Predicting fertility across 11 European countries

Date: 17th of September 2025
Name: Jasper van der Veen
E-mail address: j.van.der.veen.2@student.rug.nl
University: Rijksuniversiteit Groningen
Faculty: Gedrags- en Maatschappijwetenschappen
Course: Master thesis

Abstract

This thesis investigates the predictability of short-term fertility outcomes across European countries using three machine learning models: penalized logistic regression, support vector machines, and extreme gradient boosting (XGBoost). Drawing on harmonized international survey data, the analysis explores both theoretical and technical factors that may influence model performance, including sample size, class balance, and model-specific characteristics. To ensure robustness, all models are evaluated through repeated cross-validation and assessed using multiple performance metrics, including accuracy, ROC-AUC, F1 score, and Brier score. Findings indicate that predictive performance varies substantially across countries, with class distribution and model architecture playing a central role. XGBoost consistently outperforms the other models, particularly in countries with balanced class distributions. Variable importance analyses reveal that while fertility intentions are key predictors in linear models, age-related variables dominate in tree-based approaches. The study highlights the need to complement explanatory approaches with predictive frameworks and offers methodological insights for future research on demographic behaviour.

Contents

Chapter 1: Introduction

In 2012, Charles Duhigg wrote in the *New York Times* about how much companies know about individuals. Statistician Andrew Pole revealed that Target's statistics department could predict when a customer was pregnant, based on purchase history. This led to an awkward moment when a father confronted a store employee because his high school daughter received pregnancy-related coupons. He later apologized after learning she was indeed pregnant. In this case, a supermarket knew about the pregnancy before her own father (Duhigg, 2012).

This example highlights two key points for this thesis. First, behaviour prediction has become central in science and business, signalling a shift in statistical approaches (Rahal et al., 2022). Predictive methods often rely on complex models capable of uncovering intricate relationships. Drawing on machine learning and techniques from other scientific fields can enrich sociological methods (Breiman, 2001b). The reasons are manyfold. First, evaluating a statistical model on out-of-sample prediction reduces overfitting and thus leads to a more robust empirical literature. Second, these techniques are more capable of modelling non-linear relationships and interactions improving scientific understanding. Third, if predictive performance is poor, it could be sign that current theories are not sufficient in explaining a phenomenon (Salganik et al, 2020). Fourth, calculating the relative importance of different variables could give more context into how behaviour is shaped, which could help formulate new theories..

The second key point of the example is that this is a successful case of predicting who will have a birth which begs the question of how predictable births are. Fertility is a widely studied life outcome (Mills et al., 2011; Balbo et al., 2012), especially in light of its demographic, economic, and policy implications. Declining fertility rates across much of Europe have prompted increased scholarly attention, yet few studies have explored fertility using predictive modelling techniques (Sivak et al., 2024).

This study

This thesis will compare the predictability of fertility across eleven different European countries. Specifically, I will compare the performance of three predictive models across fertility data from eleven European countries, using data from the Generations and Gender Survey (GGS). Using the same sets of variables (e.g., partnership status, sex, age) in the prediction models, I will further be able to compare variable importance across countries. This research thus addresses three research questions. First, how predictable are fertility outcomes across Europe? Second, what are cross-country differences—in either methodology and/or culture—that affect predictability? Third, which variables are most important and how do they vary across countries. The comparison of multiple different countries based on fertility predictability has not been attempted before, therefore adding to the scientific literature.

Scientific relevance

While an increasing number of researchers are engaging with prediction methods, these approaches remain underutilized within sociology and demography (Arpino et al., 2021; Salganik et al., 2020; Stulp et al., 2023). Moreover, systematically quantifying predictive ability is closely linked to advancing our scientific understanding of social phenomena, yet this has not been extensively applied to fertility outcomes (Garip, 2020).

This thesis addresses that gap by evaluating the predictability of fertility behaviour across multiple European countries using machine learning models. By doing so, it not only contributes to methodological innovation but also provides insights into the complexity of fertility decisions in diverse social contexts. Prediction allows for the identification of non-linear relationships and complex interactions between factors influencing fertility, which traditional linear models may overlook. Additionally, the use of out-of-sample evaluation reduces the risk of spurious findings, strengthening the robustness of empirical conclusions (Molina & Garip, 2019).

A focus on prediction is useful for assessing how well fertility can be predicted, which offers an indirect test of existing theoretical frameworks. If models perform poorly, this may indicate gaps in our understanding of the determinants of fertility, pointing to areas for further theoretical development (Salganik et al., 2020). This work thus advances both the scientific methods and substantive knowledge within fertility research.

In the case of fertility, this type of systematic predictive evaluation has been largely absent. Without such efforts, it remains unclear whether low predictive performance reflects inherent unpredictability in individual fertility choices or shortcomings in existing theoretical models. By evaluating the predictability of fertility across multiple national contexts, this thesis contributes to clarifying the scope and limits of current explanations, and ultimately strengthens the scientific foundations of fertility research.

Societal relevance

Fertility behaviour is an increasingly pressing societal issue. In many Western countries, fertility rates are below replacement level, leading to population ageing, labour shortages, and pressure on welfare systems (Bloom et al., 2009). At the same time, individuals face growing barriers to having children, such as economic insecurity, housing shortages, work-family conflict, and shifting gender norms. Understanding who is likely to have children, and under what conditions, can inform public debate and support policy interventions that reduce slowing fertility rates which could help solve these issues.

Chapter 2: Theory

Orthodox Modelling in Social Science

Social science models typically aim to reveal relationships between variables, using simple models like linear or logistic regression, with the aim of making causal inferences. These models have interpretable parameters that reflect the strength of a particular variable and help test hypotheses (Agresti & Finlay, 2013). This approach has culminated in a large body of research on fertility (see e.g., Balbo et al., 2012 for review). However, establishing causality is challenging without experimental designs and often relies on theory construction, which is subjective and difficult to prove (Watts, 2014). Perhaps this is one reason why a replication crisis is observed throughout the social sciences and other fields (Bem, 2011; Novella, 2012; Aarts et al., 2015). This crisis refers to the many findings of studies that failed further replication. Another reason for the replication may be the uncritical use of p-values. Regression models test hypotheses via p-values, with the test being whether a parameter is different from zero. To aid with this decision, a threshold (usually 5 percent) is used as evidence and the finding is considered “statistically significant”. This practice has faced criticism due to the fact that the p-value is highly sensitive to statistical practices like sample selection, variable inclusion, and outlier removal (Gadbury & Allison, 2012). Since researchers can choose how to apply these statistical practices and journals prefer to publish significant results, many argue that researcher degrees of freedom contribute to the replication crisis by making p-values unreliable indicators of true effects. These issues have spurred reassessment of statistical significance as a reliability marker and prompted calls for methodological improvement (Aarts et al., 2015). It has been suggested that a way to relieve the uncertainty, is to focus on prediction (Yarkoni & Westfall, 2017)

A Focus on Prediction

A first and crucial component of prediction is that a model’s quality should be evaluated based on its performance on cases it has not seen before, known as out-of-sample prediction. Models that perform well on the data they were trained on, or in-sample, may perform poorly on new data because they have captured noise specific to the training set that does not generalize. This phenomenon is called overfitting. Machine learning models are specifically designed to prevent overfitting and to optimize out-of-sample predictive accuracy. Such models may also help to reduce underfitting, which occurs when the model lacks sufficient complexity to capture the relationship between variables and therefore performs poorly across both training and test data. Underfitting is particularly relevant in the context of social behaviour, where relationships between predictors and outcomes are often non-linear. People do not evaluate situations in a consistent or proportional way, and this inconsistency can be attributed to cognitive biases in decision-making. For instance, individuals tend to overestimate the likelihood and impact of rare events (Kahneman, 2011). These distortions in perception and judgment challenge the assumptions of linear relationships, which assume stable and predictable responses to contextual changes. As a result, behaviour may shift in abrupt or context-dependent ways that linear models struggle to detect.

Machine learning models are capable of addressing these challenges by allowing for the inclusion of a large number of variables and by effectively modeling non-linear effects and interactions. In doing so, they help mitigate underfitting and reveal patterns that may be obscured in simpler statistical models.

Beyond their ability to balance overfitting and underfitting, machine learning model predictions provide a valuable benchmark. These predictions enable researchers to assess whether there is potential for model improvement. For example, if a relatively simple model performs only slightly worse than a much more complex one, it may be concluded that adding more variables is unlikely to yield meaningful gains (Shmueli, 2010). Salganik and colleagues (2020), for instance, found that in predicting five life outcomes, more complex machine learning models did not substantially outperform simpler benchmark models.

A drawback of machine learning models is that their increased complexity often makes interpreting specific effects more difficult. As models become more complex, they generally become less transparent. Although complexity can enhance predictive power, it does not necessarily improve our understanding of particular relationships. To address these challenges, various interpretability methods have been developed.

In summary, the social sciences can benefit from a focus on prediction because of machine learning's ability to improve out-of-sample accuracy, handle complex and non-linear relationships, and offer empirical benchmarks for model performance—ultimately complementing traditional explanatory approaches by revealing patterns that may otherwise remain hidden, while also encouraging the development of more robust and generalizable theories.

Prediction in Fertility Research

Fertility outcomes are studied by many different disciplines because of their importance for individuals and populations (Sivak et al., 2024). Although the field is heavily statistics-minded and demographers have long been concerned with population forecasts, a focus on prediction remains surprisingly limited. Despite this, recent predictive work—such as that by Arpino et al. (2021)—has revealed heterogeneous and non-linear patterns in demographic outcomes, suggesting clear potential for advancing fertility research. Yet these insights have received relatively little attention. Many researchers have participated in large-scale collaborative efforts, including data challenges like the Fragile Families Challenge (Salganik et al., 2020) and fertility-focused projects by Sivak et al. (2024). While such initiatives have contributed valuable methodological innovations, they also consistently expose the low predictive power of existing models. This raises important questions about the strength of prevailing theories and the adequacy of current measurement strategies (Salganik et al., 2020). Because of this reason we will be using 13 already well known variables that have influence on fertility behaviour. These will be: Education, Number of marriages, Current employment status, Occupational category, Health, Religion, Current partnership status, Marriage, Satisfaction with the relationship, Fertility intentions, Age of the youngest child, Gender and Age based on the research of Balbo and colleagues (2012).

Theoretical Expectations for Variability in Predictability

There is still limited research on prediction in the social sciences, and this holds even more true for fertility research. The absence of existing literature makes it difficult to establish clear benchmarks or informed expectations for predictive performance across different contexts. Nevertheless, it is still possible to consider the factors that constrain or enhance predictions. In machine learning, prediction error can be divided into two categories (Lundberg et al., 2024). The first is learning error, which stems from the model's learning process. Factors such as sample size, model choice, and noise from data collection can influence how well a model make predictions from the available data. For example, country-level sample sizes in the Generations and Gender Survey (GGS) dataset (see Table 2.1) shows that sample size vary considerably across countries (from 1035 to 5679). A larger sample size can reduce learning error by providing more information during model training. While overall sample size is important because larger samples can reduce learning error by providing more training information, the distribution of the outcome variable is equally critical. Rare events, such as low fertility, are inherently more difficult for models to learn. In the Generations and Gender Survey (GGS) dataset (see Table 2.1), Bulgaria has the largest sample size among the countries included in this study, which would generally support model performance. However, Bulgaria also has a low fertility rate, with only 10 percent of respondents having had a child during the observation window. This creates a substantial class imbalance that limits the model's ability to learn meaningful patterns, particularly due to the small number of positive cases. As a result, the model may struggle to distinguish between individuals who had a child and those who did not, despite the large overall sample size.

The second source of error is irreducible error, which refers to the portion of error that remains even when all learning error is eliminated. Irreducible error largely depends on the task itself. Its size reflects how much the outcome is shaped by variables we can actually measure. This type of error can not be reduced by making different methodological choices. In some domains, outcomes may be highly influenced by unpredictable events, or unmeasured factors, all of which contribute to higher irreducible error. Such unpredictability is hard to quantify and in the same way it is hard to assess which unmeasured factors are most problematic (e.g., the known unknowns: (epi)genetics certainly play a role in fertility, but they are rarely measured in survey, and the unknown unknowns; variables whose importance we do not (yet) realise). The difference found between different model types can only be attributed to learning error, while the difference between countries is mostly irreducible error depending on how similar the data collections is.

Predictability of behaviour can also vary across different social and cultural settings. Because behaviour is strongly shaped by social norms and institutional frameworks, the predictability of such behaviour may vary accordingly. In contexts where pronatalist norms are strong, such as societies where marriage almost inevitably leads to parenthood or where religious and cultural expectations promote early childbearing, fertility outcomes tend to follow more uniform patterns, and standard predictors like age and marital status probably perform well. For instance, (multilevel) analyses of European data show that similarity in age norms across countries strengthens the link between marital status and fertility behaviour (Liefbroer et al., 2014).

In contrast, in individualised societies with more relaxed fertility norms, personal values and choices play a greater role. Cases such as childfree couples or single women opting for sperm donation illustrate increased behavioural divergence, which complicates prediction or at least necessitates additional variables to account for this divergence properly. This aligns with findings that in settings with strong community education or cultural transmission, individual-level variability increases, reducing predictability based on typical demographic features (Henrich & McElreath, 2003).

This distinction is especially relevant in fertility research: variations in normative strength across countries directly influence the predictive relevance of common demographic variables. Where norms are cohesive, individual variation in behaviour due to preferences may be constrained and fertility becomes easier to forecast. Where norms are weaker, forecasts become more uncertain, reflecting greater behavioural heterogeneity.

In the following section, we describe two strongly varying population-level measures that may relate to the predictability of fertility outcomes, namely a country’s opinion on childlessness and socially accepted age ranges for childbearing.

Social fertility period

In regard to fertility, culture is important to understand the differences between different countries. In many societies there exists a cultural timetable on which important life event expectations are based (Settersten & Hagestad, 1996). When people reach a certain age, society will expect them to have children. If they do not match this expectation they will be sanctioned (Billari et al., 2010; Lazzari et al., 2022; Diaz & Fiel, 2016). I expect that in the countries in which this pressure is more intense, fertility should be more predictable. More predictable means that a model should make less mistakes and is more able to distinguish having a child and not having a child, which would show a difference in irreducible error, when using the same model. In general, fertility norms are stricter in more traditional and conservative societies (Norris & Inglehart, 2004). In keeping with cultural strictness, variables like age, marital status and number of marriages should also be more potent predictors for fertility in stricter countries. To operationalise cultural strictness in regard to fertility behaviour, Liefbroer et al. (2014) used a measure called a "social reproductive period", which is the number of years in which it's socially acceptable to have children (based on data collected in 2006 and 2007). The shorter this lifespan is, the stricter the fertility norms should be in a country, which would make fertility more predictable. This would make Hungary the most predictable country and Austria the least predictable, as the former has the shortest reproductive period and the latter the longest (Table 2.1). Across Europe there does not seem to be a lot of cross-country variation in social fertility period except for Hungary and Austria.

Hypothesis 1: A country with a shorter social reproductive period should be more predictable than a country with a longer social reproductive period.

Opinion on childlessness

In addition to examining the social fertility lifespan, understanding residents' attitudes toward childlessness provides valuable insight into the societal pressure to have children. If remaining childless is widely accepted, individuals are likely to feel less compelled to have children, even when circumstances are favourable. To assess this, I utilized data from the European Values Survey (EVS) conducted in 2017, which asked respondents to rate their agreement with the statement: "It is a duty towards society to have children." Responses were scored on a scale where a higher score (5) indicates stronger disagreement with this statement, meaning greater acceptance of childlessness. As shown in Table 2.1, Western countries such as the Netherlands, France, and Germany exhibit the highest acceptance of childlessness, while countries like Bulgaria and Georgia show the least acceptance. This pattern aligns with broader cultural trends, as Western European societies tend to emphasize individual autonomy and personal choice, which supports greater acceptance of diverse life trajectories, including childlessness. In contrast, Eastern European countries often place a stronger emphasis on traditional family values, where having children is viewed not only as a personal milestone but also as a social expectation and moral responsibility (Merz & Liefbroer, 2012). These cultural differences help explain the regional variation in how childlessness is perceived and the degree of social pressure individuals may experience.

Hypothesis 2: A country which is less accommodating to childlessness should be more predictable than a country with is more accommodating.

Table 2.1: Country Sample size, proportion of People who had child within timeframe and length of social reproductive period

COUNTRY	SAMPLE SIZE	PROPORTION OF PEOPLE WHO HAVE HAD A CHILD IN SAMPLE	AVERAGE OPINION ABOUT CHILDLESSNESS	SOCIAL FERTILITY PERIOD
BULGARIA	5679	0.100	1.913	21.25
HUNGARY	4996	0.445	2.909	20.80
POLAND	4726	0.512	2.913	23.25
GEORGIA	4401	0.175	2.112	21.25
AUSTRIA	3912	0.198	3.308	26.95
RUSSIA	3752	0.438	2.850	22.35
FRANCE	3183	0.192	3.662	24.45
NETHERLANDS	3071	0.198	4.231	22.65
CZECH REPUBLIC	1522	0.142	2.466	22.35
GERMANY	1431	0.224	3.443	23.85
LITHUANIA	1035	0.443	2.722	21.58

Chapter 3: Method

This chapter examines three machine learning methods: ridge regression, support vector machines (SVM), and extreme gradient boosting (XGBoost). It covers their theoretical foundations, implementation on fertility data, and evaluation through cross-validation and performance metrics.

The first model employed is a penalised logistic regression, which serves as a baseline due to its simplicity and interpretability. The second model is a support vector machine, included for its greater complexity and its distinct underlying methodology compared to the other models. The third model is an extreme gradient boosting (XGBoost) algorithm, selected for its strong performance in recent classification tasks (Park & Lee, 2022; Niazkar et al., 2024). Given its proven effectiveness in similar contexts, we expect XGBoost to achieve the highest predictive performance in this study.

Cross-Validation

All three models rely on hyperparameters, which are values that control how the model learns from the data and generates predictions. To determine the optimal values for these hyperparameters, a portion of the training data must be set aside to assess model performance. Instead of relying on a single split of the data, this study uses cross-validation, a widely recommended resampling method (Rooij and Weeda 2020). In this approach, the training data are divided into equally sized subsets. The amount of subsets or folds can vary, but as part of study we have chosen for five folds for all models to increase comparability. The model is trained on all but one fold and validated on the remaining fold. Its performance is measured by way of f1 score. This procedure is repeated so that each fold is used once as the validation set, and the performance metrics across all folds are averaged to produce a more stable and reliable estimate of model quality. The model will be trained with multiple different values for each hyperparameter to determine the best values. Cross-validation improves predictive accuracy, reduces the influence of random variation from a single data split, and enhances the replicability of results (de Rooij and Weeda 2020).

For hyperparameter tuning, cross-validation is used to evaluate different hyperparameter sets. Each hyperparameter or combination of hyperparameters is assessed based on its cross-validated predictive ability score, and the set yielding the best predictions is selected for final model evaluation on the holdout data which was not used in hyperparameter tuning. Once the optimal hyperparameters are identified, the model is trained on the entire training dataset and then tested based on the holdout set (which has not been used during cross validation) to generate final predictions.

Ridge Regression

Ridge regression, like standard logistic regression maximizes the maximum likelihood, but with the addition of a penalty term for the sum of all coefficients in the model, effectively constraining the magnitude of the coefficients (Hastie et al., 2013). This penalty term (λ) is determined via cross-validation. The penalty term with the best average out-of-sample performance is used, this value does change per country. As a result of the penalty term, ridge regression fits the training dataset less tightly because the coefficients are smaller. This penalty makes predictive performance worse in the training set, but it is likely to make better out-of-sample predictions.

$$1: maximize \sum_{i=1}^{N}[ln(1-pi)+yiln(pi1-pi)]-\lambda\sum_{j=1}^{p}\beta_j^2$$

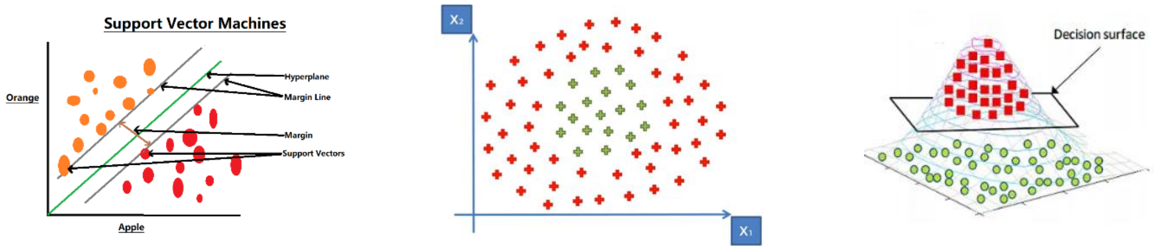
Support Vector Machine

A Support Vector Machine (or SVM) is a model that can distinguish between types of outcomes (Hearst et al., 1998). It does this by drawing a “boundary” through the data space. Figure 3.1 shows such a boundary for the simplest case of two variables on the two axis (plus the outcome inidicated by colour) in two-dimensional space. When additional variables are involved, the boundary line turns into a hyperplane. The area around this boundary to the closest datapoint of each group is called the margin and has to be maximized. The closest data points are the so-called support vectors on which the boundary is based, which is where this model gets its name from.

Choice of kernel

When an outcome is not linearly separable (compare Figures 3.1a and b), a kernel function, which is a mathematical formula, is used to transform the data into a higher-dimensional feature space, enabling the support vector machine (SVM) to construct a linear decision boundary (Figure 3.1c). In this study, I employ the radial basis function (RBF) kernel, chosen for its strong predictive performance despite its reduced interpretability (Pan, 2023). The RBF kernel maps predictors into an infinite-dimensional space, with its flexibility governed by two critical parameters: gamma (γ), which controls the radius of influence of individual data points, and C, the regularization parameter that balances margin width against classification error. A high gamma yields complex, localized decision boundaries, risking overfitting, while a low gamma produces smoother, more generalized boundaries that may underfit (sci-kit learn, n.d.). Conversely, a high C prioritizes perfect training classification (narrow margin, potential overfitting), whereas a low C tolerates misclassifications to widen the margin and improve generalization. Other SVM hyperparameters include the kernel degree for polynomial kernels, class weights for imbalanced data, and the tolerance parameter for optimization convergence, though these have less impact on RBF kernel performance. The optimal γ and C are determined through cross-validation.

Figure 3.1:
(a) A linearly separable distribution with a standard Support Vector Machine (SVM) decision boundary (Janbandhu, 2024);
(b) a non-linearly separable distribution (Saxena, 2021);
(c) transformation into a higher-dimensional space with a non-linear decision surface (Anshul, 2025).



Extreme Gradient Boost

The third model used in this thesis is Extreme Gradient Boost or XGBoost (Chen & Guestrin, 2016). This model uses a large number of decision trees, which will be explained first.

Decision Tree

A decision tree is a predictive model that divides a dataset into smaller and more homogeneous subgroups in order to make classifications or predictions. Given a dataset with a binary outcome, the decision tree algorithm begins by splitting the data into two groups. This initial split is made based on one of the input features and a chosen threshold, which is a specific value that best separates the data into distinct groups (whether a group is made out of either mostly 0's or 1's). If for example a medicine has a positive effect on all women but not on one man, then gender would be great split.
To determine whether a split is useful, the model uses a measure of impurity such as the Gini impurity score (see Formula 1). This score reflects how mixed the classes (either 0 or 1)

are within a group. A lower Gini score means the group is more “pure,” meaning that most of its elements belong to the same class.

