilization method for the transferred embryo (FET5): IVF
Fertilization method for the transferred embryo (FET6): ICSI
Number of culture days at fresh ET: 1-3
Number of culture days at fresh ET: 4-8
Number of culture days at FET1: 1-3
Number of culture days at FET1: 4-8
Number of culture days at FET2: 1-3
Number of culture days at FET2: 4-8
Number of culture days at FET3: 1-3
Number of culture days at FET3: 4-8
Number of culture days at FET4: 1-3
Number of culture days at FET4: 4-8
Number of culture days at FET5: 4-8
Number of culture days at FET6: 4-8
Number of embryos transferred at fresh ET: DET
Number of embryos transferred at fresh ET: SET
Number of embryos transferred at FET1: DET
Number of embryos transferred at FET1: SET
Number of embryos transferred at FET2: DET
Number of embryos transferred at FET2: SET
Number of embryos transferred at FET3: SET
Number of embryos transferred at FET4: SET
Number of embryos transferred at FET5: SET
Number of embryos transferred at FET6: SET
Number of ETs (only when dropouts are included)
Number of failed ETs (only when dropouts are excluded)

C

Appendix: Model training specifics

In this Appendix, the hyperparameters evaluated using GridSearchCV are specified. These are listed in table C.1, and includes the hyperparameters used for the models with the highest average test AUROC across all research questions, with and without dropouts.

Table C.1: List of hyperparameters, their respective values used in GridSearchCV for each ML model (RF, SVM, logistic regression, and logistic regression where age is adjusted using splines), and the used values.

Model	**Hyperparameter**	**Search space**	**With dropouts**	**Without dropouts**
RF	`max_depth`	3, 5, 10, None	10	None
	`min_samples_leaf`	1, 5, 10, 20	10	1
	`min_samples_split`	2, 4, 6	2	6
	`min_weight_fraction_leaf`	0, 0.01, 0.05	0.01	0.01
	`class_weight`	balanced, None	balanced	None
SVM	`C`	0.01, 0.1, 1, 10	1	1
	`class_weight`	balanced, None	balanced	balanced
LR	`penalty`	l1	l1	l1
	`C`	0.01, 0.1, 1, 10	0.1	0.1
	`solver`	liblinear	liblinear	liblinear
	`class_weight`	balanced, None	balanced	None
LR (adjusted)	`penalty`	l1	l1	l1
	`C`	0.01, 0.1, 1, 10	0.1	0.1
	`solver`	liblinear	liblinear	liblinear
	`class_weight`	balanced, None	None	balanced

DEPARTMENT OF ELECTRICAL ENGINEERING
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY