I Z A Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17046

Machine Learning and Multiple Abortions

Pradeep Kumar
Catia Nicodemo
Sonia Oreffice
Climent Quintana-Domeque

JUNE 2024

I Z A Institute of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 17046

Machine Learning and Multiple Abortions

Pradeep Kumar
University of Exeter

Catia Nicodemo
University of Oxford, University of Verona and IZA

Sonia Oreffice
University of Exeter, HCEO and IZA

Climent Quintana-Domeque
University of Exeter, HCEO and IZA

JUNE 2024

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Machine Learning and Multiple Abortions*

This study employs six Machine Learning methods—Logit, Lasso-Logit, Ridge-Logit, Random Forest, Extreme Gradient Boosting, and an Ensemble—alongside registry data on abortions in Spain from 2011-2019 to predict multiple abortions and assess monetary savings through targeted interventions. We find that Random Forest and an Ensemble method are most effective in the highest risk decile, capturing about 55% of cases, whereas linear models and Extreme Gradient Boosting excel in mid to lower deciles. We also show that targeting the top 20% most at-risk could yield cost savings of 5.44 to 8.2 million EUR, which could be reallocated to prevent unintended pregnancies arising from contraceptive failure, abusive relationships, and sexual assault, among other factors.

JEL Classification: I12, I18, C53, J13, C55

Keywords: Lasso, Logit, multiple abortions, random forest, Ridge, Extreme Gradient Boosting, Ensemble, reproductive healthcare

Corresponding author:
Climent Quintana-Domeque
University of Exeter
Rennes Drive
Exeter, EX4 4PU
United Kingdom

E-mail: C.Quintana-Domeque@exeter.ac.uk

* We thank Giuseppe Cavaliere, Samuel Engle, Aureo de Paula, Dario Sansone and participants at the University of Exeter internal seminar and the Third Catalan Economic Society Conference for comments and suggestions. The code used to generate all the figures and tables in this study will be made available in a public repository upon publication. All errors and omissions are solely our responsibility.

1 Introduction

This study employs a range of statistical and machine learning techniques, including Logit, Lasso-Logit, Ridge-Logit, Random Forest, Extreme Gradient Boosting, and an Ensemble method, to predict multiple induced abortions, or repeat abortions, which occur when a woman has more than one induced abortion during her lifetime. These methods enable us to uncover complex relationships between variables, enhance predictive accuracy, and provide robust insights into the predictability of multiple abortions.

Using administrative data on reported induced abortions in Spain from 2011 to 2019, we address two key questions: First, can we predict multiple abortions? Second, how much monetary savings could be redirected towards enhancing women's reproductive healthcare and support services, such as information dissemination, counselling, and other essential services, to avoid elective abortions that arise from undesirable situations?

The phenomenon of multiple abortions is prevalent not only in Spain but also globally. In Spain, the abortion rate in 2021 was 10.7 per 1,000 women aged 15-44, with 34.65% being multiple abortions (IVE, 2021). This pattern is mirrored, and even stronger, in other countries: in England and Wales, the 2021 abortion rate was 18.6 per 1,000 women, with 43% being multiple abortions (Abortion Statistics, England and Wales, 2021); while in the United States, the rate was 11.6 per 1,000 women, with 42.7% being multiple abortions (Abortion Surveillance, United States, 2021).

Understanding the patterns and circumstances leading to multiple abortions is vital for developing effective reproductive healthcare that integrates contraception and support services, thereby reducing the need for repeat procedures and enhancing women's reproductive autonomy. Our main analysis focuses on induced, elective multiple abortions by women aged 18 or above, excluding multiple abortions induced for medical reasons.

The novelty of our study lies in the innovative application of machine learning to predict multiple abortions among women in Spain, leveraging Bayesian adjustment to derive policy relevant insights from a unique dataset. Specifically, while existing data enables the estimation of the probability of a woman having had a previous abortion given an abortion today, our research seeks to predict the probability of a woman having a future abortion given her history and abortion today. This distinction is critical for understanding the dynamics of abortion recurrence and informing public health interventions.

We use data from 2011-2015 as a training set (453,364 abortions, 35.0% multiple abortions), 2016 and 2017 for hyperparameter tuning (159,775 abortions, 35.7% multiple abortions), and 2018 and 2019 as an out-of-sample test set (168,261 abortions, 34.3% multiple abortions). A decile-based prediction analysis, ranking women by their predicted probability of having a multiple abortion, reveals that the Random Forest and Ensemble methods outperform in the highest risk decile, capturing about 55% of cases, while linear models excel in the mid-range deciles. Notably, the performance of the Extreme Gradient Boosting method is more aligned with linear models than with the Random Forest, particularly in lower-risk deciles.

We also find that implementing targeted policies towards the top 20% most at-risk could yield cost savings ranging from €5.44 to €8.2 million using an Ensemble method. The primary aim of preventing multiple abortions should be to avoid unintended pregnancies arising from undesirable situations, such as contraceptive failure, abusive relationships, or sexual assault, while abstracting from situations dependent on health or medical conditions.

Abortion is a complex and often divisive issue, extensively debated in recent years. Women may experience varied abortion-related care pathways, influenced by factors such as medical needs, contraceptive failure, or experiences of violence like abuse or rape (Coast et al., 2018). Abortion serves as a crucial safeguard for women's health, well-being, and autonomy (Levine and Staiger, 2002), distinctly different from contraceptive methods which aim to prevent pregnancy before it occurs.

Our paper enriches the existing literature on abortion and its accessibility by focusing on multiple abortions. Abortion access profoundly impacts maternal health (Clarke and Mühlrad, 2021) and significantly influences women's education, workforce participation, income, and family dynamics (Bailey and Lindo, 2017; Knowles Myers, 2017; Lindo et al., 2020; Oreffice, 2007; Pop-Eleches, 2010). Furthermore, the rising costs of abortion, particularly in contexts requiring out-of-state travel like in the US, disproportionately affect women of lower socioeconomic status (Lindo et al., 2020). These increasing expenses risk deepening the existing disparities in women's health, well-being, and autonomy.

The application of machine learning to the study of multiple abortions introduces a new dimension to an area that has seen significant growth. This growth is especially notable in identifying risk factors and predicting outcomes related to unintended pregnancies, pregnancy complications, and miscarriages (Liu et al., 2021; Kranker et al., 2020).

While our focus is on predicting induced, elective abortions in adult women who have previously undergone abortions to better understand and potentially prevent pregnancies driven by non-health-related factors, we consider extensions of our analysis to include non-elective abortions and women aged 15-17.

Our findings are intended to inform targeted reproductive healthcare policies and interventions that reduce multiple abortions arising from undesirable situations, such as contraceptive failure, abusive relationships, or sexual assault, while promoting women's health and autonomy. In this regard, the estimated resources saved due to a reduction in multiple abortions could support initiatives such as the development of mobile applications for pregnancy prevention (Mangone et al., 2016; Stifani et al., 2023) aimed at high-risk groups, providing preventive support tailored to their needs.

2 Data

2.1 Available information

Access to the registry data of all induced abortions in Spain was granted by the Spanish Ministry of Health in June 2021.[1] For each year from 2011 to 2019, we have the individual data from the notification questionnaire[2], which contains information on both the pregnant woman and the induced abortion. In the data, it is not possible to identify the affected individual, as the notification form does not include the name, family name, or address. Hence, an institutional review board approval was not required.

In Spain, each voluntary induced abortion must be notified by the responsible physician to the health authority of the region (autonomous community) where the induced abortion took place – the doctor will enter the data online.[3] Information on the pregnant women includes: date of birth, living arrangement (alone, in a couple, etc.), place of residence (province), educational attainment, employment status, number of children alive, number of previous abortions, and type of contraceptive method (if used). Information on induced abortion includes: the source of information on pregnancy termination (public health center, private health center, citizen information helplines, friends or family, media, internet), weeks of gestation at the time of the induced abortion, date of the abortion, reasons

[1]Data can be obtained by writing an application to the Ministerio de Sanidad, Servicios Sociales e Igualdad (Paseo del Prado 18-20, 28014 Madrid, Spain). More information is available on their website.

[2]Last accessed: November 24, 2023.

[3]This online app is available since January 1, 2011.

(requested by the woman, medical reason), publicly vs. privately funded, and type of abortion (surgical, medical).

2.2 Abortions and multiple abortions during 2011-2019

Out of the 910,743 induced abortions recorded during the period 2011-2019, 819,739 are elective (requested by the woman and excluding those performed for medical reasons). Among the women who undergo an abortion in a given year, 63.19% have had no previous abortions, 24.41% have had one previous abortion, 8.03% have had two, 2.62% have had three, and 1.77% have had four or more.

As indicated in Figure 1, we exclude from our analysis: abortions with more than 24 weeks of gestation (n=1,187), abortions from women residing in the provinces of Ceuta and Melilla (n=1,082), those with three previous abortions or more (n=39,883), and non-elective abortions (n=87,191). Hence, the final sample comprises 781,400 observations.

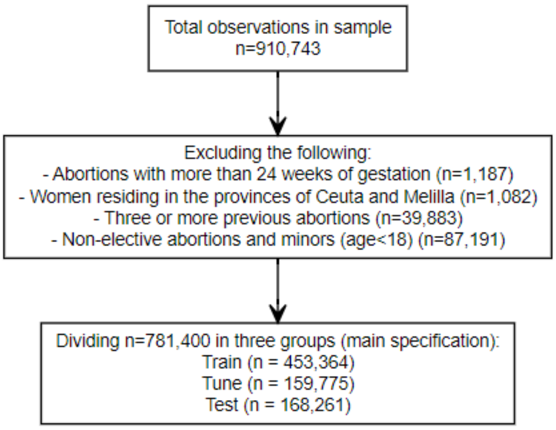

Figure 1: Flow chart on sample selection

Our outcome variable is coded as 1 if the unit of observation represents a multiple abortion (i.e., the woman had a previous abortion), and 0 if it is the first abortion for the woman in a given year. We focus on predicting the likelihood of a woman having a subsequent abortion, rather than the total number of abortions she will have.

2.3 Splitting the Data: Training, Tuning, and Testing

One of the primary objectives of the machine learning approach is to mitigate the overfitting problem, which occurs when a model with excellent in-sample performance results in poor predictive power for out-of-sample data (Mullainathan and Spiess, 2017). As indicated in

Figure 1, we divide our final sample into three distinct sets: training, tuning, and testing. The training set consists of 453,364 observations from the years 2011-2015 and is utilized to estimate the relevant model. The validation set, encompassing 159,775 observations from 2016 and 2017, is used to tune the model's hyper-parameters, such as the penalization term in Lasso-Logit. The third set, comprising 168,261 observations from the years 2018 and 2019, serves as out-of-sample test data.

While it is common for practitioners to divide their data into two sets (training and testing), using the training set for both model estimation and hyper-parameter tuning through cross-validation is typically reserved for instances where the dataset is not sufficiently large. This is not the case here. Furthermore, using only two sets can be problematic, as tuning hyper-parameters on the same data used for training the model can lead to biased results (Russell and Norvig, 2015; Ng, 2016; Sansone, 2019).

2.4 Predictors and descriptive statistics

The predictors consist of the variables described and summarized in Table 1, as well as the woman's place of residence among 50 Spanish provinces. We believe that splitting the data based on time periods is a policy-relevant approach, as policymakers rely on past data to make predictions for the future. The spatial dimension is accounted for by including the Spanish province of residence indicator(s).

Table 1 shows that the conditional prevalence of multiple abortions across predictors is similar over time, that is, across the training, tuning and testing datasets. The mean (or fraction) of multiple abortions within each age category shows no significant variation across the three data groups. The Table reveals some interesting patterns among women undergoing an abortion in a given year during the period 2011-2019: for instance, women aged 25-29 and 30-34 exhibit a higher likelihood of multiple abortions compared to other age groups. Similar trends are observed for other predictors. Women with primary education, living as a couple, having dependent children, using hormonal contraceptives, receiving information from private health centres, and foreign-born women are more likely to experience multiple abortions than their counterparts in other categories.

		Training (2011-15)		Tuning (2016-17)		Testing (2018-19)	
		Mean	Freq.	Mean	Freq.	Mean	Freq.
Age							
	15–19	0.15	53,180	0.15	16,960	0.14	17,838
	20–24	0.32	100,450	0.31	35,035	0.30	37,722
	25–29	0.40	102,205	0.40	35,114	0.38	37,522
	30–34	0.42	95,974	0.43	32,725	0.41	33,882
	35–39	0.39	70,704	0.40	27,337	0.40	27,905
	40–44	0.37	26,652	0.37	10,997	0.38	11,788
	Missing	0.20	4,199	0.20	1,607	0.22	1,604
Education							
	Primary	0.41	98,218	0.42	28,360	0.39	27,369
	Secondary	0.36	288,247	0.37	106,096	0.36	111,566
	University	0.24	58,387	0.24	23,065	0.23	27,197
	Missing	0.34	8,512	0.34	2,254	0.33	2,129
Employment status							
	Unemployed	0.33	218,035	0.34	67,236	0.33	65,002
	Employed	0.37	230,229	0.37	90,432	0.36	100,294
	Missing	0.36	5,100	0.37	2,107	0.34	2,965
Living arrangement status							
	Alone	0.36	105,634	0.36	39,747	0.35	38,535
	As a couple	0.40	214,441	0.40	73,942	0.39	77,370
	With family/parents	0.26	111,826	0.27	38,236	0.27	43,620
	With others	0.33	12,293	0.32	4,241	0.27	5,664
	Missing	0.32	9,170	0.35	3,609	0.35	3,072
Dependent children							
	No	0.25	161,747	0.26	57,908	0.24	70,898
	Yes	0.44	188,143	0.45	65,472	0.44	76,434
	Missing	0.35	103,474	0.35	36,395	0.34	20,929
Type of contraceptive							
	Natural	0.39	18,085	0.40	5,452	0.36	6,330
	Barrier	0.29	120,476	0.31	40,536	0.30	45,296
	Mechanical	0.35	4,071	0.38	1,412	0.36	2,061
	Hormonal	0.41	63,572	0.43	22,334	0.42	26,727
	Irreversible	0.36	4,223	0.35	594	0.32	637
	No contraceptive	0.35	96,337	0.35	35,011	0.34	29,091
	Missing	0.37	146,600	0.36	54,436	0.35	58,119
Source of information							
	Public health centre	0.34	286,593	0.36	104,026	0.35	114,876
	Private health centre	0.44	48,001	0.44	15,526	0.41	15,224
	Over the phone	0.38	2,653	0.41	907	0.38	1,411
	Family or friends	0.36	44,035	0.38	12,232	0.37	11,602
	Media	0.40	2,815	0.35	375	0.33	283
	Internet	0.26	42,414	0.27	20,402	0.26	20,904
	Missing	0.42	26,853	0.35	6,307	0.33	3,961
Foreign born							
	No	0.30	281,086	0.32	102,593	0.31	105,541
	Yes	0.43	171,335	0.43	56,584	0.40	61,734
	Missing	0.27	943	0.27	598	0.25	986
Publicly funded							
	No	0.34	142,898	0.32	41,119	0.30	38,845
	Yes	0.36	310,466	0.37	118,656	0.36	129,416
	Total	0.35	453,364	0.36	159,775	0.34	168,261

Table 1: Summary Statistics. The table reports the mean of the multiple abortion indicator (=1 if multiple abortion; =0 if first abortion) and the count of total abortions in each category for the three data sets.

3 Methods

3.1 Predictive methods

Our prediction problem is a binary classification task aimed at obtaining robust predictions for out-of-sample data. To achieve this, we employ six predictive methods based on five models:

- **Logit**, **Lasso-Logit** (Least Absolute Shrinkage and Selection Operator, Tibshirani (1996)), and **Ridge-Logit**, which produce a linear decision boundary in the predictor space, classifying each observation into one of two categories (1 or 0). These models may not be optimal if the true decision boundary is non-linear. Among them, Lasso-Logit is notable for its ability to select the most crucial variables while excluding irrelevant ones. Additional details are provided in the online appendix, section A1.1.

- **Random Forest** (Breiman, 2001) and **Extreme Gradient Boosting** (Chen and Guestrin, 2016), which are non-parametric methods that do not assume any functional form. These decision tree-based methods excel at capturing non-linearities and interactions among predictors (Berk et al., 2008; Hastie et al., 2009), resulting in a non-linear decision boundary. Random Forest, in particular, is constructed by averaging a specified number of classification trees, each grown from a bootstrapped sample and a different random sub-sample of inputs. Each tree predicts whether observations are multiple abortions or not, with classification based on a majority vote across the trees. The main distinction between Random Forest and Boosting is that while trees in Random Forest are grown in parallel from bootstrapped samples, trees in Boosting are grown sequentially, starting with a simple decision tree and gradually learning the prediction function from the data. Additional details are provided in the online appendix, sections A1.2 and A1.3.

- **Ensemble Method** (Dietterich, 2000; Athey and Imbens, 2019), which combines predictions from the aforementioned five methods. We explore two approaches for combining predicted probabilities: a simple average and a logistic regression. Additional details are provided in the online appendix, section A1.4.

3.2 Bayes Correction

Among women who have an abortion in a given year, our data allow us to identify the probability of having had a previous abortion. However, we are interested in predicting the probability of a woman having an abortion today, given that she had a previous abortion. To infer the latter from the former, we perform a Bayesian adjustment. Formally, we write

$$P(A_{i,t} = 1 \mid A_{i,t-s} = 1, X_{i,t}) = P(A_{i,t-s} = 1 \mid A_{i,t} = 1, X_{i,t}) \frac{P(A_{i,t} = 1 \mid X_{i,t})}{P(A_{i,t-s} = 1 \mid X_{i,t})}$$

where $A_{i,t} = 1$ indicates the woman i had an abortion in year t, and $A_{i,t-s} = 1$ indicates she had an abortion in year $t - s$, with $X_{i,t}$ representing a vector of predictors.[4] While the conditional probability $P(A_{i,t-s} = 1 \mid A_{i,t} = 1, X_{i,t})$ can be directly estimated with our data and has been the object of previous research (see Nicodemo et al. (2022)), we require assumptions to estimate $P(A_{i,t-s} = 1 \mid X_{i,t})$ and $P(A_{i,t} = 1 \mid X_{i,t})$ in order to recover the conditional probability of interest $P(A_{i,t} = 1 \mid A_{i,t-s} = 1, X_{i,t})$.

The challenge with our dataset is its limitation to only include cases of women who underwent an abortion within the year, omitting those with zero abortions at time t. Consequently, our dataset at time t is left-truncated at zero, lacking both the dependent variable and the predictors for women without abortions. Figure 2 illustrates that the proportion of abortion counts is consistent across the years. During 2011-2019, 63.95-66.25% of observations are first abortions, 25.31-27.06% are second abortions, and the remaining 8.24-9.19% are third abortions. The observed persistence in our data suggests that the stationarity condition is a reasonable assumption, which is a requirement for being able to consistently estimate $P(A_{i,t-s} = 1 \mid X_{i,t})$ and $P(A_{i,t} = 1 \mid X_{i,t})$ through a zero-truncated Poisson regression.

Given that zero-truncated Poisson models are not recommended for small samples, we restrict our analysis to women with a maximum of 3 induced abortions, representing more than 95% of the total number of reported abortions during the period 2011-2019. While our dataset may include multiple entries for the same woman, the high proportion of first-time abortion cases (63.95-66.25%) implies limited duplication. Although this duplication could potentially introduce some bias in coefficient estimation, its impact is likely to be minimal. This is because our analysis primarily focuses on the predicted probabilities, which are less

[4] Age, education, employment status, living arrangement status, dependent children, type of contraceptive, source of information, foreign born, publicly funded, and province of residence.

sensitive to such biases than the coefficients.

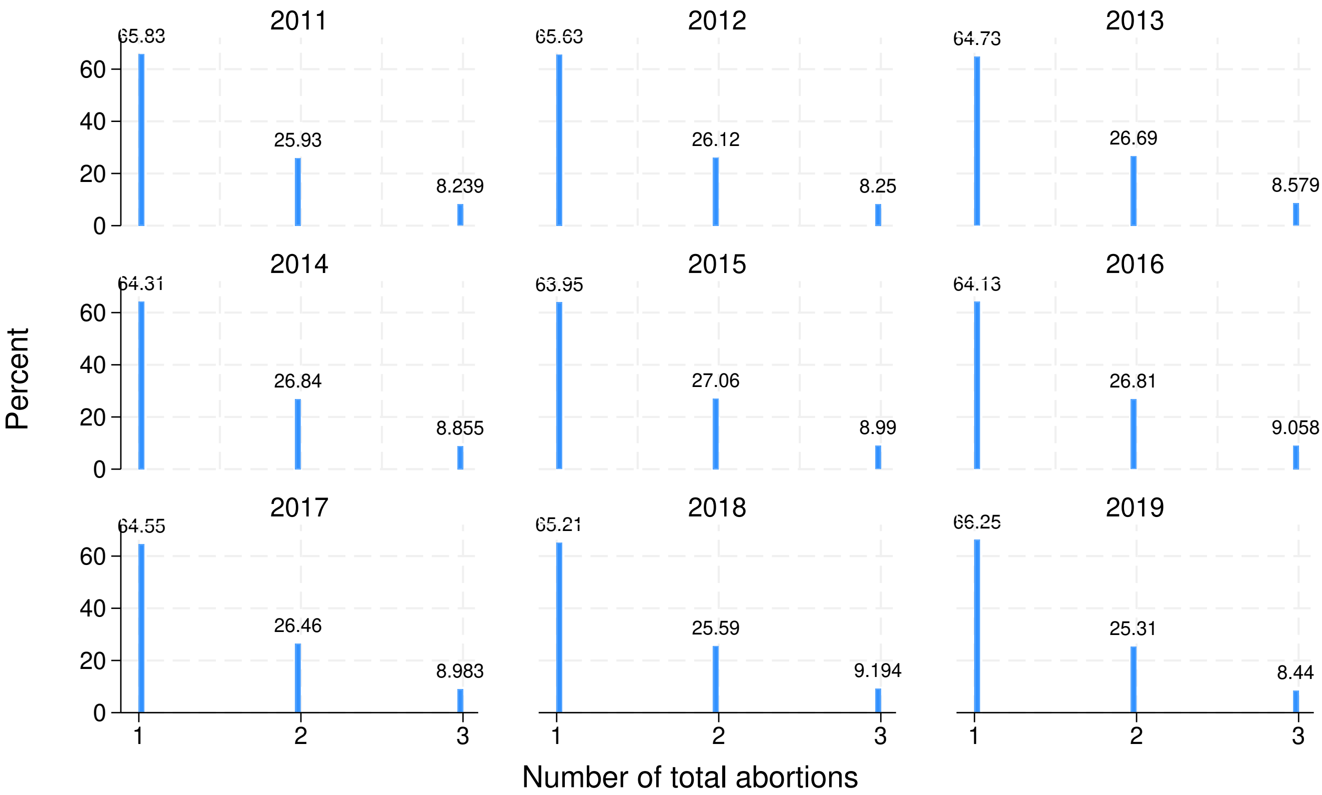

Figure 2: Percentage of first-time (1) and multiple abortions (2 and 3) by year.

4 Predicting multiple abortions

We implement the prediction of multiple abortions using `RStudio` version 2023.09.1 and `R` version 4.3.2 for Windows.[5] Implementation details are reported in the online appendix A1. All methods employ nine categorical variables listed in Table 1 and dummy variables for provinces. Given the high number of missing values in two of our variables, namely type of contraceptive and dependent children, which have 33.17% and 20.58% missing observations respectively,[6] we are treating the missing values in each variable as a distinct category for implementation purposes.

The predicted probability of each method is weighted by the Bayes adjustment. We use a zero-truncated Poisson regression model to calculate the components of the Bayes adjustment, $P(A_{i,t} = 1|X_{i,t})$ and $P(A_{i,t-s} = 1|X_{i,t})$. Details are provided in the online

[5]Last accessed: 6 December 2023. RStudio and R.

[6]The remaining variables each have less than 5% missing observations. See Table A2 in the online appendix A2.

appendix A1.5.

In evaluating the performance of classification models, scholars frequently employ a range of metrics, such as Area Under the Curve (Area Under the ROC (Receiver Operating Characteristic) Curve) and sensitivity (proportion of multiple abortions that are correctly classified) or accuracy (proportion of total abortions that are correctly classified) at different thresholds. However, these measures may not always be conducive to the design or execution of an intervention. For instance, if policymakers have resources to enact an intervention for 10,000 women, equivalent to approximately 10% of total annual elective abortions, it is crucial to utilise our model to identify the 10% of women who are most likely to undergo multiple abortions. To achieve this, we conduct a decile analysis on our out-of-sample data, which involves segregating the data into ten distinct bins and ordering them according to predicted probability. Subsequently, we ascertain the proportion of multiple abortions within each bin across the various machine learning models.

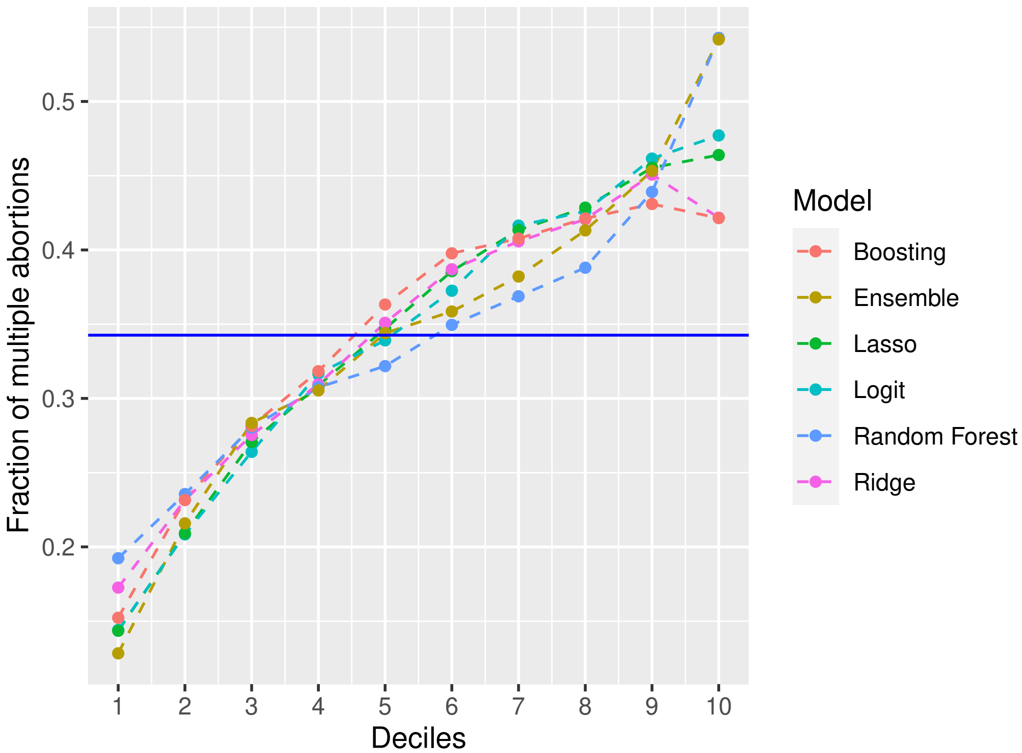

Figure 3: Decile analysis on out-of-sample data. The data has 168,261 observations with 34% multiple abortions. The horizontal blue line denotes the prediction based on the unconditional mean (0.34).