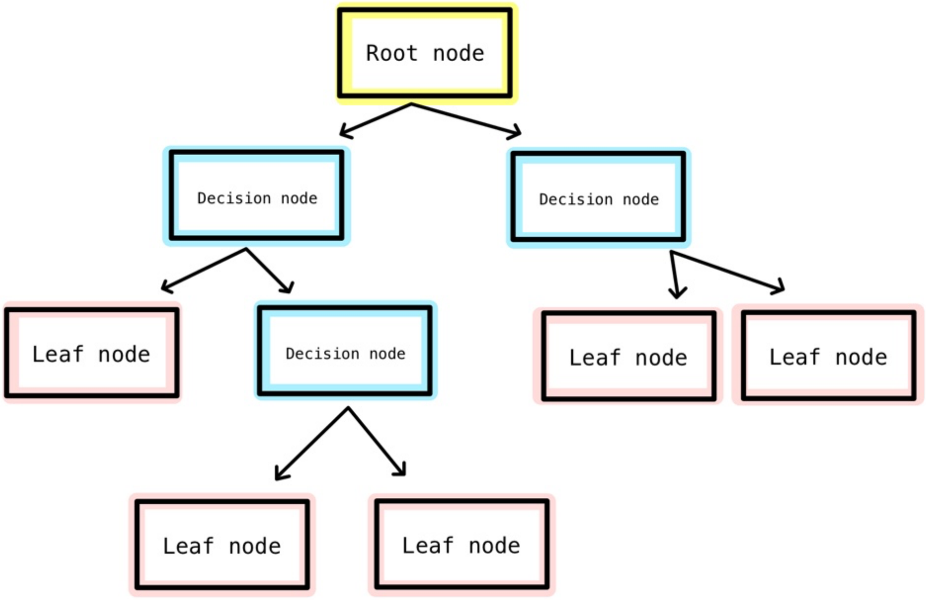
Each resulting subgroup is then evaluated in the same way. If further splitting leads to a meaningful reduction in impurity, the group is split again using a new threshold. This process continues until no further beneficial splits can be made. The final groups, which are no longer split, are called leaves. These are the endpoints of the decision tree (see Figure 3.2).

$$2: Gini\ impurity\ score = 1 - (probability\ of\ “1”)^2 - (probablity\ of\ “0”)^2$$

This often results in decision trees where some paths, also called *branches*, are deeper than others. Predicting the class of new observations is done in the following way: starting at the top of the tree, the observations are routed through the branches based on the learned threshold values. When an observation reaches the end of a branch, known as a *leaf*, a prediction is made based on the training examples that also ended up in that leaf. Because this method does not rely on a linear relationship between variables, it is more flexible in capturing complex patterns.

The total number of subgroups, and the minimum number of observations allowed per leaf are all hyperparameters that can all be determined. Limiting the complexity of a decision tree is often done to prevent overfitting.

Figure 3.2: Graphical representation of a Decision Tree (Kim, 2022)



Many of the better performing models build many decision trees, which outperform the development of singles trees. Therefore, models are used that build on the decision tree principle. Examples include: AdaBoost, Gradient Boost, Random Forest, and Extreme Gradient Boost, also known as ensemble methods. These models use the fundamentals of a decision tree to make better predictions by adding multiple decision trees in a multitude of ways.

Application of Decision Trees in XGBoost

Extreme Gradient Boosting (XGBoost) is a machine learning algorithm that builds an ensemble of decision trees in a sequential manner. It is particularly well-suited for tabular data (which is data organized in rows and columns), where it consistently outperforms more complex models such as deep neural networks (Grinsztajn et al., n.d.). XGBoost operates by constructing multiple decision trees, where each new tree aims to correct the errors (residuals) made by the previous one.

The process begins with a simple prediction—typically a constant value of 0.5 for all observations. The difference between this initial prediction and the true outcome forms the residuals, which reflect the prediction errors. A new decision tree is then trained to predict these residuals in the same training data. Each subsequent tree focuses on the remaining errors, gradually improving the overall model fit. This iterative approach allows XGBoost to approximate complex patterns in the data.

To prevent overfitting, XGBoost incorporates several regularization techniques. One of them is a similarity score used to evaluate potential splits in the tree, analogous to the Gini impurity in standard decision trees. The similarity score incorporates a penalty term (λ) to discourage overly complex splits:

$$3{:}\, Similarity\ score = \frac{(\Sigma\, residual\ i)^2}{\Sigma\,[previous\ probability\ i * (1 - previous\ probability\ i)] +\ \lambda}$$

This score is used to decide whether a split improves model performance. A gain score quantifies the improvement obtained by a split, calculated as:

$$4{:}\, Gain\ score = LEFTsimilarity + RIGHTsimilarity - ROOTsimilarity$$

This formula uses the similarity score of the original leaf, denoted by the root similarity, and the new similarity scores denoted by the left and right similarity. If the gain score exceeds a certain threshold, denoted by γ (gamma), the split is accepted and the tree continues to grow. After each tree is built, its output is multiplied by a learning rate (ε), typically set to 0.3, and added to the previous model. This lessens the contribution of each tree and reduces the risk of overfitting.

The process continues until either a specified number of trees is reached or the model performance on validation data no longer improves. The final model is a weighted sum of all individual trees, each one trained to model the residuals of the previous ensemble.

XGboost hyperparameters

A key hyperparameter is the maximum tree depth, which imposes a strict limit on how deep each decision tree can grow. A greater depth permits the construction of more complex trees, which may improve in-sample performance but also increase the risk of overfitting. Conversely, a tree that is too shallow may not capture more complex patterns in the data, leading to underfitting.

Another important parameter is the minimum number of observations per leaf, which also influences tree complexity. This threshold determines the smallest allowable leaf size, and its ideal value depends on the characteristics and size of the dataset. Notably, this parameter may contribute to the formation of asymmetrical trees, in which certain branches extend deeper than others.

The gamma (γ) parameter plays a pivotal role in the decision-making process of the tree. It determines whether a given branch is allowed to split further by evaluating the gain score. If the gain score falls below the specified gamma value, the branch is designated as a terminal leaf. Thus, a higher gamma value leads to more conservative, simpler trees, whereas a lower value encourages more complex, granular splitting.

Additionally, the penalty term (λ) affects how tightly a model fits the training data. When λ is set to zero, it exerts no influence. As λ increases, achieving high similarity scores becomes more difficult, resulting in reduced model complexity and greater resistance to overfitting.

Lastly, the learning rate (ε) controls the extent to which each individual tree contributes to the overall model prediction by minimizing individual contributions. While a standard value of 0.3 is commonly used, lower values tend to yield higher accuracy by requiring more iterations, albeit with increased computational demands. Higher learning rates, on the other hand, allow the model to converge more quickly on the correct prediction in theory but often at the expense of overshooting optimal values, which can hinder predictive performance.

In addition to these core hyperparameters, several others are relevant to controlling complexity and improving generalization. The subsample parameter determines the proportion of training instances used to build each tree, introducing randomness that helps prevent overfitting. Similarly, colsample_bytree controls the fraction of features randomly selected for each tree, while colsample_bylevel does so for each tree level. The total number of trees is governed by n_estimators, and in cases of class imbalance, the scale_pos_weight parameter adjusts the loss function to give more importance to the minority class. These parameters are all determined by cross validation, creating the predictive performance.

Interpretation

Interpreting an Extreme Gradient Boost model is difficult due to the aforementioned complexity. One can examine the thresholds calculated by the individual decision trees, but as the number of trees increases, so too does the number of thresholds. This, combined with a lack of clear interpretability of the thresholds themselves, makes substantive interpretation of the model difficult. Although direct interpretation of these results is no necessary for this thesis.

Model evaluation for the three models

At present, there is no universally accepted standard for evaluating predictive models within the scientific community. This study adopts out-of-sample evaluation, using 75 percent of the data for training and the remaining 25 percent as out-of-sample data for evaluating final model performance. While this is a widely used approach, there are multiple evaluation metrics available, each suited to different objectives and contexts. These differences require researchers to make subjective choices about what kind of performance is most relevant for their specific research goals.

In light of this, four distinct evaluation metrics are used in this study: accuracy, ROC-AUC, F1 score, and Brier score. Each metric captures a different aspect of model performance. Unlike traditional hypothesis testing, these measures do not yield binary outcomes such as "significant" or "not significant." Instead, they provide a useful measure of the strength of an entire model.

Accuracy

The first metric used to evaluate predictive performance is accuracy, defined as the proportion of correctly predicted outcomes over the total number of predictions (Zheng, 2015). Accuracy is intuitive and easy to communicate, making it a commonly used metric in predictive modelling. However, it has notable limitations in the context of class imbalance—when the event of interest occurs either relatively rarely or frequently. In such cases, models that consistently predict the majority class may still yield deceptively high accuracy. This is relevant in our study, where the fertility outcome of having a child in 3 to 4 years occurs relatively infrequently (e.g., roughly 28%). Consequently, although accuracy will be reported to provide general context, it should be interpreted with caution and supplemented by other, more informative performance metrics.

$$Accuracy = \frac{True\ positive + True\ negative}{True\ positive + False\ positive + True\ negative + false\ negative}$$

ROC-AUC

The second metric employed is the Receiver Operating Characteristic – Area Under the Curve (ROC-AUC), which quantifies the model's ability to distinguish between classes. It is calculated by plotting the true positive rate (TPR) against the false positive rate (FPR) at various classification thresholds and computing the area under this curve (Zheng, 2015). The TPR represents the proportion of correctly identified positive cases (i.e., individuals who had a child) out of all actual positive cases, while the FPR denotes the proportion of negative cases incorrectly classified as positive, relative to all actual negative cases. By summarizing performance across all possible thresholds, the ROC-AUC metric mitigates the influence of class imbalance, which is particularly relevant in the context of fertility prediction, where positive outcomes are relatively rare. Despite its usefulness, the ROC-AUC is more abstract than metrics such as accuracy, which can make its interpretation less intuitive. The resulting value ranges from 0 to 1, where 0.5 indicates random performance and 1.0 denotes perfect discrimination.

Brier-score

The third metric used is the Brier score, which measures the performance of probabilistic predictions. It is calculated as the mean squared difference between the predicted probability assigned to an outcome and the actual binary outcome (Redelmeier et al., 1991). Unlike classification metrics that require thresholding of predicted probabilities into binary outcomes, the Brier score retains the continuous nature of probability estimates. This makes it particularly useful for evaluating the calibration of a model—how closely predicted probabilities reflect actual outcomes. A lower Brier score indicates better predictive performance, with values ranging from 0 (perfect prediction) to 1 (completely inaccurate prediction for a binary outcome). In the context of fertility prediction, this metric provides insight into how confident the model is in its predictions, making it a valuable complement to threshold-dependent measures such as accuracy and ROC-AUC.

$$Brier - score = \frac{1}{n} * \Sigma(probability\ i - observed\ class\ i)^2$$

F1 score

The final metric included is the F1 score, which serves to better assess the impact of class imbalance on model performance. The F1 score is the harmonic mean of precision and recall, two complementary metrics that capture different aspects of prediction quality (Zheng, 2015). Precision reflects the proportion of predicted positive cases that were indeed positive (i.e., proportion of individuals predicted to have a child that indeed had a child), while recall indicates the proportion of actual positive cases that were correctly identified by the model (i.e., proportion of correct predictions who had a child relative to everyone in the sample who had a child or true positive rate). By combining these two metrics, the F1 score balances the trade-off between over-predicting positive outcomes (which inflates false positives) and under-predicting them (which inflates false negatives). This makes it a particularly suitable metric in scenarios with class imbalance, such as who has a child in a narrow timeframe, where the positive outcome is relatively rare.

$$Precision = \frac{True\ positive}{True\ positive + false\ positive}$$

$$Recall = \frac{True\ postive}{True\ postive + false\ negative}$$

$$F1\ score = \frac{2 * Precison * Recall}{Precision + Recall}$$

Variable importance

Parameter estimates for most machine learning models are difficult to interpret because there is a large amount of parameters to consider with more complex models and these are often not intuitive. To address this, several interpretable ML techniques have been developed. One relatively simple yet effective approach is variable importance (Breiman, 2001). This method works by randomly permuting a single predictor in the holdout data, making predictions with this modified dataset, and then measuring how much the model's accuracy declines. A greater drop indicates that the variable is more important. This technique allows us to identify which predictors carry the most weight in each model and how this differs across countries.

Dataset

The data used in this study is derived from the Generations and Gender Survey (GGS) (Fokkema et al., 2016), a large-scale, cross-national, longitudinal dataset designed to improve understanding of demographic behaviours and family dynamics. The GGS is a cross-national panel survey on life-course and family dynamics of individuals aged 18-79 years, with follow-up surveys at three and six-year intervals to track how people's lives unfold. The aim of this programme is to provide researchers with up-to-date, internationally comparable family demographic data, coordinated by a team based at the Netherlands Interdisciplinary Demographic Institute (NIDI).

