Predicting Contraception Methods for Indonesian Families in 1987

Edie Espejo
Isabel Serrano
Asem Berkalieva

DS200: Principles and Techniques of Data Science
Graduate Version Final Project

13 May 2020

Abstract

To better understand access and use of contraceptives in Indonesia in 1987, a survey of married households collected data on various demographic and socioeconomic factors. Our paper uses logistic regression and random forests models to predict contraceptive use. Our first approach classifies between households that use either no, short-term, or long-term contraceptive methods. Our second approach uses a two-step binarization where general contraceptive use is predicted first, then contraceptive type is predicted thereafter. Accuracies were compared on training, cross-validation, and the final validation sets. Ultimately, we find that the two-step binarization approach and random forests perform best, where the regression approach identifies husband education, education gap, and estimated years married to be the factors most indicative of contraceptive use.

Introduction

Motivation

As early as 1965, Indonesia witnessed a large-scale effort to regulate population growth via a National Family Planning Program (1). This movement, which was largely motivated by the economic intent to raise the standard of living and avoid overpopulation, gained support from numerous community leaders, including local government and religious figures (2). **The National Indonesia Contraceptive Prevalence Survey of 1987** examined the success of the Family Planning Program. The survey determined that 95% of women were aware of contraceptive measures and 48% of married women used some form of contraception at the time of the survey (3). Despite this feedback and the substantial national decline of average children per family, 4 out of 10 women reported not using contraceptive measures despite wanting to delay or prevent births (3). Additionally, women popularly cited reasons such as lack of accessibility, financial burden, or general inconvenience as reasons for discontinuing use of contraceptives (3). Thus, although the National Family Planning Program showed success in lowering birth rates through contraception use, there was a clear need to improve availability of contraceptives as of 1987. Unsurprisingly, education and geography were features associated with the likelihood of engaging in family planning practices (3). In a similar vein, the type of contraceptives available were unevenly distributed across resources (3). Together, these features indicate unequal access to forms of contraception.

Description of Data

The Survey of 1987 conducted interviews of 11,884 women who were married at least once in their lifetime. The women spanned the ages of 15 to 49 and represented 93% of the Indonesian population, providing a sample representative of the country's varying socioeconomic conditions (3). For this project we use a subset of the data collected in this survey. By contrast, our data contains information for 1,473 women. This dataset contains information for women's age, level of education, work status, religion, husband's level of education and occupation, number of children, standard of living, media exposure to family planning, and contraceptive use. While the original dataset demonstrates a diverse and even sampling of women in Indonesia, our dataset contains an uneven distribution of women with respect to contraceptive use, education, standard of living, and religion (Figs. in Notebook). Specifically, the majority of women in our data have completed primary school, have a high standard of living, and are Muslim. Overall, this dataset provides a platform to study the relationship between socioeconomic factors and contraceptive use.

General Objective

Acknowledging that access and cost were factors influencing the use of contraception, we aim to understand how socioeconomic features impact contraceptive use. The goals of our project are two-fold: **1) predict contraceptive use through a variety of socioeconomic features 2) identify the best modeling approach for predicting contraceptive use.** We begin our work by understanding the available dataset and identifying how the composition of our sample impacts predictive ability. Based on additional information in the Survey of 1987 Summary Report (3), we introduce new modeling features that account for changes in family planning culture within the Indonesian population. Lastly, since the dataset is skewed and does not represent the diversity in socioeconomic factors, we vary our modeling approaches to address the resulting reduction in prediction accuracy. Broadly, we compare the predictive abilities of logistic regressions and random forests. Within our logistic regression approach, we model the data in two ways 1) by including the three categories of contraceptive use in our response variable and 2) in a two-step approach that binarizes the **general use** of contraception and then the **type** of contraceptive use. Ultimately through our modeling approaches, we aim to identify a bias in the Family Planning Program in order to improve the allocation of resources and outreach efforts regarding contraceptive use in the Indonesian population.

EDA

Subset Verification Fertility rate is the number of children a woman has in her lifetime. The 1987 National Indonesia Contraception Survey states that the average fertility rate was 3.3 children in 1987 which was an improvement from two years prior where the fertility rate was 5.5 (3). Our data subset has a mean fertility rate of 3.26 and a median of 3.0. A Wilcoxon Ranked Sign test for whether our data subset has an average of 3.3 yielded a p-value of 0.07, indicating a difference in average fertility rate between our sample and the survey population. Our subset has a mean of 3.36 with an SD of 2.36. This p-value is best explained by the removal of women below the age of 20 from our data.

Summary Statistics Within the training set (n=1,056), 57.84% of women use some form of contraception. The average wife age is 33.18 years and the median is 32 years. Our subjects are predominantly Muslim (83.81%), not employed (73.20%), and have good media exposure (92.23%). Most husbands and wives are equally educated (41.87%). 29.45% of husbands are more educated than the wives. The wives are more educated than the husbands in the remaining 3.62%.

Age and Number of Children (Figures 1-2)
Women of different age groups show different preferences for contraception methods. Seven equally-spaced age groups were created based on the standardized wife age values in **Figure 1**. The relative frequencies of contraception between age groups differ. In the younger age groups, no contraception and short-term contraception are used at similar rates. Older age groups predominantly did not use contraception. Notably, the middle age group (-0.0199, 0.489] shows a relatively uniform spread between all three contraceptive options. **Figure 2** shows the distribution of the number of children within each wife age group. Recall that the age groups are standardized and the minimum wife age in our dataset is 20 years. Clearly, women in older age groups have a larger range and more children on average. The R^2 value between the number of children and wife age is approximately 0.268. The relationship here does not seem to be linear as expected due in part to natural limitations of fertility.

Kids Per Married Year (KPMY) (Figures 3-6)
We are interested in the relationship between our engineered feature, KPMY, and the existing covariates of wife age and number of children. From **Figure 3**, we see KPMY has a linear offset for each level of number of children. There are also outliers where some women had children at a fast rate, evidenced by a high KPMY value. The R^2 value between these two variables is 0.060 which provides evidence for a nonlinear relationship. Figure 4 reflects the relationship between KPMY and wife age. The relationship between these two variables exhibits qualities of an exponential decay curve shifted across multiple starting points. The R^2 value between these two variables is 0.157 which provides evidence for a nonlinear relationship.

To motivate our selection of features for the prediction models, we conducted several tests for significance on our training set. Since our data were unbalanced (43.37% of women in the training set do not use contraception, 32.95% use a short-term method, and 23.67% use a long term-method) and represented counts, we chose nonparametric tests. KPMY per contraceptive group was visualized in Figure 5. The four distributions of KPMY per contraceptive group all have outliers, therefore are skewed right. We will use a Kruskal-Wallis test to test the hypothesis of whether or not the three groups share the same KPMY distribution. **The resulting p-value of 2.772e-25 leads us to conclude that there are distributional differences of KPMY between the three classes.**

Since we are also interested in the binary prediction between contraceptive use or not, we will test the hypothesis of whether KPMY is distributed similarly between these binary groups. Based on a Mann-Whitney U test, circumventing distributional assumptions, **we conclude that there are distributional differences for KPMY between the no contraception and the contraception groups (p-value=1.123e-20). Figure 5** shows the relationship visually.

Education Gap (Figures 7-8)

Education gap distributions per contraceptive method are not uniform as shown in **Figure 7**. When husbands are more educated than their wives, there is a higher proportion of women that do not use contraception. When the husbands and wives are equally educated, the distribution of preferences in contraceptive use is more uniform; however, there are still more subjects within the "None" group than the other two. When the wife is more educated, short-term contraceptives are most popular. It is unclear by the boxplots in **Figure 8** that the mean education gap magnitudes are the same between levels of contraception. However, we do see that the long-term group has more equally educated husbands and wives.

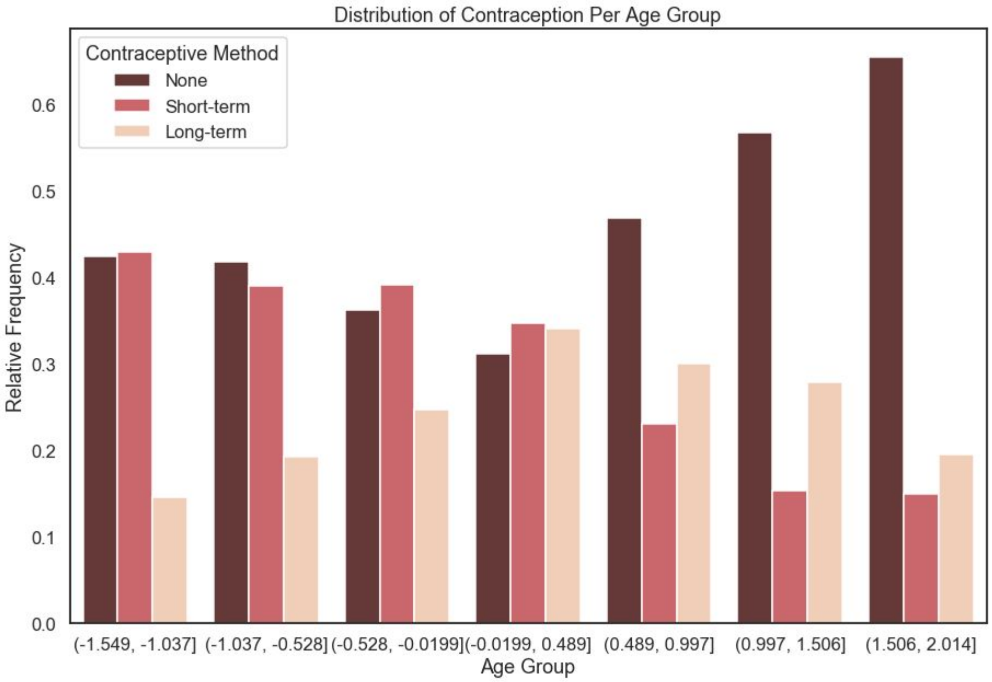

Figure 1

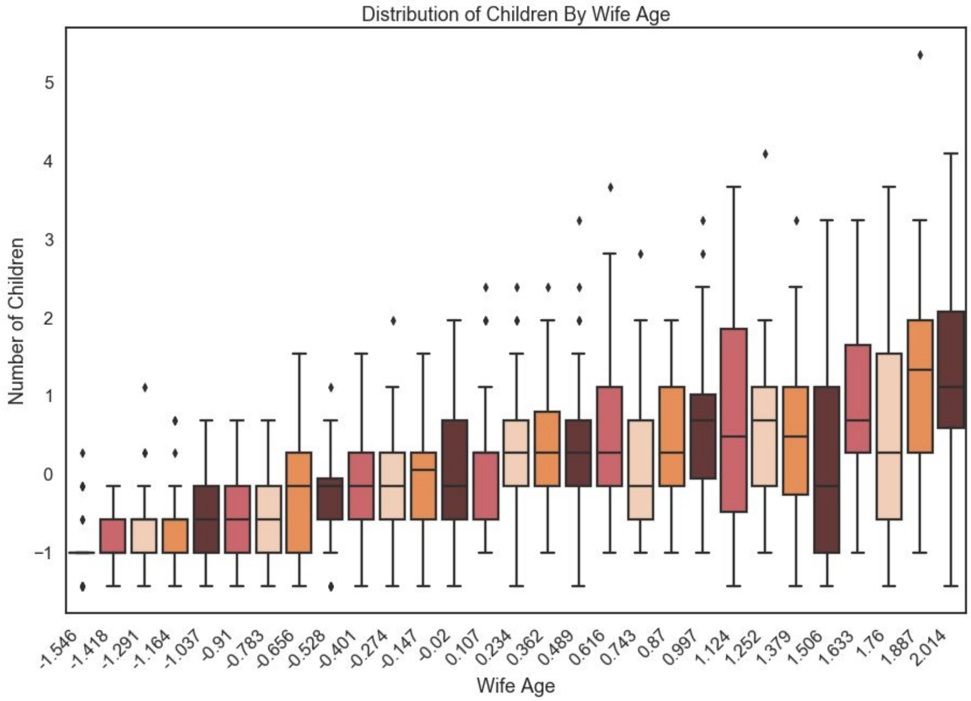

Figure 2

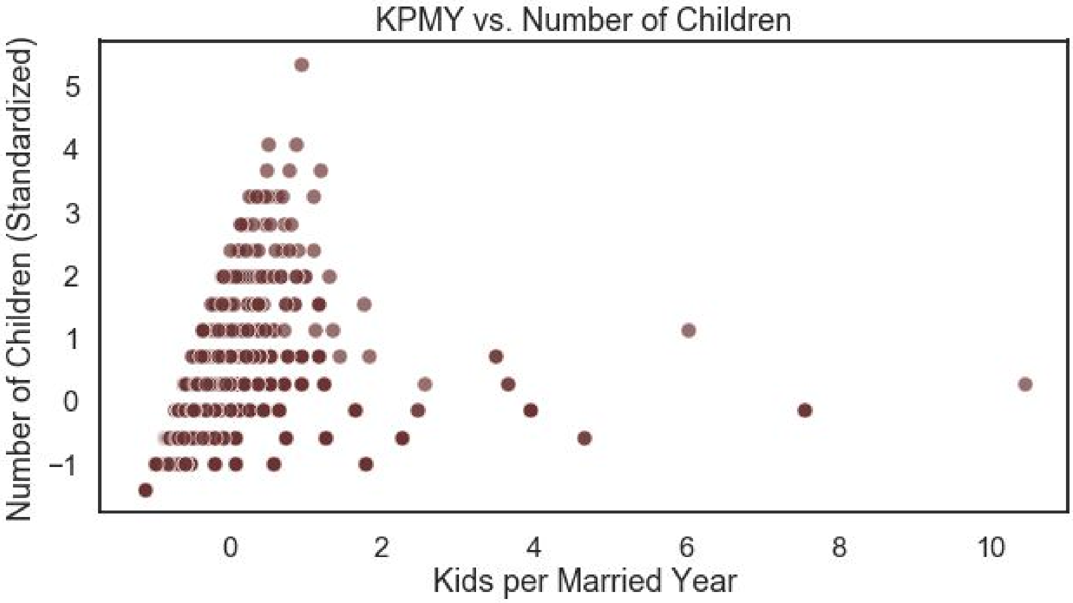

Figure 3

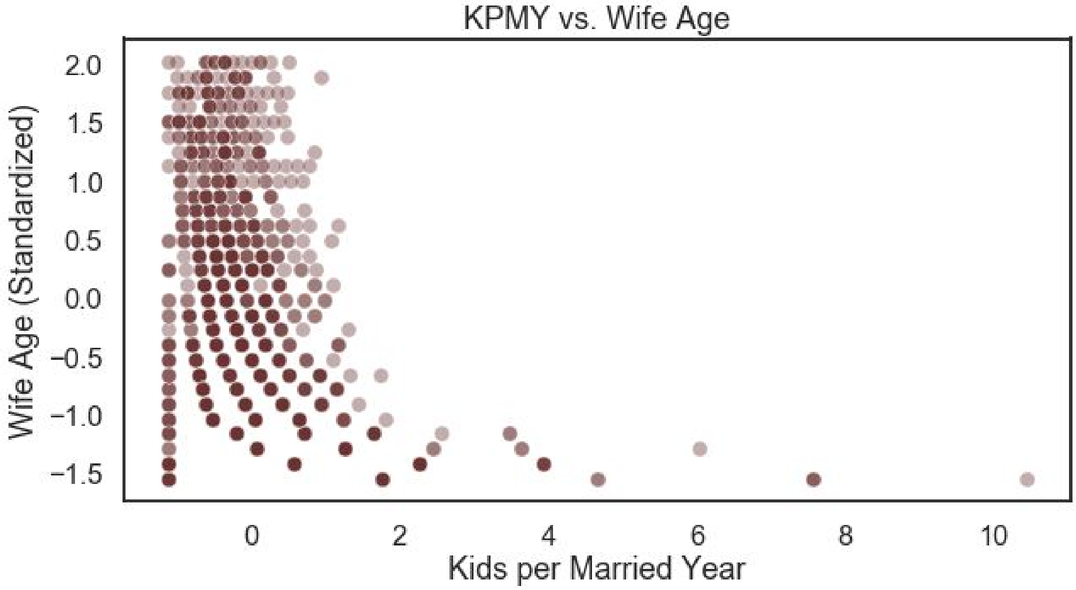

Figure 4

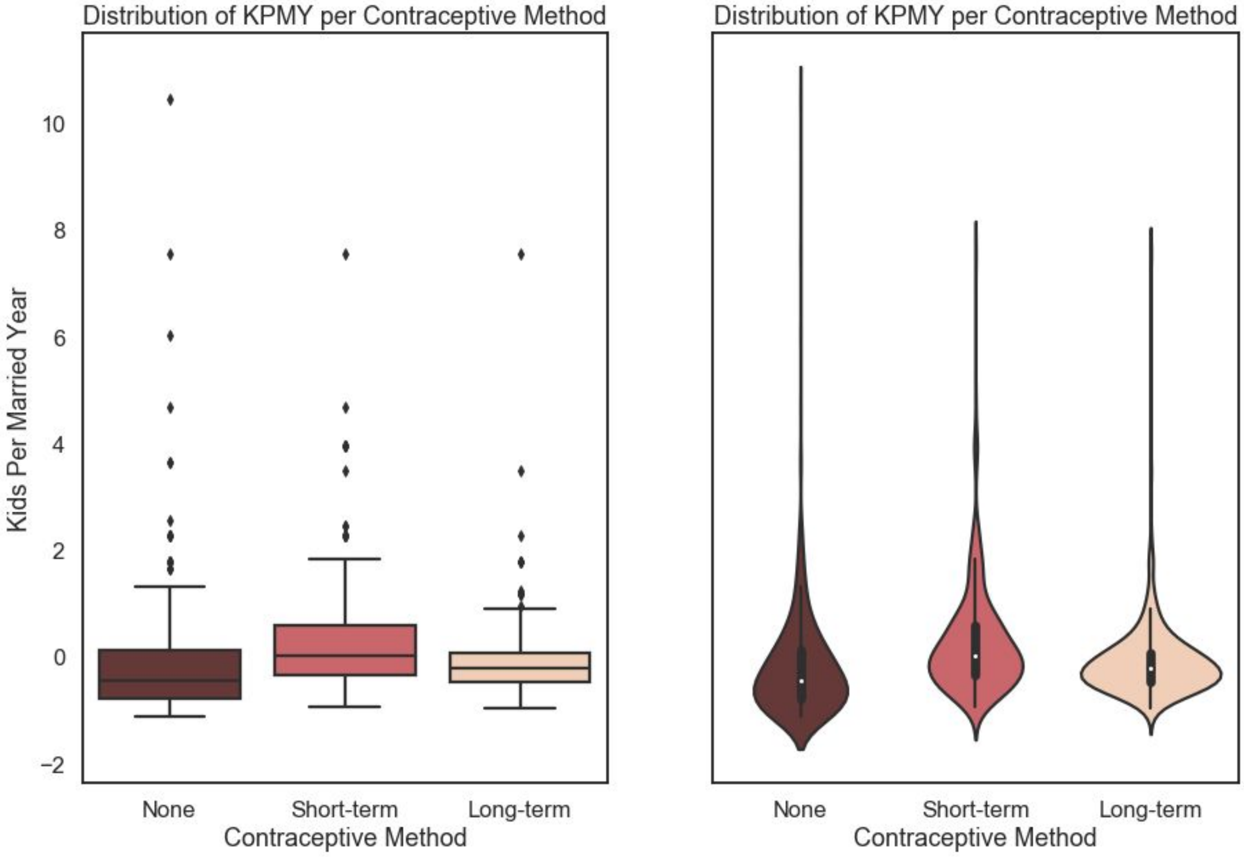

Figure 5

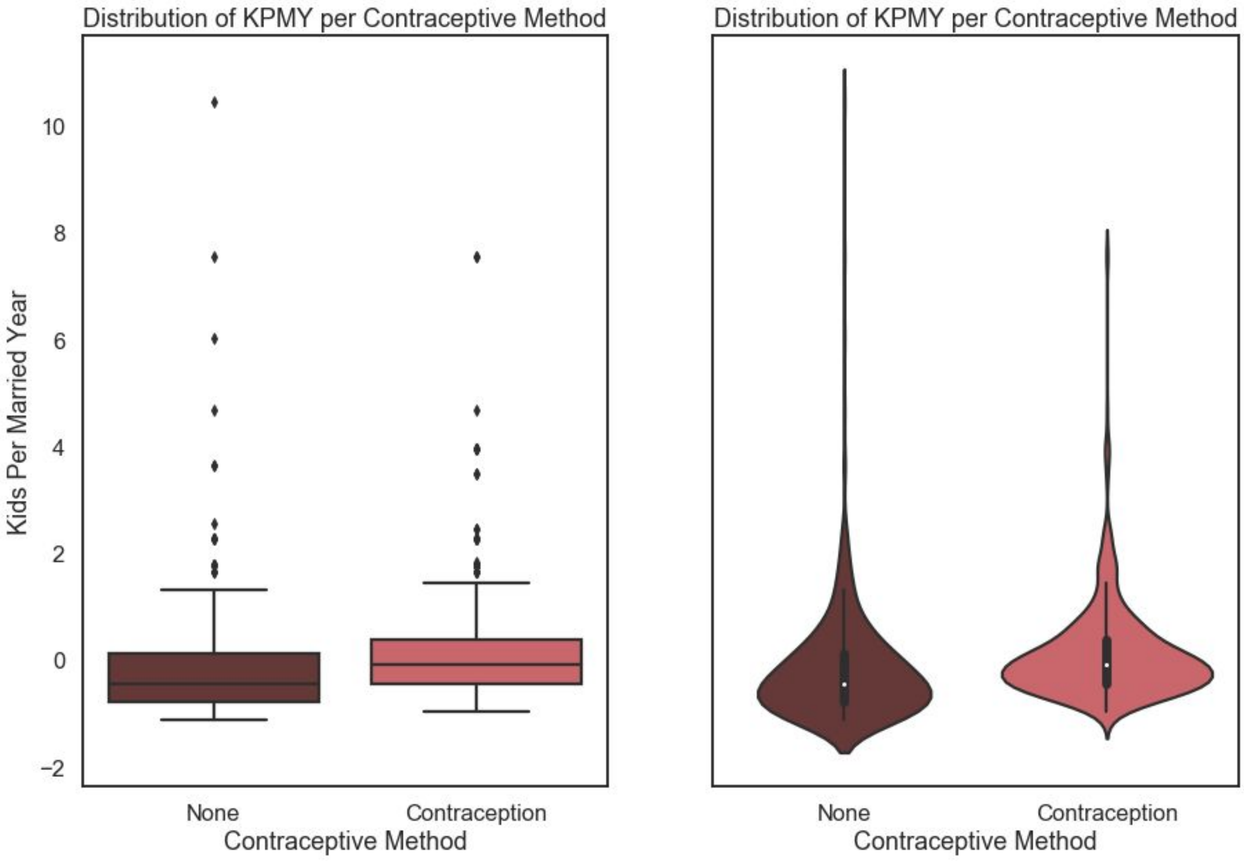

Figure 6

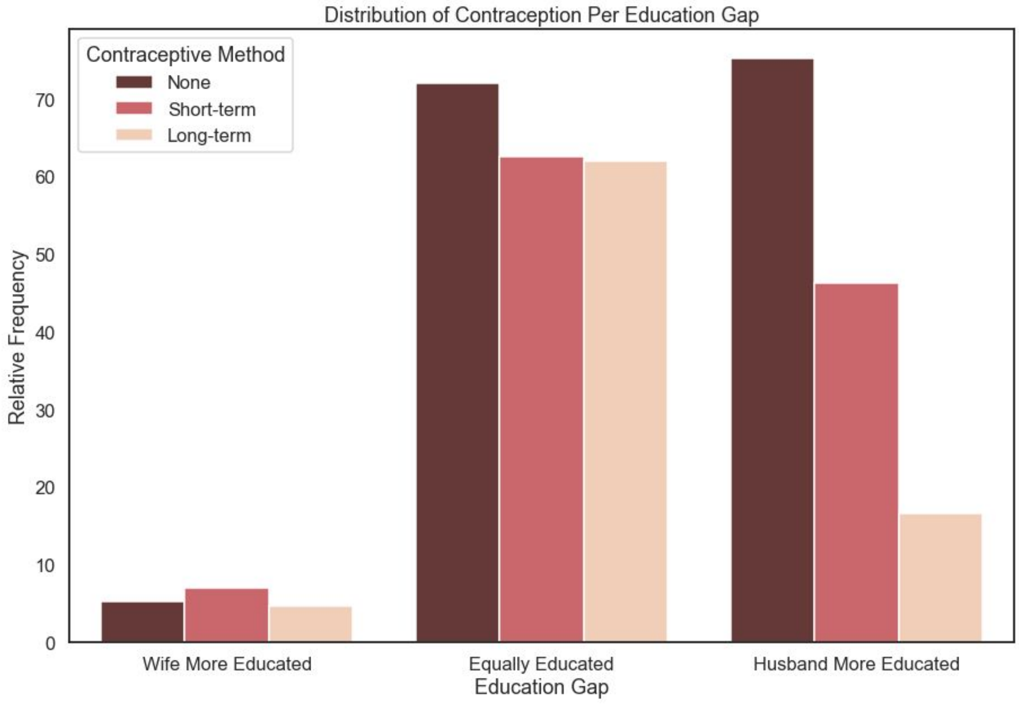

Figure 7

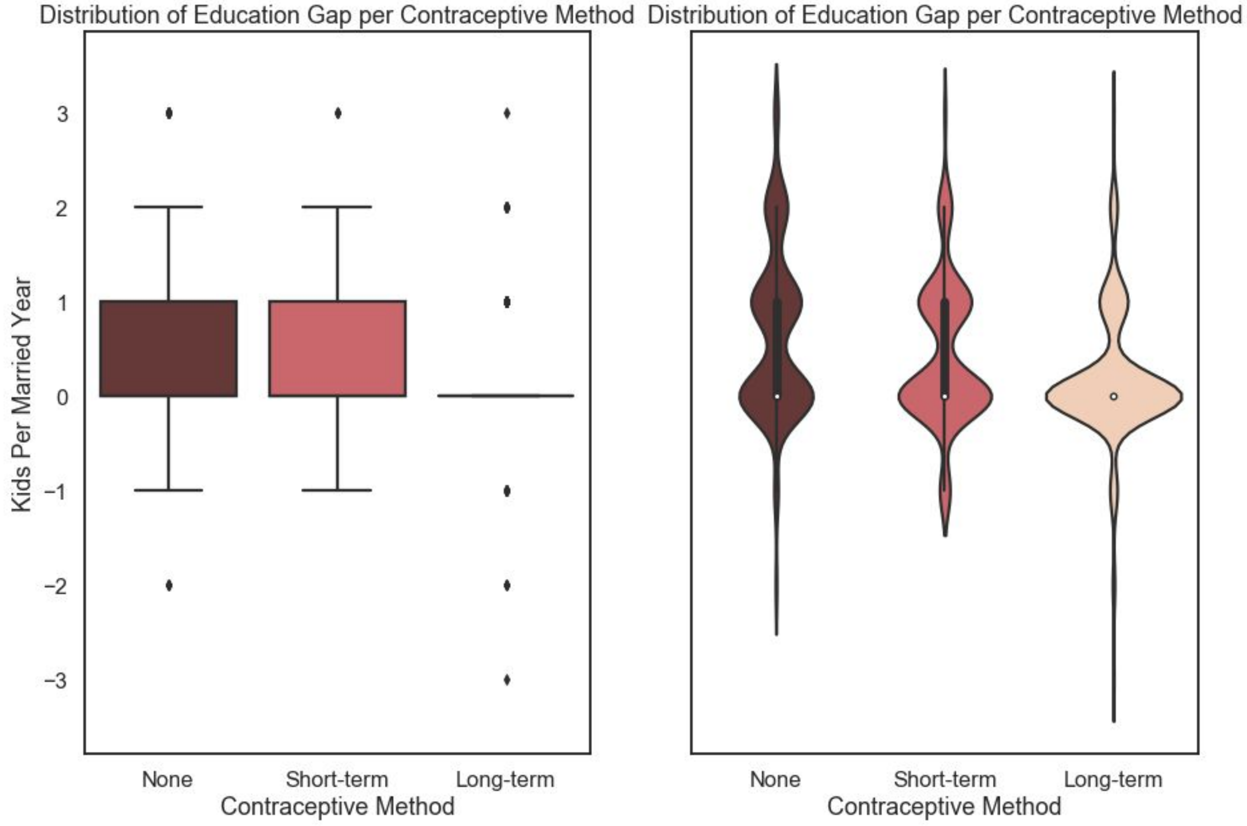

Figure 8

Methods

Data Cleaning

Our data was divided into training and test sets. The training set contained 75% of the data (1,104 records) while the test set encompassed the remaining 25% (369 records). In order to introduce a new feature (see below), we removed individuals below the age of 20 (64 records, 2% of the data). We ensured this filtering would not cause loss of information by conducting exploratory data analysis (EDA) and confirming that relationships between important features remained fairly similar. The “contraceptive” feature in the dataset was revised to reflect ordinal levels of contraception. In our revised dataset, a 0, 1, and 2 code for no, short-term, and long-term use of contraception, respectively. Standard of living was condensed from four to three levels, where middle-low and middle-high classes were merged into a single middle class. Overall, the original dataset was fairly organized and only called for recoding variables to suit our modeling needs.

Feature Engineering

To enhance our models’ predictive power, we engineered new features that captured the change in family planning practices among women across different age groups. We initially assumed that the number of children would be indicative of contraceptive use; however, our EDA revealed that the number of children per couple was distributed similarly across all three contraceptive method groups. We introduced the feature, **“kids per married year” (KPMY)**, which is defined as the number of children divided by the number of estimated years married. We used the estimated median age of marriage given in the Survey Summary Report to build this feature (3). While the number of children is a feature of the original dataset, our new feature takes into account changes in marriage patterns and the age at which women began having children. Moreover, the distribution of KPMY between the three contraceptive groups was significantly different according to a Kruskal-Wallis test, demonstrating the strength of our newly created feature.

Additionally, we considered the difference between levels of education among husbands and wives. This was calculated by subtracting the ordinal level of the wife’s education by the level of the husband’s education. In addition to the education gap, a numeric quantity, we were also interested in creating an ordinal variable that indicated who had the higher level of education. Thus, the **difference in education (ordinal variable)** is coded as -1, 0, and 1 if the wife has a higher level of education, there is no gap in education, or the husband has a higher level of education, respectively. Additionally, the **“wife education”** feature was converted into a binary variable, where 0 represents not having completed primary school while 1 indicates completion. The level of education was binarized since this difference in educational attainment has been associated with the likelihood of using contraception (3).

Finally, we applied one-hot encoding for categorical variables and standardized continuous variables. We one-hot encoded the education levels of husbands and wives. This created a binary variable for each category in the variables. We applied scaling for wife’s age, number of children, kid per married year, and estimated years married. Since the original features differed in ranges and variances, scaling was applied to ensure each feature equally contributed to the model.

Model Construction