Figure 3 displays the proportion of multiple abortions identified by each machine learning method across distinct deciles, ranking women by their predicted probability of having a multiple abortion. Out-of-sample data for 2018 and 2019 show that 34% of the cases are multiple abortions, serving as a benchmark for evaluating the predictive methods. In the highest decile, the Random Forest and Ensemble methods capture approximately 55% of multiple abortions, whereas the linear decision boundary models (Logit, Lasso-Logit, and Ridge-Logit) predict around 45%. Performance across all models is similar in the second highest risk decile, clustered around 45%. In the subsequent five deciles, linear models predict more effectively, while Random Forest excels in the two least risky deciles. Above the unconditioned mean (solid horizontal line), an Ensemble model acts as an effective compromise between Random Forest and the linear models. Notably, the performance of the Extreme Gradient Boosting method aligns more closely with the linear models than with the Random Forest method.[7]

The ROC analysis in Figure A5 in the online appendix A3 reveals overall performance among the six classification methods. Interestingly, an Ensemble method appears to perform at least as well as, if not better than, the rest from an overall (cumulative) standpoint. The Random Forest model shows varied effectiveness at different cut-offs (better performance at higher cut-offs), as indicated by its Area Under the Curve confidence interval and alignment with previous findings. As seen in the deciles plot, an Ensemble method's performance does not decline as much as the Random Forest's in the middle deciles; it is able to perform better cumulatively than all the other methods.

5 Robustness checks

In the online appendix A4, we conduct several robustness checks. First, for completeness, we present the analysis after the inclusion of non-elective induced abortions. This analysis shows that the Random Forest and Ensemble models are most effective in extreme risk deciles, with results similar to the initial analysis (see Figure A6).

Second, it is crucial to acknowledge that the performance of these machine learning methods is influenced by the quality of our data. In our main analysis, we address the missing data by creating an additional category for each variable. This approach was chosen over deleting all observations with missing data, as the latter would result in the loss of over half of our observations. After excluding missing data, method performance

[7]The Random Forest model was configured with two randomly selected predictors at each split, with error rates calculated on the tuning sample. See Table A1.

varied across deciles (see Figure A7); the removal of two variables with the most missing observations ('type of contraceptive' and 'dependent children') resulted in a performance decline (see Figure A8).

Finally, in the last robustness check, we show that Random Forest models with and without dummy variables yielded remarkably similar results (see Figure A9).

6 Potential Monetary Savings Through Targeted Interventions

How much money could be saved by preventing multiple abortions, so that these savings could be reallocated to avoid unintended pregnancies arising from situations such as contraceptive failure, abusive relationships, and sexual assault? A first-order approximation to provide an answer to this question needs two pieces of information: an estimated price of an abortion and an estimated number of reduced multiple abortions.

The cost of an abortion in Spain ranges from €325 to €490.[8] Our test data from 2018 to 2019 include 57,643 cases of multiple abortions. Using these figures, we can calculate the minimum and maximum monetary costs associated with multiple abortions as predicted by different ML methods. The total cost of multiple abortions in 2018-2019 is estimated to be between €18,733,975 and €28,245,070. Dividing this number by two gives an estimate of the annual costs.

Table 2 presents the potential monetary savings achievable through various intervention strategies. When targeting the top two deciles, an Ensemble method is the most effective, predicting cost savings between €5.44 million and €8.2 million. Expanding the intervention to include more deciles could increase these savings, but it would also lead to more false positives in our intervention sample.

Note that in our sample of 781,400 abortions, 71.48% are publicly funded. Within the subset of 273,484 multiple abortions, 73.35% are publicly funded. In the specific case of our test sample, which includes 57,643 multiple abortions, 79.78% are publicly funded. To focus solely on public costs, one can approximate the figure by multiplying the estimates in Table 2 by a factor of 0.8.

While the optimal way to design interventions would be to target women in decreasing order of their predicted risk scores, based on our machine learning methods, the execution of such interventions can be constrained by political and implementation frictions. In such

[8] https://www.clinicasabortos.com/precio-del-aborto (Last accessed: November 24, 2023).

	Number of top risky deciles				
	Top 2	Top 4	Top 6	Top 8	All
Logit	[5.13,7.74]	[9.74,14.68]	[13.63,20.55]	[16.81,25.34]	[18.73,28.25]
Lasso-Logit	[5.03,7.58]	[9.63,14.52]	[13.64,20.56]	[16.81,25.34]	[18.73,28.25]
Ridge-Logit	[4.77,7.19]	[9.29,14.01]	[13.33,20.09]	[16.52,24.91]	[18.73,28.25]
Random Forest	[5.37,8.1]	[9.51,14.34]	[13.18,19.87]	[16.39,24.72]	[18.73,28.25]
Boosting	[4.66,7.03]	[9.19,13.86]	[13.35,20.14]	[16.63,25.08]	[18.73,28.25]
Ensemble	[5.44,8.2]	[9.79,14.76]	[13.63,20.55]	[16.85,25.41]	[18.73,28.25]

Table 2: How much money can be saved by predicting multiple abortions per year (in million €)? The cost savings are based on our out-of-sample data predicted number of abortions (years 2018 and 2019) and the price of an abortion (minimum: €325, maximum: €490).

scenarios, it may be useful to know the top predictors of the risk scoring methods to design simpler interventions, instead of targeting women in a non-linear, yet optimal, manner.

In the online appendix A5, we discuss the top predictor variables using LASSO-Logit and Random Forest models. Both methods identify four common predictors: a woman's age, province of residence, source of information about pregnancy termination, and education level. These factors not only highlight the unequal distribution of multiple abortions among the population of women but also suggest that constrained, tailored interventions to women with these characteristics can hopefully provide support to reduce the risk of pregnancies arising from undesirable situations such as contraceptive failure, abusive relationships, sexual assault, and other factors, and effectively reduce multiple abortions.

7 Conclusion

Our study utilises Machine Learning methods combined with registry data on abortions in Spain from the Ministry of Health, Social Services, and Equality for the period 2011-2019. Our objectives are to predict multiple abortions using various models (Logit, Lasso-Logit, Ridge-Logit, Random Forest, Extreme Gradient Boosting, and Ensemble), identify key predictors, and estimate potential monetary savings through targeted interventions.

We find that the Random Forest and Ensemble models capture 55% of multiple abortions in the top decile, while the remaining models predict around 45%. Although all models perform similarly in the second decile, linear models outperform Random Forest in the middle six deciles, with Random Forest excelling in the two lowest-risk deciles. an Ensemble model, acting as a compromise between linear methods and the Random Forest, suggests potential monetary savings of between €5.44 to €8.2 million. The Logit model projects savings of €5.13 to €7.74 million, targeting the top 20% most at-risk women.

We hope that the findings of this study can be utilised by national and regional health authorities, enabling policymakers to use preferred algorithms for screening and effectively supporting women at risk of multiple abortions due to unintended pregnancies. This strategy echoes recent developments and suggestions in the field of mobile phone apps and public health information (Mangone et al., 2016; Stifani et al., 2023).

Online Appendix

A1 Implementation details

We use six methods to predict multiple abortions: Logit, Lasso-Logit, Ridge-Logit, Random Forest, Extreme Gradient Boosting, and an Ensemble method. There was no tuning involved in the Logit method, and since it is a widely used method, we do not discuss it further here.

A1.1 Lasso-Logit and Ridge-Logit

While the model parameters, denoted as β, in Logit regression are estimated by minimising the negative log-likelihood function (L), the Lasso-Logit and Ridge-Logit regression work by adding a penalty term to the negative log-likelihood function. The expressions to be minimised in the two cases are:

$$L + \lambda \sum |\beta| : \text{ for Lasso-Logit Regression}$$
$$L + \lambda \sum \beta^2 : \text{ for Ridge-Logit Regression}$$

where λ is a non-negative parameter selected by either minimising the cross-validation error or out-of-sample error on some validation set. We create a grid of 100 equally spaced numbers between 10^{-5} (0.00001) and $10^{0.5}$ (3.16) to choose an optimal λ. For each λ in the grid, model coefficients are estimated for both the Lasso-Logit and Ridge-Logit regression using the training sample (2011-15). Figures A1 and A2 show how the value of model coefficients change with the value of the tuning parameter λ. All predictor coefficients tend towards zero as we move to higher values of λ (right-side of the figure), whereas for the smaller values of λ we move closer to the Logit regression coefficients. For $\lambda = 0$, both Lasso-Logit and Ridge-Logit models are exactly the same as the Logit regression model.

Subsequently, we calculate the misclassification error rate for the 100 models on the tuning sample (2016 and 2017). The λ corresponding to the minimum error rate in this tuning sample is chosen as the optimal tuning parameter. The optimal λ for the Lasso-Logit model is 0.00021 and for the Ridge-Logit model is 0.00316. They are denoted by vertical dashed lines in Figures A1 and A2. As the optimal λ is closer to zero for both models, it is not surprising that the performance of both models is similar to the Logit model.