Over the past 20 years, the GGP has collected survey data in 25 countries in Europe and beyond. For this study, 11 countries were selected from the 18 countries where GGS has been conducted: Austria, the Netherlands, Hungary, Poland, Russia, Bulgaria, Georgia, Lithuania, Germany, France, and the Czech Republic. Only countries for which two waves of data were available were included, as the dataset's panel structure allows us to measure fertility outcomes—specifically the occurrence of childbirth between survey waves, typically spaced 3 to 4 years apart depending on the country. In total, these 11 countries covered 37.708 individuals who were present in both waves (see also Appendix 1).

Survey Design and Methodology

The GGS questionnaire, created by an international team of social scientists, captures comprehensive individual-level variables essential for fertility analysis, including age, education, employment, relationship status, fertility intentions, and number of children. The infrastructure includes a register-linked survey of adults to which a contextual database covering their adult lives is matched. The GGP's overarching goal is to provide data infrastructure for improved understanding of the causes and consequences of dramatic declines in fertility and changes in partnership behavior in Europe and other affluent countries. The survey's longitudinal design enables researchers to examine both retrospective life histories and prospective behavioural outcomes, making it particularly valuable for studying intention-behaviour consistency in fertility decisions.

After its first round of face-to-face implementation, the second round has been implemented on the web, though this methodological shift has required careful assessment of data quality. Recent validation studies have verified the accuracy of fertility histories by comparing GGS data with population-based estimates from the Human Fertility Database (HFD) and the United Nations Population Division, confirming the reliability of the web-based data collection approach (V. A. Leocádio et al., 2023).

Choice of variables

For comparability across all models and countries, I have chosen 13 variables based on the meta research of Balbo and colleagues (2012) and the availability in the GGS dataset. These were:

1. Education: highest achieved level of education. In order to make this variable comparable across the countries the original coding was changed, and the final

variable included three levels: 1) pre-primary, primary, or lower secondary education, 2) upper-secondary and post-secondary non-tertiary education, and 3) higher education.

Education is a major driver of fertility behaviour and differently so for men and women (Beck et al., 2024). For example, highly educated women are more likely to be childless, whereas the opposite is true for men (Jalovaara et al., 2018).

2. Number of marriages: the total number of marriages that a person had, including the current one.
3. Marriage: whether a person is currently married or not
4. Current partnership status: whether a person currently has a co-habiting partner, a non-cohabiting partner, or no partner.
5. Satisfaction with the relationship measured from 0 (not at all satisfied) to 4 (completely satisfied), plus “no partner”.

Relationship context is a central determinant of fertility. Being in a stable union, especially marriage, is traditionally associated with higher fertility across most societies (Kuang et al., 2025). While cohabitation is increasingly common and serves as a setting for childbearing, marriage remains a stronger predictor of fertility intentions and behavior, particularly for higher-order births. The number of marriages captures complex life-course dynamics, such as remarriage, which can influence fertility through blended family motivations and partnership renewal (Elleamoh & Dake, 2019). Additionally, relationship satisfaction plays a critical role in childbearing decisions. Individuals in higher-quality relationships are more likely to desire and plan for children, while those in less satisfying or unstable relationships may postpone or forego fertility (Testa & Basten, 2014).

6. Current employment status (employed, unemployed/retired/currently on leave).
7. Occupational category: an ISCO category of current occupation (if employed) or previous occupation (if currently not employed)

Employment status affects fertility through financial stability and work-life balance. Employment can act as a barrier to fertility for women or a prerequisite depending on welfare services in a country (Kreyenfeld, 2009). For men, employment is more consistently viewed as a prerequisite for fatherhood, as stable income and job security are often seen as necessary to support a family (Tragaki & Bagavos, 2014).

8. Health: whether a person has a long-standing illness or chronic condition

Health status is a critical factor in fertility behavior. Individuals with chronic illnesses or long-term health conditions may experience reduced biological capacity to conceive or raise children and may also choose to delay or avoid childbearing due to health-related concerns about parenting capacity and longevity. Health can also affect the stability of partnerships and employment, both of which further influence fertility intentions and outcomes (Lazzari & Beaujouan, 2025).

9. Religion (none, Orthodox, Catholic, other Christian, other religion)

Religious beliefs often guide values related to marriage, contraception, and family size. More religious individuals or communities tend to have higher fertility intentions and outcomes, influenced by traditional gender roles and family norms. Religious affiliation thus remains a key socio-cultural factor in understanding fertility behavior, even in secularizing societies (Skirbekk, Kaufmann, & Goujon, 2010).

10. Fertility intentions: whether a person wants to have a(nother) child in the next 3 years

Stated fertility intentions are one of the strongest predictors of short-term fertility behaviour. Although not all intentions are realized, they provide valuable insight into motivational and contextual factors underlying future childbearing decisions. Fertility intentions are especially useful when analysed alongside age, partnership status, and employment (Schoen et al., 1999).

11. Age of the youngest child. (if a person Is childless, they will be coded as -1)

The age of the youngest child helps to capture spacing patterns and progression to higher-order births. Parents with very young children often delay further childbearing, whereas those whose youngest child is older may be more likely to plan another child. It is a dynamic factor reflecting life course timing (Van Bavel & Różańska-Putek, 2010).

12. Gender.

Fertility behavior and the social meaning of parenthood differ significantly by gender. Women's fertility is more biologically time-limited and more directly affected by employment, relationship dynamics, and policy supports. Men's fertility tends to occur later and is more closely linked to stable employment and union formation. Gender thus serves as a key moderating variable in fertility analysis (Goldscheider, Bernhardt, & Lappegård, 2015).

13. Age.

Age is a foundational variable in fertility research. It reflects both biological constraints and social expectations around parenthood. Younger individuals are more likely to intend future childbearing, while older individuals face declining fecundity and narrower opportunities for family formation. Age also interacts with other key variables such as education, partnership, and employment (Mills et al., 2011).

While it is true that using a relatively limited set of predictor variables may constrain the performance of more complex models such as XGBoost and Support Vector Machines (SVM), this trade-off was considered acceptable in order to ensure a fair and consistent comparison across models and builds on existing theoretical work. Prioritizing comparability over maximizing predictive accuracy was deemed more appropriate for the purposes of this analysis.

14. New child

We measured a couple having a new child by the calculating the difference between the amount of children a person had between the first and the second wave of the GGS dataset. When this number is higher than 0, it was denoted as a 1 otherwise it is a 0.

Sample selection

After the age of 45, very few people have children, mainly due to biological limitations (Eijkemans et al., 2014). For this reason, individuals over 45 were excluded from the analysis. This also ensures a clearer focus on the sociological factors influencing fertility behavior among those still of childbearing age and helps partially reduce variation in age distributions across countries. Additionally, respondents with missing values on the outcome variable were excluded (~N = 7150) as such cases provide no usable information for model estimation and can impair the accuracy and validity of predictive analysis

Handling missing values

Some models, such as XGBoost, can automatically handle missing values by treating them as meaningful information. For instance, if a survey question is left blank, the missing response might actually be useful for prediction, perhaps indicating a specific behaviour or trait. However, in this analysis, we chose to manually impute missing values for all variables. While this makes the models more comparable (since not all algorithms handle missing data the same way), it also means we lose XGBoost's natural ability to learn from missing data patterns. This trade-off was intentional for consistency, but it's important to recognize that imputation removes one of XGBoost's strengths.

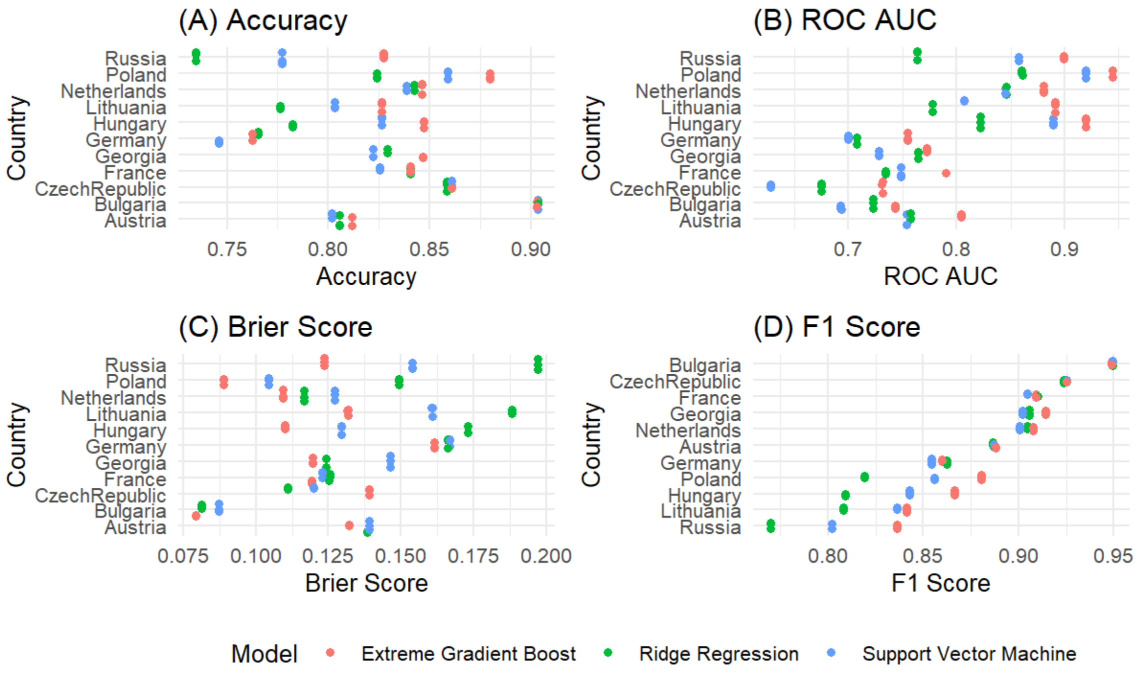
To preserve as much data as possible, I imputed all missing numeric variables using the median and all categorical variables using the mode, both calculated from the training set. The number of missing values for each variable is detailed in Appendix A. This approach prevents the substantial case loss that would occur from listwise deletion, thereby maintaining sample size and enhancing the reliability of predictive modelling. This can exacerbate possible bias in a dataset but for datasets of this size it is less likely.

An exception to this strategy is the variable for age of the youngest child. In approximately 90 percent of cases, a missing value for this variable reflects childlessness rather than non-response. For this reason, it was not imputed using the standard procedure. Instead, these missing values were treated as meaningful and handled separately, by assigning -1 as an indicator to capture the presence or absence of children.

Chapter 4: Results

We report all four key evaluation metrics, accuracy, precision, recall, and F1 score, to provide a comprehensive assessment. Since models can perform well on some metrics but poorly on others, it is essential to identify these variations and explore potential explanations for why differences occur across countries.

Figure 4.1: for each country and model (a) accuracy, (b) ROC AUC, (c) Brier score, and (d) F1 score. Performance is compared across three models: Ridge Regression (green), Support Vector Machine (blue), Extreme Gradient Boosting (red). Each model was run three times.



There is considerable variation across each of the four metrics in terms of performance. Accuracy ranges from 0.73 To 0.93. These high accuracies are partly driven by the highly imbalanced datasets that imply that even a model that predicts that no one would have a child would do well. The ROC AUC scores vary from 0.63 to 0.94; the former considered poor performance and the latter excellent performance (Çorbacıoğlu & Aksel, 2023). The Brier score ranges from 0.076 to 0.180. Taking the square root of 0.076 equals 0.27, which can be interpreted as the predicted probability on average being 0.27 off from the correct class. The F1 scores ranges from 0.73 to 0.94.

No country performs consistently well on all of evaluation metrics. However, some countries, such as Bulgaria, Poland, and the Netherlands, tend to perform relatively well overall, even if they fall short on specific metrics. Bulgaria stands out with the highest F1 score, strong accuracy, and the lowest Brier score. This indicates that the model performs well in both classification accuracy and probability calibration for this country. It does have comparatively low ROC AUC which could have been caused by class imbalance.

Poland shows consistently high scores across all four metrics, suggesting a well-balanced model performance. The Netherlands also scores highly on accuracy, Brier score, and ROC AUC, but its lower F1 score indicates some difficulty in balancing precision and recall, even if the overall probability estimates and rankings are strong.

In contrast, countries like Russia, Germany, and Georgia generally show weaker performance, especially on F1 score and ROC AUC. This may reflect challenges in capturing patterns in the data, poor class balance, or issues related to data quality in those countries.