We employed two approaches to predict contraceptive use: logistic regression and random forests. Both methods are used to classify categorical responses, but differ in their parametric and nonparametric nature, respectively. A logistic regression fits a model to predict the probability of some event occurring given the features. While in its simplest form, a logistic regression is used on binary data, it can be extended to model several classes of events. As a generalized linear model, the logistic regression is expressed as the following:

$$\hat{\mathbf{P}}_\theta(Y = 1 \mid x) = \sigma(\phi(x)^T\theta) \ = \frac{1}{1 + \exp\left(-\phi(x)^T\theta\right)}$$

where σ represents the parameterization vector. Assumptions for logistic regression include independent observations, a binary outcome (for binomial) or categorical outcome (for multinomial), little or no multicollinearity among the independent variables, and linearity of independent variables and log odds. While we can guarantee the first two assumptions have been met, we did not guarantee the third and acknowledge this as a limitation. Our analysis used both multinomial logistic regression when predicting from the three levels of contraceptive use (no use, short-term, long-term) and binomial logistic regression when predicting general use of contraception (use vs. no use) and contraception type (short-term vs. long-term). When a multinomial logistic regression is conducted for a 3-class classification, two simultaneous GLM equations are solved. The highest predicted probability is used to assign a class. We selected logistic regression because our outcome variable is categorical (and binary for the second two cases), the model returns probabilities for ease of interpretation, and it allows for understanding the relationship between the response and independent variables.

Unlike a logistic regression’s use of a linear decision boundary to classify a categorical variable, random forests use nonlinear boundaries. Random forests consist of building numerous decision trees, where each tree begins with all data in one node. The data is then subsequently split, where each split is chosen to minimize the model’s loss. Numerous trees are built in order to prevent overfitting of the training data. Each decision tree is applied to a bootstrapped sample of the training data; repeating this process is called bagging. The individual predictions are then averaged to determine the final model. A notable process that prevents overfitting is the sampling of features at each split, which is often the square root of the number of features. No formal assumptions exist for random forests, as it utilizes a nonparametric approach. We selected random forests because it is nonparametric, easy to interpret, and avoids overfitting.