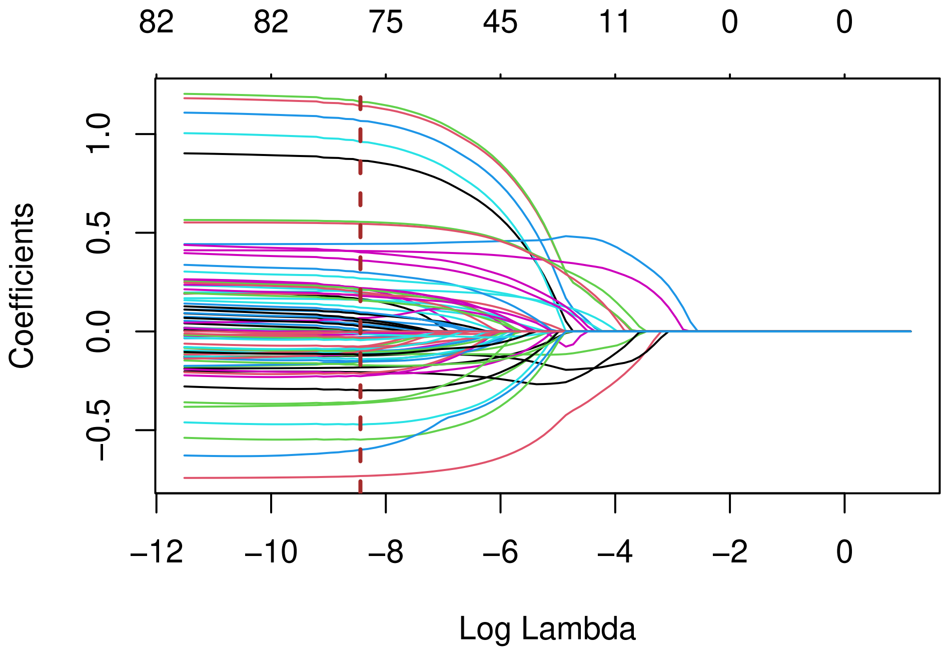

Figure A1: Tuning parameter for the Lasso-Logit model. Log(λ) vs. model coefficients. The dashed vertical line shows the optimal value of the tuning parameter. The top line shows number of non-zero coefficients.

Finally, we make predictions for the out-of-sample test data (2018 and 2019) using the model with the optimal λ and report the findings in Figure 3.

A1.2 Random Forest

Random Forests are built by combining predictions from several decision trees constructed using bootstrapped training samples. Each time a split is considered in a decision tree, a random sample of m predictors is considered from the full set of p predictors. The split is allowed to use one of the m chosen predictors. The Random Forest method tends to perform better than individual trees by generating multiple de-correlated trees (through bootstrap and random sampling), which reduces variance—one of the components of the test error rate. As the Random Forest method is built by combining decision trees, there are at least two benefits to it. First, a decision tree will incorporate interactions between predictors, which are driven by the data and not chosen by the practitioner. It is common in social sciences to include interaction terms to account for synergies between two variables, but the choice of variables is often based on the practitioner's intuition, which can sometimes be ad-hoc. Second, it can be argued that humans are more likely to

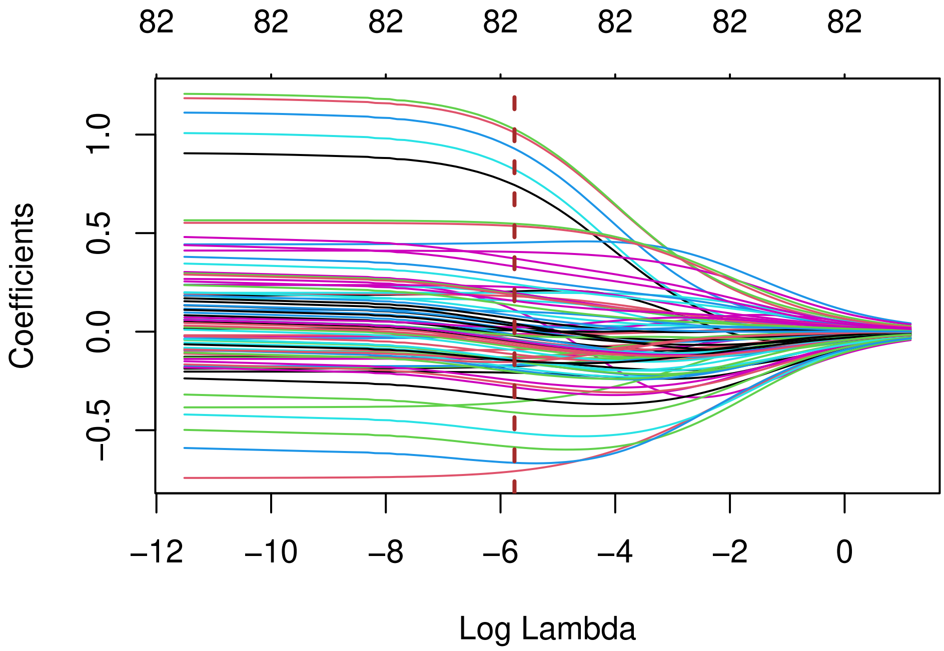

Figure A2: Tuning parameter for the Ridge-Logit model. Log(λ) vs. model coefficients. The dashed vertical line shows the optimal value of the tuning parameter. The top line shows number of non-zero coefficients.

make decisions by following certain rules (decision trees) rather than assigning weights to each possible predictor (regression methods). It is not surprising that Random Forest (and other tree-based methods such as Boosting) performs well in social science settings.

Our data has $p = 10$ total predictor variables (all categorical) as reported in Table 1. We use the `R` command *randomForest*, which recognises categorical (factor) variables and constructs trees accordingly by dividing categories at each split. We use the training sample to build the Random Forest model considering $m = 2, 3, 5, 7, 10$ predictors at each split of the decision tree. As increasing the number of trees does not lead to overfitting, it is not considered a tuning parameter and is often set to a high enough number (Probst and Boulesteix (2017) and Hastie et al. (2009)). The number of trees we use in our analysis is 500. The optimal number of predictors at each split, $m = 2$, is found by calculating the error rate on the tuning sample as reported in Table A1. Finally, we use the Random Forest model with two randomly selected predictors at each split to calculate the out-of-sample error rate.

m	Tuning error rate
2	0.3328
3	0.3347
5	0.3444
7	0.3489
10	0.3504

Table A1: Random Forest: Error rate on the tuning data

A1.3 Extreme Gradient Boosting

Extreme Gradient Boosting is an efficient implementation of the gradient boosting method proposed by Chen and Guestrin (2016). We use the *xgboost* package in R for our implementation. Traditionally, there are three tuning parameters in a boosting method: learning rate δ, number of iterations, and depth of each tree. We set the number of iterations to 100. As the boosting tree is grown sequentially, it is not necessary to tune with respect to both the number of iterations and the learning rate. We tune the boosting model using the learning rate. We create a grid of 50 points between 0.1 and 0.7 and train boosting models for each possible δ. We choose the depth of the tree to be 6. We found no significant effect on the predictive performance by changing the depth of the tree. Figure A3 shows the tuning error rate for different learning rates. The 'U' shape of the error rate curve clearly illustrates a trade-off between bias and variance as we increase the flexibility of our method. Subsequently, we choose our model ($\delta = 0.43$) whose error rate is minimum on the tuning dataset. The chosen model is used to predict multiple abortions on the test data.

We also explored whether regularization of the boosting tree could improve the performance (using the parameter γ in the *xgboost* function), however, we found no significant difference in our results.

A1.4 Ensemble

An Ensemble method (Dietterich, 2000; Athey and Imbens, 2019) is derived by merging predictions from various models. We explore two different approaches. First, a simple average of the predicted probabilities of the five models: Logit, Lasso-Logit, Ridge-Logit, Random Forest, and Extreme Gradient Boosting. The outcome of this approach has been discussed in the main section of the paper. Second, we use logistic regression to combine the five methods. We employ a logistic regression on the tuning data with the binary multiple abortions indicator as the dependent variable and the predicted probabilities from

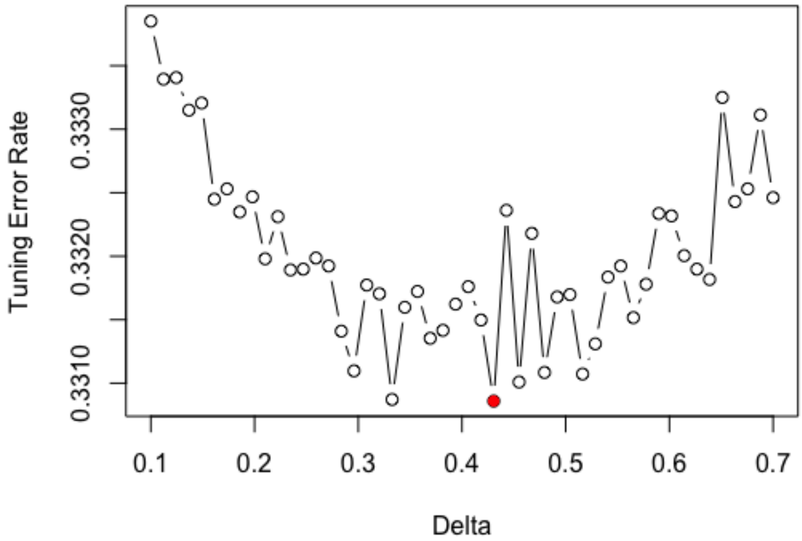

Figure A3: Tuning for the Extreme Gradient Boosting model: Learning Rate δ vs. Tuning Error Rate. The red dot denotes the optimal model with $\delta = 0.43$.

the five methods as the predictor variables. Next, using the estimated coefficients of this logistic regression, we predict probabilities on the test data. Figure A4 shows the deciles plot with the Logistic Ensemble model. Comparing with the simple average Ensemble in Figure 3, we can clearly see that the Logistic Ensemble model places a higher weight on the linear methods, hence the performance is not as good as the simple average in the most risky deciles.

A1.5 Zero-truncated Poisson Regression

In our baseline specification, we analyse data on women having their first, second, or third abortion. We estimate the Zero-truncated Poisson regression on the training and tuning data (2011-17) and use it to predict the probability of a woman having her first, second, or third abortion on the test data (2018-19). We employ the same 10 categorical predictors for this regression as used in the machine learning methods.