Czech Republic presents a particularly interesting case. It performs reasonably well on accuracy, F1 score, and Brier score, but has a notably low ROC AUC. This could mean that the model makes correct binary predictions but struggles with ranking cases correctly by predicted probability. One possible explanation is that most predicted probabilities are clustered near the decision threshold, which can result in good accuracy but poor ranking performance. Alternatively, this pattern might stem from class imbalance or limited variance in the predicted scores.

These results highlight why it is important to evaluate models using multiple metrics. Each one captures a different aspect of model behaviour, and relying on only one may obscure important performance trade-offs.

Comparing Models

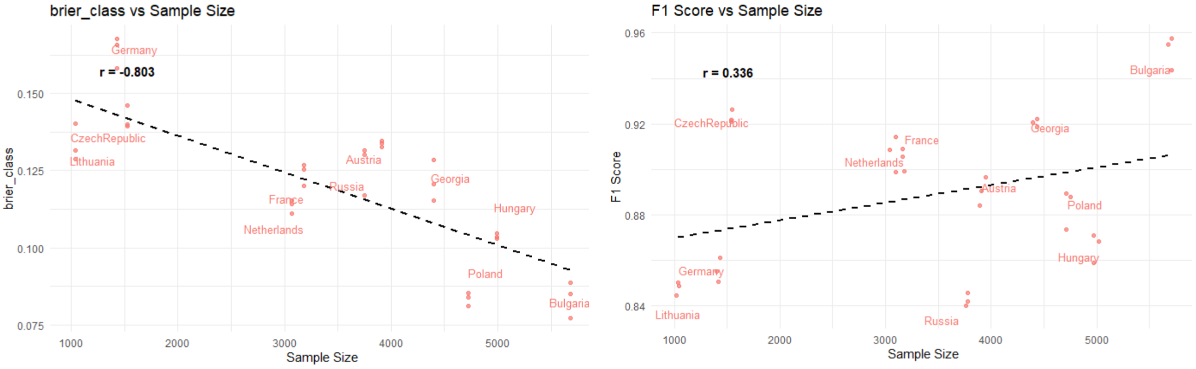
Across all countries and evaluation metrics, XGBoost performs best overall, followed by support vector machines and ridge regression. XGBoost shows clear advantages in both classification performance and probability calibration, particularly in countries with more balanced class distributions. It generally produces higher ROC-AUC and F1 scores, and lower Brier scores, indicating strong discriminative ability and well-calibrated predictions. Support vector machines and ridge regression tend to yield similar results, though they vary slightly depending on the country. In some cases, these models perform competitively, but they are less flexible in capturing complex relationships in the data. Accuracy scores are highest in countries with strong class imbalance, but this mostly reflects a bias toward predicting the majority class and should be interpreted with caution. Overall, while all three models offer some predictive value, XGBoost provides the most consistent and reliable performance across different settings and evaluation criteria.

We ran each model three times to assess variability in model performance depending on how the data is split into training and holdout data. We observed consistent results across all three iterations and models. The relative performance ranking remained stable—for instance, if XGBoost outperformed SVM, which outperformed penalized regression, this pattern held across all iterations.

Assessing Influential Variables context variables

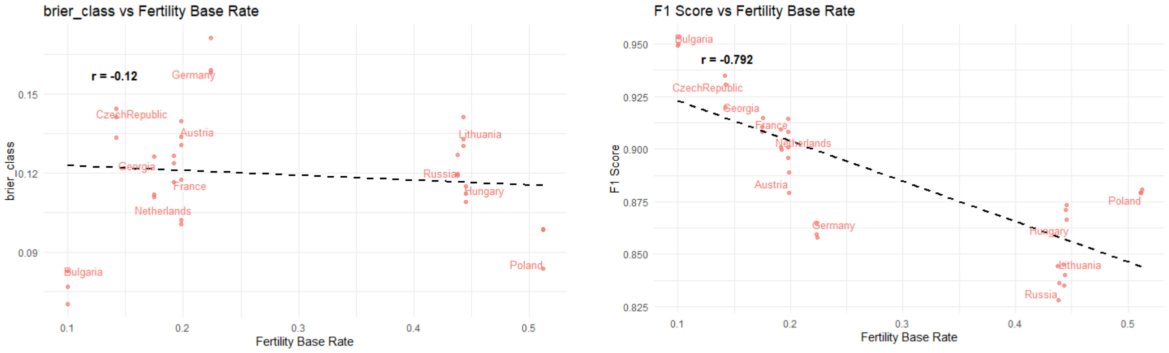
When evaluating predictive performance, it is important to take country-specific factors into account. We consider sample size and the base fertility rate to be potential constraints on model performance. We also examine the length of the social fertility period and societal attitudes toward childlessness to explore how cultural and social factors might influence predictability. In the analysis that follows, we focus on the Brier score and F1 score. These two metrics offer complementary insights: the Brier score evaluates how well the predicted probabilities align with actual outcomes, while the F1 score captures the balance between precision and recall. This is particularly important given the class imbalance that often characterizes fertility data. This focus allows for a nuanced assessment of both the reliability and practical utility of the predictions. The same scatterplots for accuracy and ROC-AUC are provided in appendix 2 for completeness. In these analyses, for comparability purposes and clarity, we only focussed on XGBoost models (which performed best) and we took the average of the three iterations as predictive performance. In the supplement we also show these results for the other types of models.

Figure 4.2: A) brier score against sample size B) f1- score against sample size



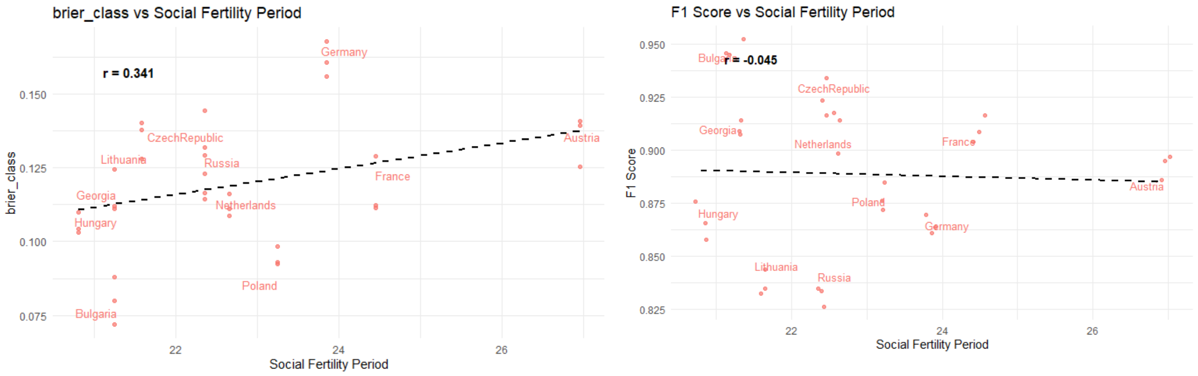
There appears to be a mild positive relationship between sample size and model performance. The correlation with F1 scores shows a weak positive association (r = 0.170), while Brier scores demonstrate a strong negative correlation (r = -0.803), indicating that larger sample sizes are associated with better performance on both metrics. This pattern aligns with machine learning expectations: larger datasets typically provide more training examples, leading to better model generalization and more reliable probability estimates. Countries with larger samples allow models to learn more robust patterns, resulting in both improved classification accuracy and better calibrated predictions.

Figure 4.3: A) brier score against fertility base rate B) f1- score against fertility base rate



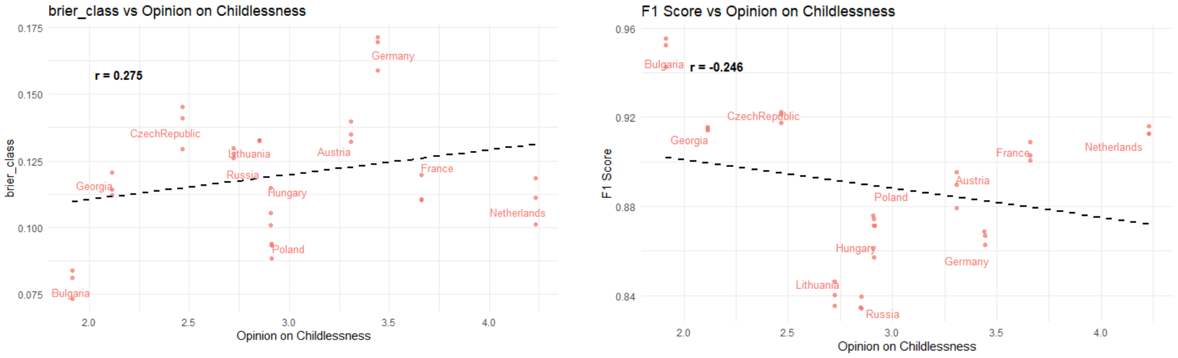
The relationship with fertility proportion reveals the strongest correlations in our analysis, though with opposing directions across metrics. F1 scores show a strong negative correlation (r = -0.792), while Brier scores show a mild negative correlation (r = -0.120). This means that countries with higher fertility rates a lower F1 scores and mildly better Brier scores. However, this pattern requires careful interpretation. The strong negative correlation with F1 scores may partly reflect the metric's behaviour when base rates are high - models that frequently predict "zero" can still achieve reasonable F1 scores even without truly learning meaningful patterns. The Brier score correlation is much weaker and the negative correlation means the opposite as a lower Brier score means a better performance, suggesting that the relationship between fertility rates and prediction difficulty is not consistent across metrics and may be an artifact of how F1 scores behave with imbalanced data rather than indicating that higher base rates genuinely make prediction more challenging

Figure 4.3: brier score and f1- score against social fertility period



The social fertility period shows minimal association with model performance across both metrics. F1 scores display a very weak negative correlation (r = -0.045), while Brier scores show a mild positive correlation (r = 0.341). The scatterplots reveal no clear linear trends, suggesting that the length of socially acceptable fertility periods has negligible impact on model predictability. This finding indicates that while fertility timing norms may influence individual reproductive decisions, they do not systematically translate into better or worse model performance across different national contexts.

Figure 4.4: brier score and f1- score against opinion on childlessness



Societal attitudes toward childlessness show weak but consistent correlations with model performance. F1 scores demonstrate a mild negative correlation (r = -0.246), while Brier scores show a mild positive correlation (r = 0.275), suggesting that more accepting attitudes toward childlessness are associated with slightly worse performance on both metrics. However, visual inspection of the data suggests a potential non-linear relationship where countries with either very accepting or very disapproving attitudes demonstrate better overall performance than those with moderate positions. This pattern implies that polarized societal norms - regardless of their direction - may create more predictable behavioural patterns than ambivalent cultural contexts.

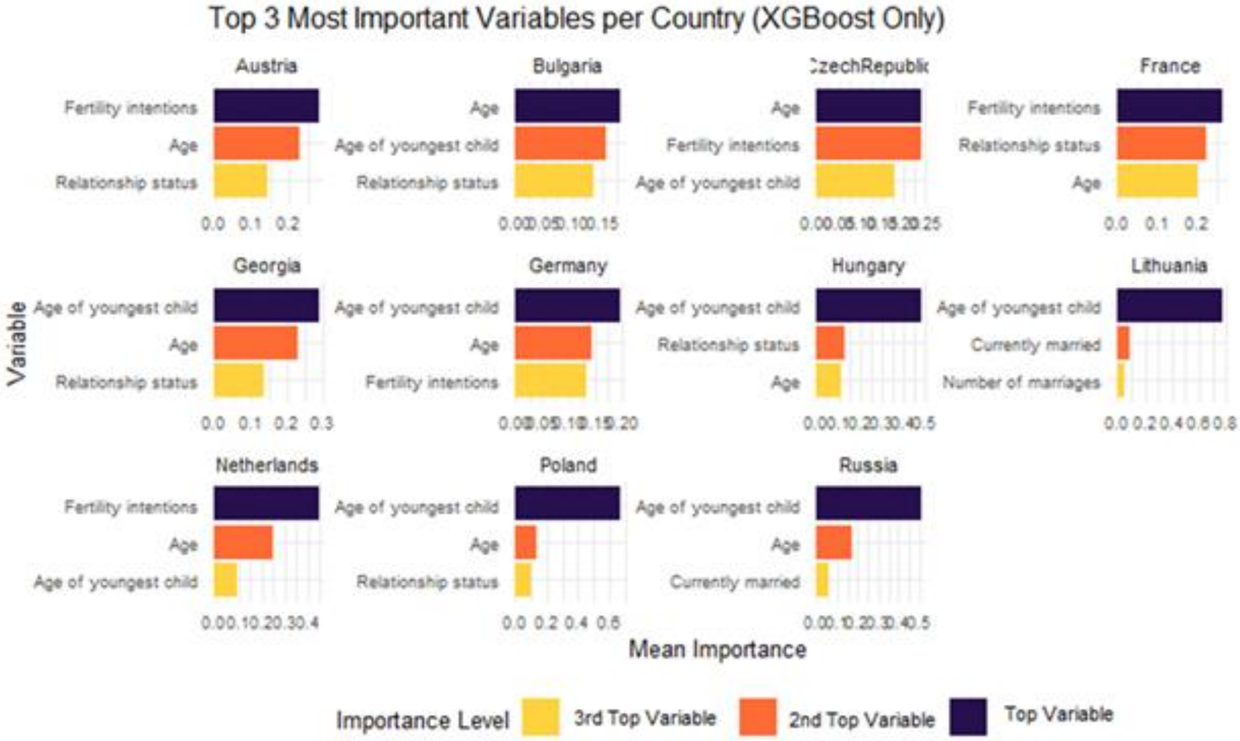
To test these patterns more formally, we created a linear regression model predicting the brier score and the f1-score using four predictors: sample size, fertility proportion, social fertility period, and opinion about childlessness. The exact values of these models can be found in appendix 3 .The results show that only sample size and fertility proportion are statistically significant predictors of performance. Social fertility period and opinion on childlessness did not have a clear independent effect once the other variables were included. This suggests that while cultural attitudes may matter, we found no evidence for their direct effect in this study.

Variable importance across different countries

Comparing variable importance across countries provides valuable insights into how fertility patterns function differently by context. Given that each country requires a visualisation and there are 13 variables available, displaying all graphs would be unwieldy. Therefore, I present only the top three most important variables per country, ranked by their average importance for XGBoost only, as it is the best performing model. Using average ranks helps mitigate the influence of extreme values and provides a more robust foundation for cross-national comparison.

The analysis reveals several compelling patterns. In countries with the most positive attitudes toward childlessness—France, the Netherlands, and Austria—fertility intentions consistently emerge as the most important predictor variable. Conversely, fertility intentions do not appear among the top three predictors in countries such as Russia, Georgia, and Hungary, where the age of the youngest child takes precedence as the primary factor. These findings highlight significant cross-national differences in fertility prediction models and underscore the critical importance of context-specific interpretations when analysing reproductive behaviour across diverse cultural and policy environments.

Figure 4.2: Top 3 most important variables per country



Chapter 5: Conclusion & discussion

This study set out to test the use of different models with the aim of predicting fertility behaviour. Predictability varied substantially across countries. There are several reasons why predictability can vary: sample size (Lundberg et al., 2024), base rate (He & Garcia, 2009), differences in data quality (Lundberg et al., 2024), and substantive reasons such as cultural variation that shape fertility behaviour. These issues make it difficult to draw clear conclusions and point to the need for more standardized data and modelling practices in cross-national research.

Direct comparison of our fertility prediction results with other machine learning studies in demography is challenging due to fundamental differences in research design, outcomes, and metrics. Arpino et al. (2021) achieved 70% accuracy predicting union dissolution in Germany using Random Survival Forests, while Stulp et al. (2023) focused on fertility preferences rather than actual fertility behaviour and did not report comparable accuracy metrics. Our fertility models show accuracy ranging from 0.73 to 0.93 and ROC AUC scores from 0.63 to 0.94 across European countries. However, these apparent differences may reflect distinct prediction tasks (union dissolution vs. fertility behaviour vs. fertility preferences), different datasets (German Socio-Economic Panel vs. GGS), varying outcome definitions, and different baselines rates rather than genuine differences in predictability. What these studies do consistently demonstrate is that machine learning approaches, whether Random Survival Forests or XGBoost, substantially outperform traditional statistical methods in demographic prediction tasks, though establishing meaningful benchmarks for "good" predictive performance in demography remains an ongoing challenge.

Effects of social pressure on predictability

We hypothesised that a shorter social reproductive lifespan would make fertility behaviour more predictable, because a shorter lifespan would make variables like age more potent. The results we found does not support this hypothesis. Instead, it suggests that factors beyond the duration of socially acceptable reproductive years may play a larger role. Which could be caused by the limited variation in reproductive lifespans.

We also did not find a support for our second hypothesis which stated that a worse opinion about childlessness would make fertility behaviour more predictable, as a bad opinion about childlessness would put more pressure on people to have children at a certain age. Methodological and data differences did account for difference in predictability, particularly sample size and fertility proportion had statistically significant impact on predictive performance. This shows that methodological factors and basic demographic conditions explain more of the cross-national variation in fertility predictability than the cultural or social variables we initially expected to be important. We are hesitant to conclude that cultural differences do not impact predictability, but is likely that cultural differences have a smaller impact relative to methodological differences.

Future cross national comparison in predictability are likely to suffer from the same methodological difficulties. These issues could be partially solved in the future by a more unified approach to collecting data or different ways of resampling collected data.

Fertility intentions more important in Western countries

Cross-country differences were observed in which variables were most important in predicting fertility, suggesting cultural differences certainly exist. Countries which had less social pressure to have children in a certain timespan, most notably relied on fertility intention as a primary predictor of fertility outcomes. This reliance on fertility intentions may explain why none of the variables based on social pressure had a significant impact in our lineair regression, though this relationship requires further investigation.

The predictive models which put emphasis on fertility intention demonstrated minimal dependence on factors conventionally associated with social pressure and structural constraints, such as marriage rates and formal partnership arrangements (Ellemann & Dake, 2019). Instead, the models emphasized individuals' explicitly stated reproductive intentions (Schoen et al., 1999). The prominent role of fertility intention in these models could have compensated for the structural factors that were theoretically expected to enhance predictive accuracy under conditions of strong social pressure. Which could be a reason there was no support for our two hypotheses.

Despite this overall pattern, a notable regional variation emerges in the data. Eastern European countries consistently demonstrate greater reliance on marital and relationship status variables rather than fertility intention measures in their predictive models. This regional distinction suggests meaningful variation in the underlying sociocultural mechanisms that drive fertility decision-making processes (Sobotka, 2004; Frejka, 2008; Billingsley, 2010). The divergent predictive patterns observed across these countries may reflect differing institutional contexts, policy environments, and cultural frameworks that shape reproductive behaviour differently across European regions. These findings confirm long held beliefs about Post Communist Eastern European nations as more pro-natal as a result of the hardship of the collapse of the Communism. This context made remaining childless while in a position to have children more difficult (Thorne, 2005).

XGboost as best performing model

Cross-model differences reveal a consistent pattern where XGBoost models outperformed the other two types in nearly every country, with only a few exceptions, despite efforts to level the playing field such as removing missing values and limiting the number of variables used across all models. This superior performance provides strong evidence that XGBoost is a more effective approach for predicting fertility outcomes in Europe, which is in line with current research putting XGBoost as most suitable to tabular data which is used in this thesis (Yarkin Yildiz & Kalayci, 2024). When interpreting variable importance across different model types, it becomes clear that the age of the person and the age of the youngest child are far more important to XGBoost models than to the other two, suggesting a non-linear effect of these age variables on the probability of having a child. This finding is expected because XGBoost, being tree-based, is better able to model non-linear relationships, and such non-linear patterns are intuitive since fertility rates likely peak around certain ages, lending additional support to this understanding of fertility dynamics. These findings confirm the theoretical understanding we have of interbirth intervals having to be either not to short or not to long (Van Bavel & Różańska-Putek, 2010). We also find confirmation in our understanding of age as an obstacle to fertility, in which being either too young or too old may hinder fertility behaviour (Mills et al., 2011).

Data collection and sampling differences

To ensure fair cross-country comparisons, data collection procedures should be as consistent as possible. Unfortunately, this is not always the case for the GGS dataset, where fieldwork protocols differ across countries. Although the GGS provides guidelines based on best practices, there is no enforced standard for how data collection should be conducted. As shown in Table 6.1. substantial variation exists in sampling methodology, including the number of sampling stages, the use and type of stratification, and the timing of data collection. These differences may limit the validity of direct comparisons between countries.

Table 6.1: The differences between countries in sampling method.

Country	# Sampling Stages	Stratification	Frame	Frame Elements	Type of Sampling
Austria	1	YES	Population register	Names	SRS
Bulgaria	2	NO	Area & Census	Dwellings	PPS + SRS

Country	# Sampling Stages	Stratification	Frame	Frame Elements	Type of Sampling
Czech Republic	2	YES	Area	Dwellings	PPS + SRS
France	2	YES	Census & update new dwellings	Dwellings	PPS + SRS
Georgia	2	YES	Census	Names	PPS + SRS
Germany	2	NO	Area (GIS)	Addresses	PPS + SRS
Hungary	2	YES	Area, Settlement	Addresses	PPS + SRS
Lithuania	2	YES	Area	Settlements	RR + RR
Netherlands	1	NO	Area	Addresses	SRS
Poland	2	YES	Census area	Dwellings	SRS
Russian Federation	2	NO	Area	Dwellings	PPS + SRS

Fokkema et al. (2016) provide a comprehensive assessment of GGS Wave 1 and 2 data collection procedures and note that while "the quality of sampling and fieldwork procedures of the GGS is generally good," cross-national differences in implementation may affect comparability. Their analysis found that after weighting, the data were generally representative in terms of age, gender, region, and household size, but showed greater variation in representativeness for marital status and educational attainment across countries.

The variation in sampling frames—ranging from population registers to area-based approaches—represents a particular challenge for cross-country comparisons. While the impact of these methodological differences on our specific predictive models was beyond the scope of this study to quantify, they represent an important caveat when interpreting cross-country variations in model performance.

Difficulty of stating conclusions due to lack of industry standards

While this thesis has demonstrated that predictive machine learning models can offer new and interesting insights, it remains difficult to reach clear conclusions and communicate them effectively due to the absence of established industry standards for what constitutes good prediction in demographic research. Although this has allowed for a more nuanced discussion of each country, it also makes it harder to clearly present the findings.

The field currently lacks consensus on key methodological questions: what accuracy thresholds should be considered satisfactory for different types of demographic predictions, how to standardize model validation procedures for cross-country comparisons, and how to interpret feature importance rankings across different cultural and institutional contexts. Without established benchmarks, it becomes challenging to determine whether observed differences in model performance across countries reflect meaningful demographic variations or simply methodological artifacts.

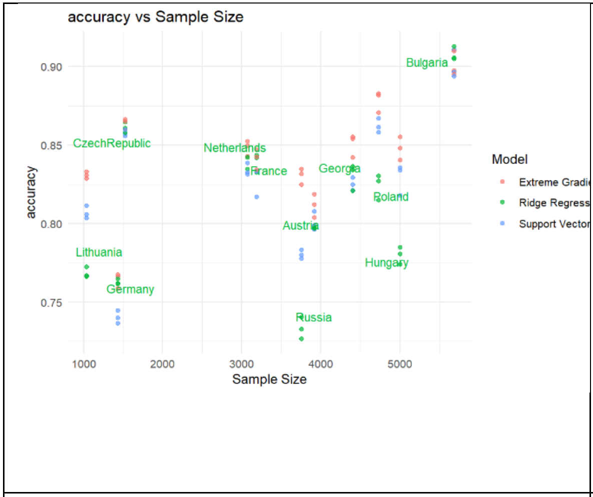
I have personally found it challenging to discuss the results of this thesis with other scientists. This difficulty is likely to decrease as prediction-focussed research is increasing and standards become more widely accepted.