Subsequently, $P(A_{i,t-s} = 1|X_{i,t})$ is calculated for the test data by adding the probabilities of a woman having her second or third abortion in year t. Similarly, $P(A_{i,t} = 1|X_{i,t})$ is calculated for the test data by summing the probabilities of a woman having her first, second, or third abortion in year t. These two probabilities comprise the Bayes adjustment,

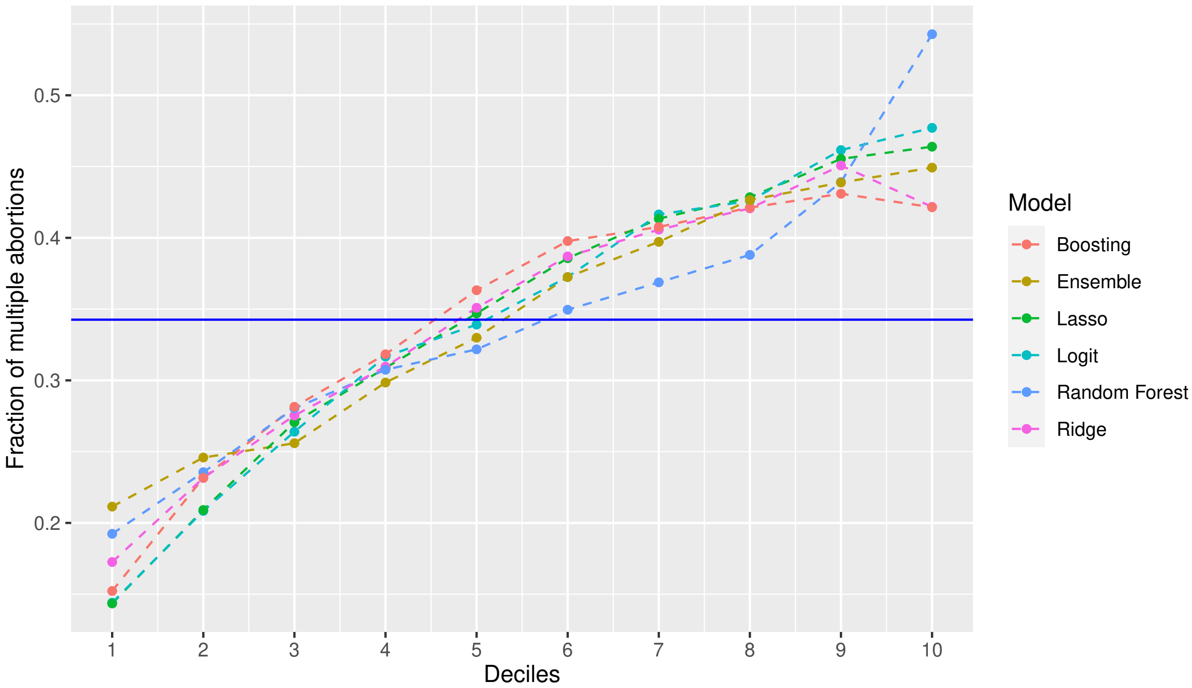

Figure A4: Deciles plot comparing Logit Ensemble with other methods.

which is then applied to the predicted probabilities of all the machine learning algorithms.

A2 Missing observations by variable

Variable	% Missing observations
Type of contraceptive	33.17%
Dependent children	20.58%
Source of information	4.75%
Living status	2.03%
Education	1.65%
Employed	1.30%
Age	0.95%
Foreign born	0.32%
Province	0%
Publicly funded	0%

Table A2: Percentage of Missing Observations

A3 ROC curve

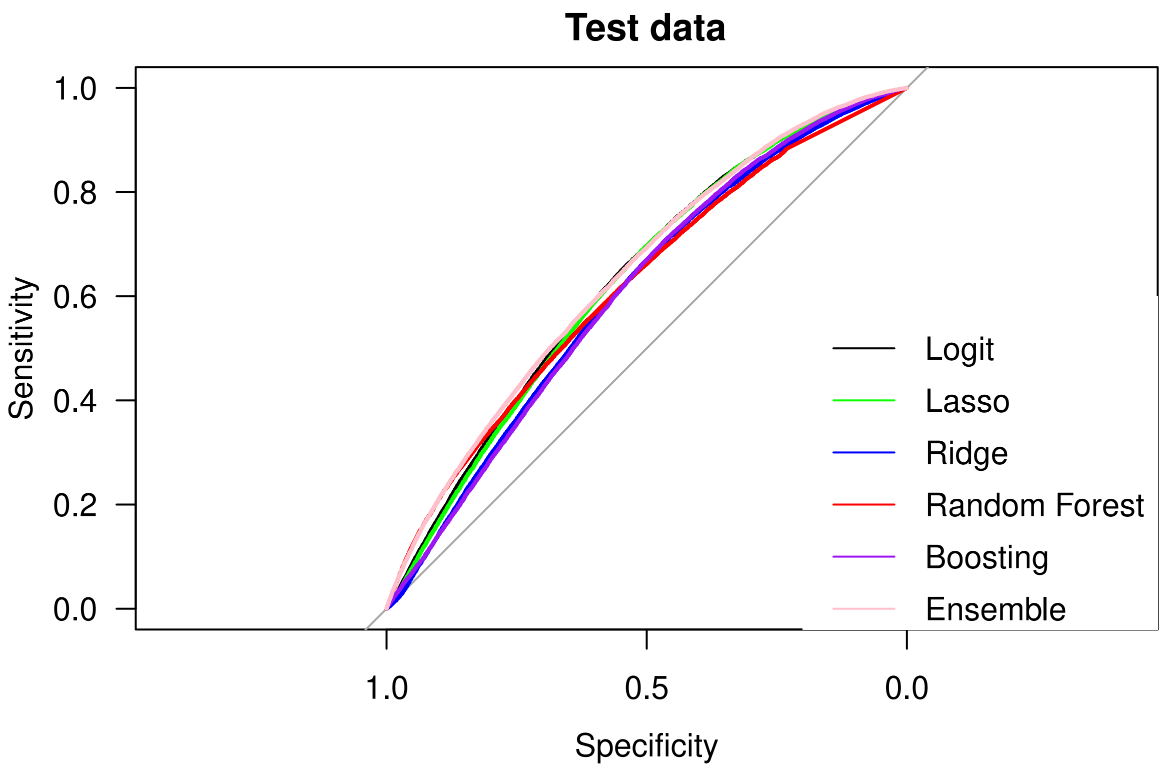

Figure A5: ROC curves for the four methods on the test data.

A popular way to compare various classification methods is through the ROC graph, as shown in Figure A5. Practically, the graphs for all six methods look very similar. The area under the ROC curve (AUC) indicates the average performance of a classifier across all cut-offs. The 95% confidence interval of the AUC measure for all six methods lies between 0.61 and 0.64. However, upon closer examination, we do see that an Ensemble method is at least weakly preferred across all cut-offs.

A4 Robustness Checks

A4.1 Including Non-elective Abortions

The primary analysis presented in the preceding sections deliberately excluded non-elective induced abortions, as we posited that predicting multiple abortions, which were deemed medically necessary, may not hold significant policy relevance. In this section, we extend our analysis to include the non-elective abortions, accounting for 87,191 cases from 2011 to 2019. The predictive performance for all six employed methods is exhibited in Figure A6. Notably, the results parallel our initial analysis in a qualitative sense. That is to say, in the highest and lowest risk deciles, the Random Forest and Ensemble methods outperform Logit, Lasso-Logit, and Ridge-Logit, whereas in the intermediate deciles, the models featuring a linear decision boundary yield superior results.

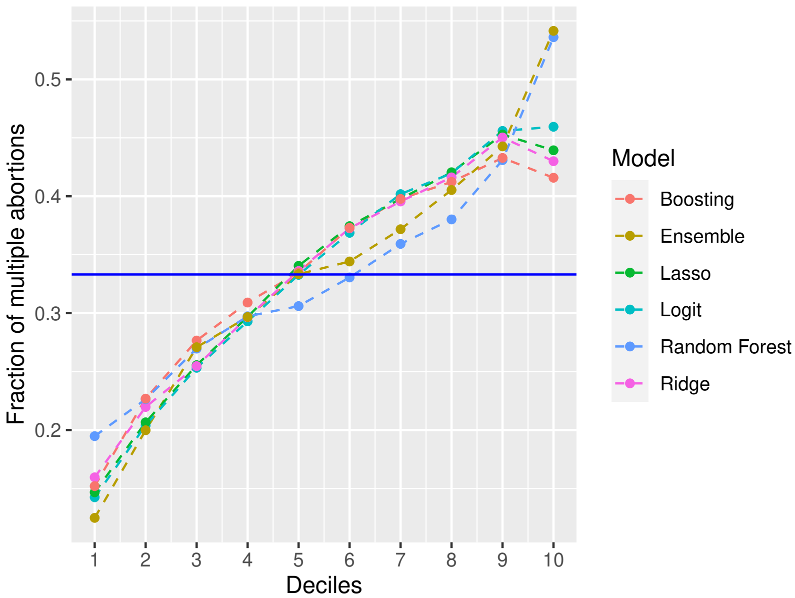

Figure A6: Out-of-sample performance of various models predicting the probability of a multiple abortion in a data set with all abortions (elective and non-elective). The test set has 185,685 observations.