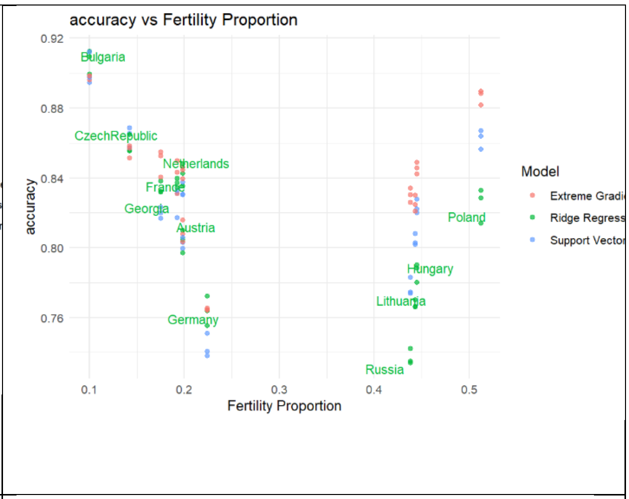
Appendix

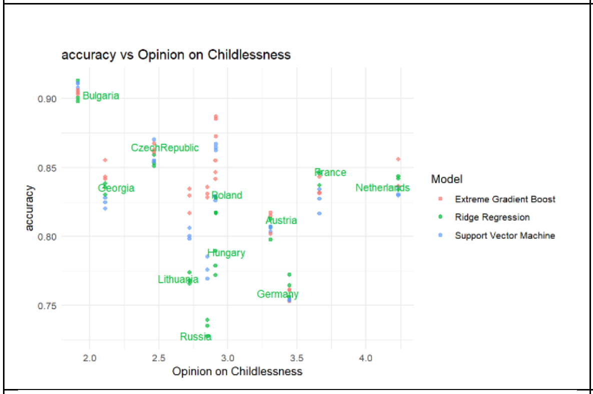
Appendix 1: Sample size per country, variable and missing percentage

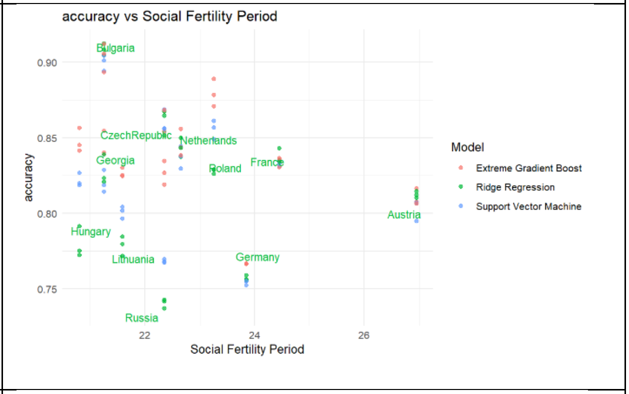
	BULGARIA	RUSSIA	GEORGIA	GERMANY	FRANCE	HUNGARY	NETHERLANDS
	(N=5679)	(N=3752)	(N=4401)	(N=1431)	(N=3183)	(N=4996)	(N=3071)
HIGHEST ACHIEVED LEVEL OF EDUCATION							
MISSING	65 (1.1%)	222 (5.9%)	0 (0%)	94 (6.6%)	0 (0%)	0 (0%)	0 (0%)
CURRENT EMPLOYMENT STATUS							
MISSING	2 (0.0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	11 (0.2%)	2 (0.1%)
CURRENT OCCUPATION							
MISSING	31 (0.5%)	12 (0.3%)	24 (0.5%)	14 (1.0%)	0 (0%)	50 (1.0%)	5 (0.2%)
HAS LONG-STANDING ILLNESS OR CHRONIC CONDITION							
MISSING	6 (0.1%)	0 (0%)	0 (0%)	4 (0.3%)	0 (0%)	8 (0.2%)	0 (0%)
RELIGION							
MISSING	17 (0.3%)	0 (0%)	0 (0%)	2 (0.1%)	115 (3.6%)	88 (1.8%)	271 (8.8%)
CURRENT RELATIONSHIP STATUS							
MISSING	18 (0.3%)	15 (0.4%)	0 (0%)	12 (0.8%)	0 (0%)	1 (0.0%)	0 (0%)
CURRENTLY MARRIED							
MISSING	122 (2.1%)	47 (1.3%)	0 (0%)	17 (1.2%)	0 (0%)	355 (7.1%)	1 (0.0%)
SATISFACTION WITH RELATIONSHIPS WITH PARTNER							
MISSING	143 (2.5%)	153 (4.1%)	0 (0%)	15 (1.0%)	124 (3.9%)	382 (7.6%)	135 (4.4%)
WANT TO HAVE A CHILD IN THE NEXT 3-4 YEARS							
MISSING	659 (11.6%)	797 (21.2%)	682 (15.5%)	400 (28.0%)	530 (16.7%)	382 (7.6%)	652 (21.2%)
AGE OF THE YOUNGEST CHILD (YEARS)							
MISSING	32 (0.6%)	66 (1.8%)	3 (0.1%)	11 (0.8%)	13 (0.4%)	59 (1.2%)	20 (0.7%)

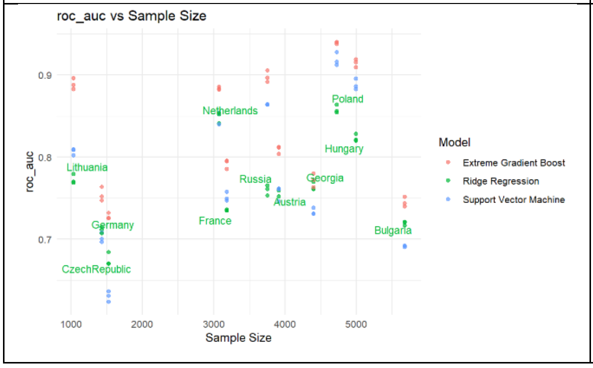
Appendix 2: accuracy and ROC-AUC scatterplots

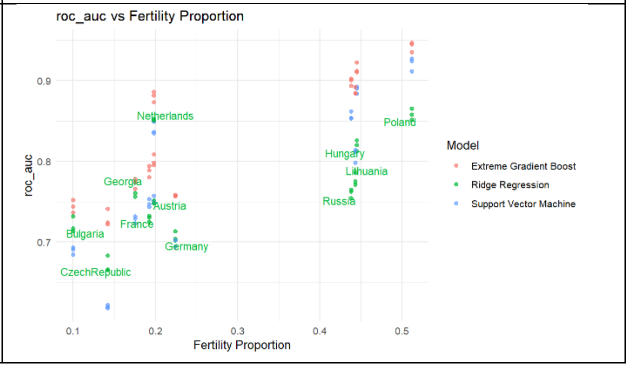

accuracy vs Sample Size

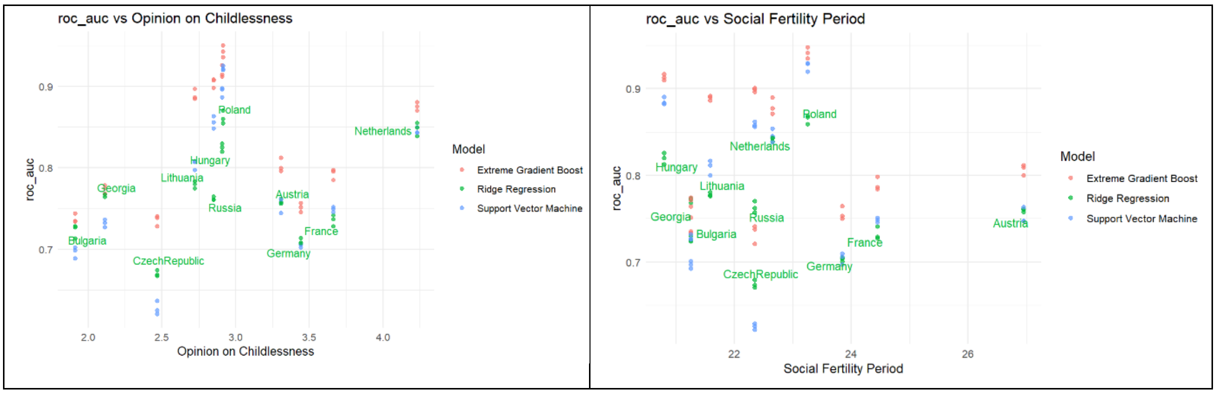

accuracy vs Fertility Proportion


accuracy vs Opinion on Childlessness


accuracy vs Social Fertility Period


roc_auc vs Sample Size


roc_auc vs Fertility Proportion



Appendix 3: linear regressions

```
Call:
lm(formula = .estimate ~ sample_size + fertility_prop + childless_opinion +
    social_fertility_period, data = brier_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.05028 -0.01470  0.00043  0.01233  0.05648

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              9.701e-02  3.465e-02   2.800  0.00621 **
sample_size             -9.559e-06  1.631e-06  -5.862 6.75e-08 ***
fertility_prop           8.116e-02  1.684e-02   4.819 5.53e-06 ***
childless_opinion       -1.811e-03  4.441e-03  -0.408  0.68436
social_fertility_period  2.196e-03  1.648e-03   1.333  0.18578
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02274 on 94 degrees of freedom
Multiple R-squared:  0.3914, Adjusted R-squared:  0.3655
F-statistic: 15.11 on 4 and 94 DF,  p-value: 1.421e-09
```

all:

```
lm(formula = f1 ~ sample_size + fertility_prop + childless_opinion +
    social_fertility_period, data = f1_data)

Residuals:
      Min        1Q    Median        3Q       Max
-0.067384 -0.011104  0.002896  0.013306  0.060049

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              9.867e-01  3.495e-02  28.234  < 2e-16 ***
sample_size              5.844e-06  1.645e-06   3.554 0.000597 ***
fertility_prop          -2.777e-01  1.699e-02 -16.349  < 2e-16 ***
childless_opinion       -9.684e-04  4.479e-03  -0.216 0.829288
```

```
social_fertility_period -2.113e-03  1.662e-03  -1.271 0.206754
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02293 on 94 degrees of freedom
Multiple R-squared:  0.7574, Adjusted R-squared:  0.747
F-statistic: 73.35 on 4 and 94 DF,  p-value: < 2.2e-16
```

Sources

Aarts, A. A., Anderson, J. E., Anderson, C. J., Attridge, P. R., Attwood, A., Axt, J., Babel, M., Bahník, Š., Baranski, E., Barnett-Cowan, M., Bartmess, E., Beer, J., Bell, R., Bentley, H., Beyan, L., Binion, G., Borsboom, D., Bosch, A., Bosco, F. A., . . . Graham, J. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251). https://doi.org/10.1126/science.aac4716

Agresti, A., & Finlay, B. (2013). *Statistical Methods for the Social Sciences* (5th ed.).

Anshul. (2025, April 21). *Support Vector Machine (SVM)*. Analytics Vidhya. https://www.analyticsvidhya.com/blog/2021/10/support-vector-machinessvm-a-complete-guide-for-beginners/

Arpino, B., Moglie, M. L., & Mencarini, L. (2021). What Tears Couples Apart: A machine learning analysis of union dissolution in Germany. *Demography*, *59*(1), 161–186. https://doi.org/10.1215/00703370-9648346

Balbo, N., Billari, F. C., & Mills, M. (2012). Fertility in Advanced Societies: A Review of Research. *European Journal of Population / Revue Européenne De Démographie*, *29*(1), 1–38. https://doi.org/10.1007/s10680-012-9277-y

Beck, K. C., Hellstrand, J., & Myrskylä, M. (2024). *More education and fewer children? the contribution of educational enrollment and attainment to the fertility decline in Norway*. https://doi.org/10.4054/mpidr-wp-2024-009

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425. https://doi.org/10.1037/a0021524

Billari, F. C., Goisis, A., Liefbroer, A. C., Settersten, R. A., Aassve, A., Hagestad, G., & Speder, Z. (2010). Social age deadlines for the childbearing of women and men. *Human Reproduction*, *26*(3), 616–622. https://doi.org/10.1093/humrep/deq360

Binningsley, S., Möllborn, S., & Jalovaarna, M. (2025). Are intensive parenting attitudes a deterrent to childbearing? *Stockholm Research Report in Stockholm*. https://su.figshare.com/articles/preprint/Are_intensive_parenting_attitudes_a_deterrent_to_childbearing_/28395011?file=52301045

Bloom, D. E., Canning, D., Fink, G., & Finlay, J. E. (2009). The cost of low fertility in Europe. *European Journal of Population / Revue Européenne De Démographie*, *26*(2), 141–158. https://doi.org/10.1007/s10680-009-9182-1

Breiman, L. (2001a). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/a:1010933404324

Breiman, L. (2001b). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3). https://doi.org/10.1214/ss/1009213726

Chen, T., & Guestrin, C. (2016). XGBoost. *KDD*, 785–794. https://doi.org/10.1145/2939672.2939785

Çorbacıoğlu, Ş. K., & Aksel, G. (2023). Receiver operating characteristic curve analysis in diagnostic accuracy studies: A guide to interpreting the area under the curve value. *Turkish Journal of Emergency Medicine*, *23*(4), 195–198. https://doi.org/10.4103/tjem.tjem_182_23

Diaz, C. J., & Fiel, J. E. (2016). The Effect(s) of Teen Pregnancy: Reconciling Theory, Methods, and Findings. *Demography*, *53*(1), 85–116. https://doi.org/10.1007/s13524-015-0446-6

Duhigg, C. (2012, February 16). How Companies Learn Your Secrets. *New York Times*. https://www.nytimes.com/2012/02/19/magazine/shopping-habits.html

Eijkemans, M. J., Van Poppel, F., Habbema, D. F., Smith, K. R., Leridon, H., & Velde, E. R. T. (2014). Too old to have children? Lessons from natural fertility populations. *Human Reproduction*, *29*(6), 1304–1312. https://doi.org/10.1093/humrep/deu056

Elleamoh, G. E., & Dake, F. a. A. (2019). “Cementing” marriages through childbearing in subsequent unions: Insights into fertility differentials among first-time married and remarried women in Ghana. *PLoS ONE*, *14*(10), e0222994. https://doi.org/10.1371/journal.pone.0222994

Gadbury, G. L., & Allison, D. B. (2012). Inappropriate Fiddling with Statistical Analyses to Obtain a Desirable P-value: Tests to Detect its Presence in Published Literature. *PLoS ONE*, *7*(10), e46363. https://doi.org/10.1371/journal.pone.0046363

Garip, F. (2020). What failure to predict life outcomes can teach us. *Proceedings of the National Academy of Sciences*, *117*(15), 8234–8235. https://doi.org/10.1073/pnas.2003390117

Goldberg, A. (2011). Mapping Shared Understandings Using Relational Class Analysis: The Case of the Cultural Omnivore Reexamined. *American Journal of Sociology*, *116*(5), 1397–1436. https://doi.org/10.1086/657976

Goldscheider, F., Bernhardt, E., & Lappegård, T. (2015). The Gender Revolution: a framework for understanding changing family and demographic behavior. *Population and Development Review*, *41*(2), 207–239. https://doi.org/10.1111/j.1728-4457.2015.00045.x

Hastie, T., Tibshirani, R., & Friedman, J. (2013). *The elements of statistical learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

He, N. H., & Garcia, E. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/tkde.2008.239

Hearst, M., Dumais, S., Osuna, E., Platt, J., & Scholkopf, B. (1998). Support vector machines. *IEEE Intelligent Systems and Their Applications*, *13*(4), 18–28. https://doi.org/10.1109/5254.708428

Henrich, J., & McElreath, R. (2003). The evolution of cultural evolution. *Evolutionary Anthropology*, *12–123*, 123–135. https://doi.org/10.1002/evan.10110

Jalovaara, M., Neyer, G., Andersson, G., Dahlberg, J., Dommermuth, L., Fallesen, P., & Lappegård, T. (2018). Education, gender, and cohort fertility in the Nordic countries. *European Journal of Population / Revue Européenne De Démographie*, *35*(3), 563–586. https://doi.org/10.1007/s10680-018-9492-2

Janbandhu, M. (2024, November 28). Understanding Support Vector Machines (SVM): A Beginner's Guide with Real-World Examples. *Medium*. https://medium.com/@mohitjanbandhu/understanding-support-vector-machines-svm-a-beginners-guide-with-real-world-examples-82a1c10bc822

Kahneman, D. (2011). *Thinking, fast and slow*. http://ci.nii.ac.jp/ncid/BB2184891X

Kim, C. (2022, July 15). Decision Tree Classifier with Scikit-Learn from Python. *Medium*. https://medium.com/@chyun55555/decision-tree-classifier-with-scikit-learn-from-python-e83f38079fea

Kreyenfeld, M. (2009). Uncertainties in female employment careers and the postponement of parenthood in Germany. *European Sociological Review*, *26*(3), 351–366. https://doi.org/10.1093/esr/jcp026

Kuang, B., Berrington, A., Kulu, H., & Vasireddy, S. (2025). The changing inter-relationship between partnership dynamics and fertility trends in Europe and the United States: a review. *Demographic Research*, *52*, 7. https://doi.org/10.4054/DemRes.2025.52.7

Lazzari, E., & Beaujouan, É. (2025). Self-assessed physical and mental health and fertility expectations of men and women across the life course. *Demography*. https://doi.org/10.1215/00703370-11873109

Lazzari, E., Compans, M., & Beaujouan, E. (2022). Changing childbearing age norms in Europe in times of fertility postponement. *BOOK*. https://doi.org/10.31235/osf.io/xbheq

Leocádio, V. A., Gauthier, A., Mynarska, M., & Costa, R. (2023). The quality of fertility data in the web-based Generations and Gender Survey. *Demographic Research*, *49*, 31–46. https://doi.org/10.4054/demres.2023.49.3

Leocádio, V., Verona, A. P., & Wajnman, S. (2024). Exploring the association between gender equality in the family and fertility intentions: an explanation of the findings in low-fertility countries. *Genus*, *80*(1). https://doi.org/10.1186/s41118-024-00234-z

Liefbroer, A. C., Merz, E., & Testa, M. R. (2014). Fertility-Related Norms across Europe: A multi-level analysis. In *Springer eBooks* (pp. 141–163). https://doi.org/10.1007/978-94-017-9401-5_6

Lundberg, I., Brown-Weinstock, R., Clampet-Lundquist, S., Pachman, S., Nelson, T. J., Yang, V., Edin, K., & Salganik, M. J. (2024). The origins of unpredictability in life outcome prediction tasks. *Proceedings of the National Academy of Sciences*, *121*(24). https://doi.org/10.1073/pnas.2322973121

Merz, E., & Liefbroer, A. C. (2012). The attitude toward voluntary childlessness in Europe: cultural and institutional explanations. *Journal of Marriage and Family*, *74*(3), 587–600. https://doi.org/10.1111/j.1741-3737.2012.00972.x

Mills, M., Rindfuss, R. R., McDonald, P., & Egbert, T. V. (2011). Why do people postpone parenthood? Reasons and social policy incentives. *Carolina Digital Repository (University of North Carolina at Chapel Hill)*. https://doi.org/10.17615/n2cm-f618

Molina, M., & Garip, F. (2019). Machine Learning for Sociology. *Annual Review of Sociology*, *45*(1), 27–45. https://doi.org/10.1146/annurev-soc-073117-041106

Niazkar, M., Menapace, A., Brentan, B., Piraei, R., Jimenez, D., Dhawan, P., & Righetti, M. (2024). Applications of XGBoost in water resources engineering: A systematic literature review (Dec 2018–May 2023). *Environmental Modelling & Software*, *174*, 105971. https://doi.org/10.1016/j.envsoft.2024.105971

Norris, P., & Inglehart, R. (2004). *Sacred and secular: Religion and Politics Worldwide*. Cambridge University Press.

Novella, S. (2012). The power of replication. *Science-Based Medicine*.

Pan, L. (2023). *Comparison of kernel functions and parameter selection of SVM classification algorithms*.

Park, H., & Lee, K. (2022). Using Boosted Machine Learning to Predict Suicidal Ideation by Socioeconomic Status among Adolescents. *Journal of Personalized Medicine*, *12*(9), 1357. https://doi.org/10.3390/jpm12091357

Rahal, C., Verhagen, M., & Kirk, D. (2022). The rise of machine learning in the academic social sciences. *AI & Society*, *39*(2), 799–801. https://doi.org/10.1007/s00146-022-01540-w

Redelmeier, D. A., Bloch, D. A., & Hickam, D. H. (1991). Assessing predictive accuracy: How to compare brier scores. *Journal of Clinical Epidemiology*, *44*(11), 1141–1146. https://doi.org/10.1016/0895-4356(91)90146-z

Salganik, M. J., Lundberg, I., Kindel, A. T., Ahearn, C. E., Al-Ghoneim, K., Almaatouq, A., Altschul, D. M., Brand, J. E., Carnegie, N. B., Compton, R. J., Datta, D., Davidson, T., Filippova, A., Gilroy, C., Goode, B. J., Jahani, E., Kashyap, R., Kirchner, A., McKay, S., . . . Yang, K. (2020). Measuring the predictability of life outcomes with a scientific mass collaboration. *Proceedings of the National Academy of Sciences*, *117*(15), 8398–8403. https://doi.org/10.1073/pnas.1915006117

Saxena, S. (2021, December 15). The Gaussian RBF kernel in non linear SVM - Suvigya Saxena - medium. *Medium*. https://medium.com/@suvigya2001/the-gaussian-rbf-kernel-in-non-linear-svm-2fb1c822aae0

Schoen, R., Astone, N. M., Kim, Y. J., Nathanson, C. A., & Fields, J. M. (1999). Do fertility intentions affect fertility behavior? *Journal of Marriage and Family*, *61*(3), 790. https://doi.org/10.2307/353578

sci-kit learn. (n.d.). *RBF SVM parameters*. Scikit-learn. Retrieved July 27, 2025, from https://scikit-learn.org/stable/auto_examples/svm/plot_rbf_parameters.html

Settersten, R. A., & Hagestad, G. O. (1996). What's the Latest? Cultural Age Deadlines for Family Transitions. *The Gerontologist*, *36*(2), 178–188. https://doi.org/10.1093/geront/36.2.178

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, *25*(3). https://doi.org/10.1214/10-sts330

Sivak, E., Pankowska, P., Mendrik, A., Emery, T., Garcia-Bernardo, J., Höcük, S., Karpinska, K., Maineri, A., Mulder, J., Nissim, M., & Stulp, G. (2024). Combining the strengths of Dutch survey and register data in a data challenge to predict fertility (PreFer). *Journal of Computational Social Science*, *7*(2), 1403–1431. https://doi.org/10.1007/s42001-024-00275-6

Skirbekk, V., Kaufmann, E., & Goujon, A. (2010). Secularism, fundamentalism, or Catholicism? the religious composition of the United States to 2043. *Journal for the Scientific Study of Religion*, *49*(2), 293–310. https://doi.org/10.1111/j.1468-5906.2010.01510.x

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

Spéder, Z., & Bálint, L. (2024). Realization of Short-Term fertility intentions in a comparative perspective: Which Macro-Level conditions matter? *Population Research and Policy Review*, *43*(5). https://doi.org/10.1007/s11113-024-09913-3

Stulp, G., Top, L., Xu, X., & Sivak, E. (2023). A data-driven approach shows that individuals' characteristics are more important than their networks in predicting fertility preferences. *Royal Society Open Science*, *10*(12). https://doi.org/10.1098/rsos.230988

Thorne, M. E. (2005). *Women in society: Achievements, Risks, and Challenges*. Nova Publishers.

Torres, A. F. C., & Akbaritabar, A. (2024). The use of linear models in quantitative research. *Quantitative Science Studies*, *5*(2), 426–446. https://doi.org/10.1162/qss_a_00294

Tragaki, A., & Bagavos, C. (2014). Male fertility in Greece: trends and differentials by education level and employment status. *Demographic Research*, *Vol. 31*, 137–160. http://www.jstor.org/stable/26350060

Van Bavel, J. (2010). Second birth rates across europe: interactions between women's level of education and child care enrolment. *Vienna Yearbook of Population Research*, *8*, 107–138. https://doi.org/10.1553/populationyearbook2010s107

Watts, D. J. (2014). Common Sense and Sociological Explanations. *American Journal of Sociology*, *120*(2), 313–351. https://doi.org/10.1086/678271

Yarkin Yildiz, A., & Kalayci, A. (2024). Gradient boosting decision trees on medical diagnosis over tabular data. *arXiv Preprint arXiv:2410.03705*. https://doi.org/10.48550/arXiv.2410.03705

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in Psychology: Lessons from Machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393

Zheng, A. (2015). *Evaluating machine learning models*.