A4.2 Dropping observations with missing predictors

As part of our robustness checks, we excluded all observations with any missing data. We lost approximately 56% of the observations from a total of 781,400, leaving us with 346,063 observations. We then repeated our analysis, as depicted in Figure A7. The performance in certain deciles shows modest improvement, while in others, it exhibits a decline when

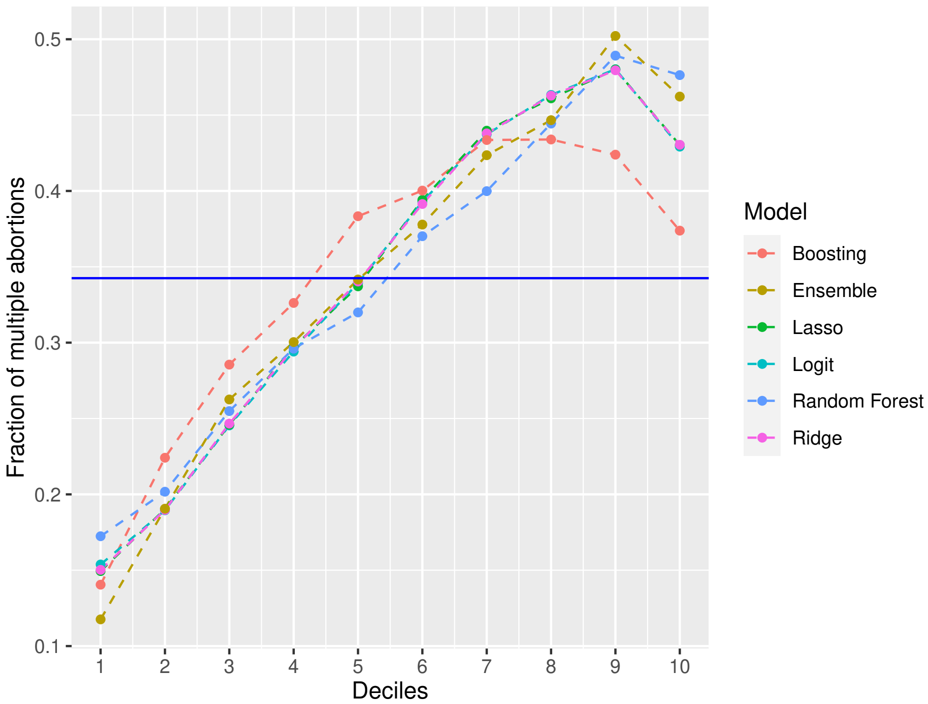

Figure A7: Decile analysis on out-of-sample data with no missing data. The horizontal blue line denotes the prediction based on the unconditional mean.

compared to the results demonstrated in Figure 3. In a separate examination of robustness, we removed the two variables—type of contraceptive and dependent children—from our model and conducted our analysis once again with the remaining eight variables. Figure A8 illustrates a decline in the performance of the methods, thereby underlining the pivotal role these two variables play in accurately predicting multiple abortions.

A4.3 Comparing Random Forest without and with dummy variables

In our preceding analysis, we employed a Random Forest model, where each variable is represented by a value determined by its numerical or non-numerical categorization. This configuration only utilized 10 categorical variables, with some categories representing missing values. As an additional robustness check, we also tested a variation of the Random Forest algorithm that used 82 binary dummy variables. Here too, some of the dummy variables represented missing values. In Figure A9 we compare the outcomes of these two different approaches. Remarkably, the results generated by both approaches are highly similar, rendering them almost indistinguishable. This suggests that our findings are robust, irrespective of whether we use a more condensed or expanded set of variables in our Random Forest model.

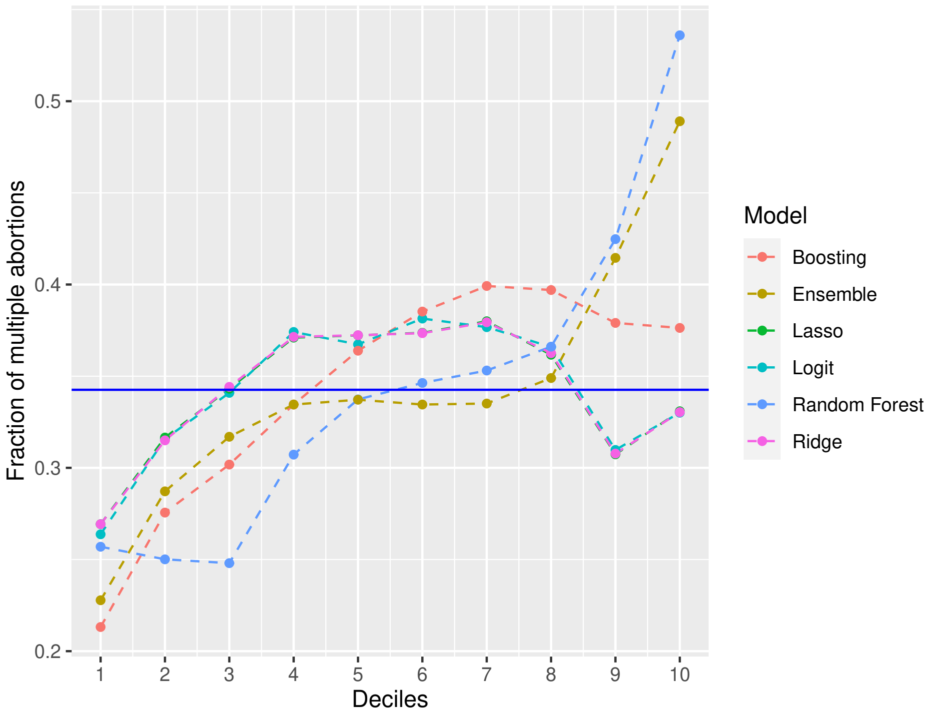

Figure A8: Decile analysis on out-of-sample data without two variables with lots of missing information: type of contraceptive and dependent children. The horizontal blue line denotes the prediction based on the unconditional mean.

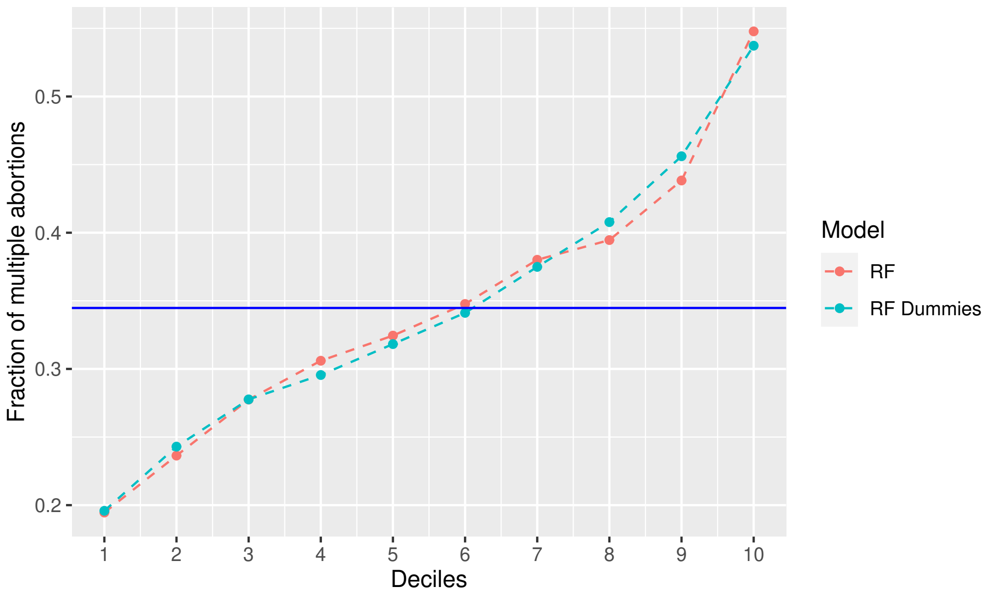

Figure A9: Decile analysis Random Forest without and with dummies.

A5 Top predictors of multiple abortions

Figure A10 displays the top-10 predictors of multiple abortions identified by the Random Forest model, with detailed tabular data presented in Table A3. Predictors are categorized into four groups based on their importance, measured by the average decrease in the Gini index across all trees. The most influential predictor is the province of residence, followed by the source of information about abortion. The third tier includes education level, foreign-born status, dependent children, age, public funding, and living status. The least influential predictors are the woman's employment status and the type of contraceptive used.

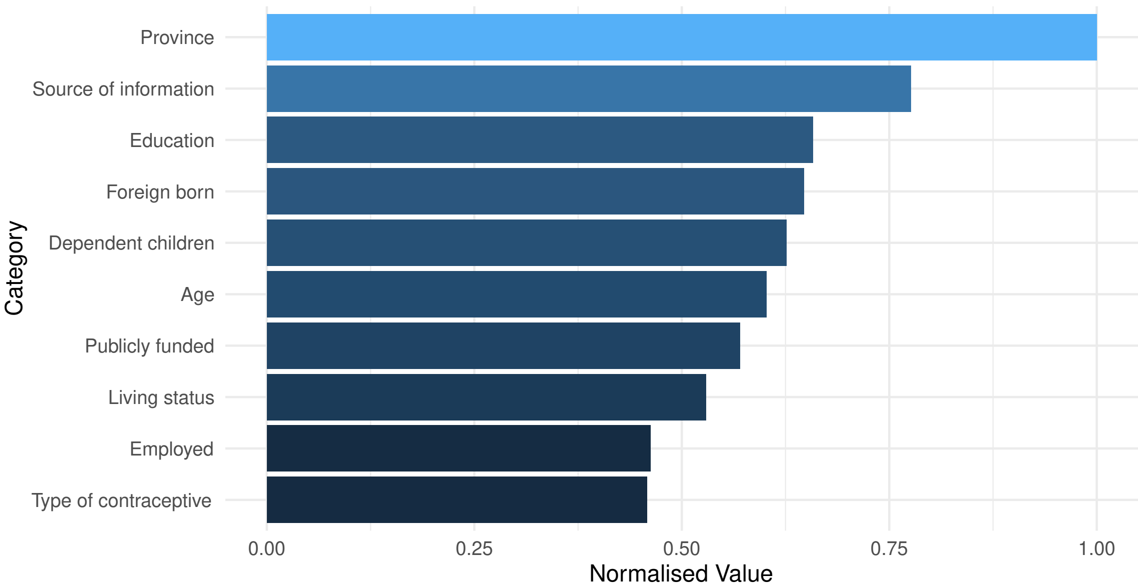

Figure A10: Importance of predictors using the Random Forest model. The most important variable is normalised to 1.

Figure A11 displays the top-10 predictors of multiple abortions according to the Lasso-Logit model, and Table A4 shows these findings in tabular format. These top predictors are selected from a total of 79 binary indicator variables, derived from 10 categorical predictor variables, as processed using the *glmnet* library in *R*.[9] All variables are standardized prior to analysis, and the importance of each is assessed by examining the absolute value of their coefficients. The Lasso-Logit model identifies age group dummy variables (except 15-19) as highly significant, positively influencing the likelihood of multiple abortions. Women with a university education are less likely to experience multiple abortions. Other significant predictors include the source of information from private health centers displaying a positive relationship, a dummy variable indicating a negative relationship with the province of La Rioja, and a missing indicator for the source of information showing a positive relationship.

[9]For the Lasso-Logit, Ridge-Logit, Logit and Extreme Gradient Boosting models, the 10 categorical predictor variables are converted into 79 binary indicators using the *glmnet* library in *R*. However, the Random Forest method directly utilizes categorical variables in tree construction with the *randomForest* library in *R*.

In summary, our analysis reveals that both the Lasso-Logit and Random Forest methods identify four common predictors of multiple abortions: a woman's age, province of residence, source of information about pregnancy termination, and education level. These factors not only highlight the unequal distribution of multiple abortions among the population of women but also suggest that targeted interventions should focus on these specific characteristics to effectively reduce multiple abortions.

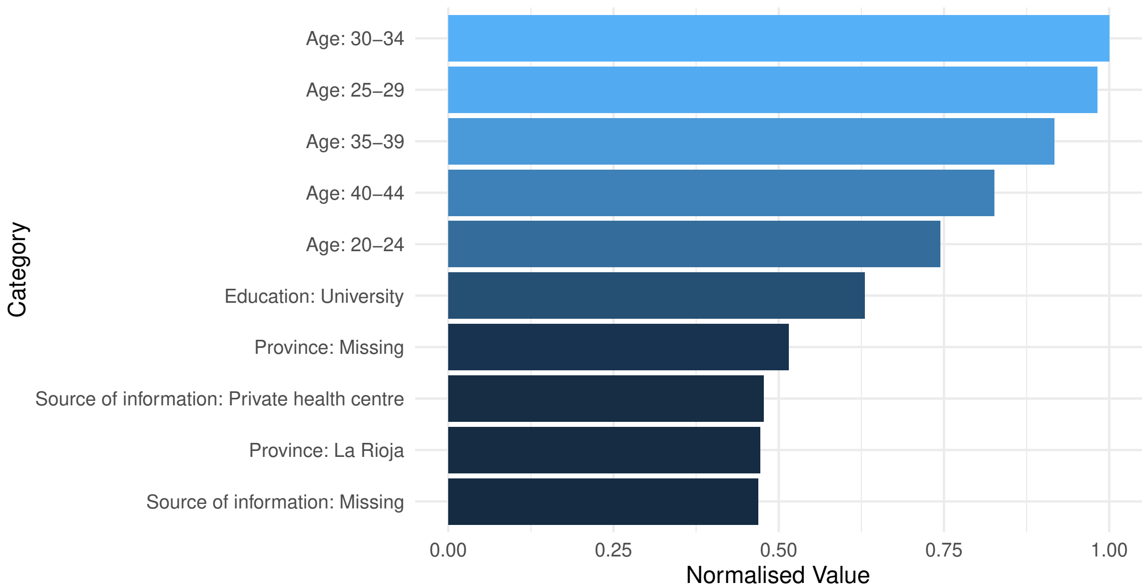

Figure A11: Importance of predictors using the Lasso-Logit model. The most important variable is normalised to 1.

Rank	Variable	Relative importance
1	Province	1.00
2	Source of information	0.78
3	Education	0.78
4	Foreign born	0.65
5	Dependent children	0.63
6	Age	0.60
7	Publicly funded	0.57
8	Living status	0.53
9	Employed	0.46
10	Type of contraceptive	0.46

Table A3: Variable importance for Random Forest method

Rank	Variable	Std. Lasso-Logit Coeff.	Relative importance
1	Age: 30-34	1.16	1.00
2	Age: 25-29	1.14	0.98
3	Age: 35-39	1.07	0.92
4	Age: 40-44	0.96	0.83
5	Age: 20-24	0.86	0.74
6	Education: University	-0.73	0.63
7	Province: Missing	-0.60	0.52
8	Source of information: Private health center	0.56	0.48
9	Province: La Rioja	-0.55	0.47
10	Source of information: Missing	0.54	0.47

Table A4: Variable importance for Lasso-Logit method

References

S. Athey and G. W. Imbens. Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725, 2019.

M. J. Bailey and J. M. Lindo. Access and use of contraception and its effects on women's outcomes in the u.s. Working Paper 23465, National Bureau of Economic Research, June 2017. URL http://www.nber.org/papers/w23465.

R. A. Berk et al. *Statistical learning from a regression perspective*, volume 14. Springer, 2008.

L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. page 785–794, 2016. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785.

D. Clarke and H. Mühlrad. Abortion laws and women's health. *Journal of Health Economics*, 76:102413, 2021. ISSN 0167-6296. doi: https://doi.org/10.1016/j.jhealeco.2020.102413. URL https://www.sciencedirect.com/science/article/pii/S0167629620310596.

E. Coast, A. Norris, A. Moore, and E. Freeman. Trajectories of women's abortion-related care: A conceptual framework. *Social Science and Medicine*, pages 199–210, 2018.

T. G. Dietterich. *Ensemble methods in machine learning*. Springer Berlin Heidelberg, 2000.

T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

C. Knowles Myers. The power of abortion policy: Re-examining the effects of young women's access to reproductive control. *Journal of Political Economy*, 125(6):2178–2224, 2017.

K. Kranker, S. Bardin, D. Lee Luca, and S. O'Neil. Estimating the incidence of unintended births and pregnancies at the sub-state level to inform program design. *PloS one*, 15 (10):e0240407, 2020.

P. B. Levine and D. Staiger. Abortion as insurance. Technical report, National Bureau of Economic Research, 2002.

J. M. Lindo, C. K. Myers, A. Schlosser, and S. Cunningham. How Far Is Too Far? New Evidence on Abortion Clinic Closures, Access, and Abortions. *Journal of Human Resources*, 55(4):1137–1160, 2020. URL https://ideas.repec.org/a/uwp/jhriss/v55y2020i4p1137-1160.html.

Z. Liu, Z. Lin, W. Cao, R. Li, L. Liu, H. Wu, and K. Tang. Identify key determinants of contraceptive use for sexually active young people: A hybrid ensemble of machine learning methods. *Children*, 8(11):968, 2021.

E. R. Mangone, V. Lebrun, and K. E. Muessig. Mobile phone apps for the prevention of unintended pregnancy: A systematic review and content analysis. *JMIR Mhealth Uhealth*, 4(1):e6, 2016. doi: 10.2196/mhealth.4846.

S. Mullainathan and J. Spiess. Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, 31(2):87–106, 2017.

A. Ng. Machine learning, coursera. *Standford University*, 2016.

C. Nicodemo, S. Oreffice, and C. Quintana-Domeque. Risk factors for repeat abortions and their spacing: Evidence from registry data in spain. *arXiv*, 2208.05567, 2022. doi: https://arxiv.org/abs/2208.05567.

S. Oreffice. Did the legalization of abortion increase women's household bargaining power? evidence from labor supply. *Review of Economics of the Household*, 5(2):181–207, 2007. URL https://EconPapers.repec.org/RePEc:kap:reveho:v:5:y:2007:i:2:p:181-207.

C. Pop-Eleches. The supply of birth control methods, education, and fertility: Evidence from romania. *The Journal of Human Resources*, 45(4):971–997, 2010. ISSN 0022166X. URL http://www.jstor.org/stable/25764796.

P. Probst and A.-L. Boulesteix. To tune or not to tune the number of trees in random forest. *The Journal of Machine Learning Research*, 18(1):6673–6690, 2017.

S. J. Russell and P. Norvig. *Artificial Intelligence: a modern approach*. Pearson, 3 edition, 2015.

D. Sansone. Beyond early warning indicators: High school dropout and machine learning. *Oxford Bulletin of Economics and Statistics*, 81(2):456–485, 2019.

B. M. Stifani, M. Peters, K. French, and R. K. Gill. There's an app for it: A systematic review of mobile apps providing information about abortion using a revised mars scale. *PLOS Digit Health*, 2(7):e0000277, 2023. doi: 10.1371/journal.pdig.0000277. URL https://doi.org/10.1371/journal.pdig.0000277.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.