We fit and assessed our models using the training data. To verify our model’s generalizability, we used cross-validation, a technique that prevents the risk of overfitting. The training data is partitioned into k-folds, where a model is fit to all but the kth fold. This is repeated until all folds have been used as a validation set once. Accuracies are averaged in order to obtain the total cross-validation accuracy. While random forest models select most features automatically, they do not tune the number of trees. Thus, we utilized cross-validation in order to select the number of trees for each random forest model.

For our three models, we selected features mimicking a leave-one-out approach and examined how features affected both the training and validation accuracies. In our approach, we fit the logistic regression model to all features before iteratively removing one feature in each successive model (See Notebook for Figs.). **Features that negatively impacted training and validation accuracies were omitted from the final model.** The list of the features we included in our model is shown in **Table 1.**

Table 1. Features Used For Each Prediction Outcome

	Multinomial (k = 17 features)	Binary: No use vs. use (k = 18 features)	Binary: Short vs. long-term (k = 16 features)
Features used	Kids per married year, Wife education, Husband education (4)[1], Categorical educational gap, Husband occupation (4), Wife religion, Wife work, Standard of living (2), Media exposure, Estimated years married	Kids per married year, Wife education, Husband education (4)[1], Categorical educational gap, Husband occupation (4), Wife religion, Wife work, Standard of living (2), Media exposure, Estimated years married, Number children	Kids per married year, Husband education (4)[1], Categorical educational gap, Husband occupation (4), Wife religion, Standard of living (2), Media exposure, Estimated years married, Number children

[1] Numbers in parentheses signify how many total features are used due to one-hot encoding.

Results

Model Accuracy Comparison

We fit two model types (logistic regression and random forests) to three prediction types: 1) all contraceptive choices 2) general use and 3) type of contraception used. Training, cross-validation, and test accuracies are presented in **Tables 2 and 3** for all prediction outcomes and each model, respectively. The multinomial model displayed the lowest accuracy scores, while the two-stage binary model and random forests demonstrate similar accuracy scores. Random forests returned higher accuracies than logistic regression models did for all three prediction outcomes. Despite attaining a high accuracy on the training sets, random forests clearly overfit despite its design to avoid overfitting. This overfitting is demonstrated by the stark contrast in accuracy between the train and test sets. When alternating the number of trees in our model, overfitting was still a prevalent issue **[Figure 9]**. Thus, while random forests return the highest training and test accuracies, the model's reliability may be subject to debate.

Model Assessment

For the multiclass prediction setting, we examined the medians and standard deviations of the predicted probabilities per subject. This measure explains how definitive our model was at predicting their classes. For example, the row `[0.629948, 0.257109, 0.112943]` has a median of 25.71% and standard deviation of 21.78%. Based on the boxplot of medians below, the median is quite small. Based on the boxplot of standard deviations of all predictions (see notebook), a definitive row of predicted probabilities such as this example has a high standard deviation in comparison to the rest of the predictions. **These two pieces of evidence show that our model is predicting probabilities that are quite uniform, demonstrative of the model's difficulty in classifying contraceptive use correctly.**

The precision and recall for the two-stage binary and random forests models were calculated.
The logistic regression model had a 67.32% precision rate and a 85.22% recall rate. The random forests model demonstrated a 70.80% precision rate and a 78.82% recall rate. These curves were compared to a "no skill" predictor, which predicts classification probabilities to be the sample average. **Both of our models predicted better than predicting at the average.**

The exponential coefficients for the logistic regression model were analyzed to assess the influential socioeconomic factors in contraceptive use. The multinomial logistic regression showed that husband education, median marriage age, estimated years married, and education gap were important predictors of contraception methods. Examining the general use of contraception features such as husband education, standard of living, and number of children were most important. Features important for predicting contraceptive type used were estimated years married, education gap, and husband education had relatively high coefficients. **Streamlining these results, we notice that husband education, the education gap, and estimated years married were influential in predicting contraceptive use**.

Table 2. Logistic Regression Model Accuracies for Each Prediction Outcome

	Multinomial	Binary Use vs. No Use	Binary Short-term vs Long-term
Training Accuracy	0.561	0.711	0.683
CV Accuracy	0.564	0.701	0.655
Test Accuracy (CV)	0.466	0.669	0.655

Table 3. Random forests Model Accuracies for Each Prediction Outcome

	Multinomial (n = 25 trees)	Binary Use vs. No Use (n = 25 trees)	Binary Short-term vs Long-term (n = 15 trees)
Training Accuracy	0.953	0.976	0.941
Test Accuracy	0.513	0.700	0.621

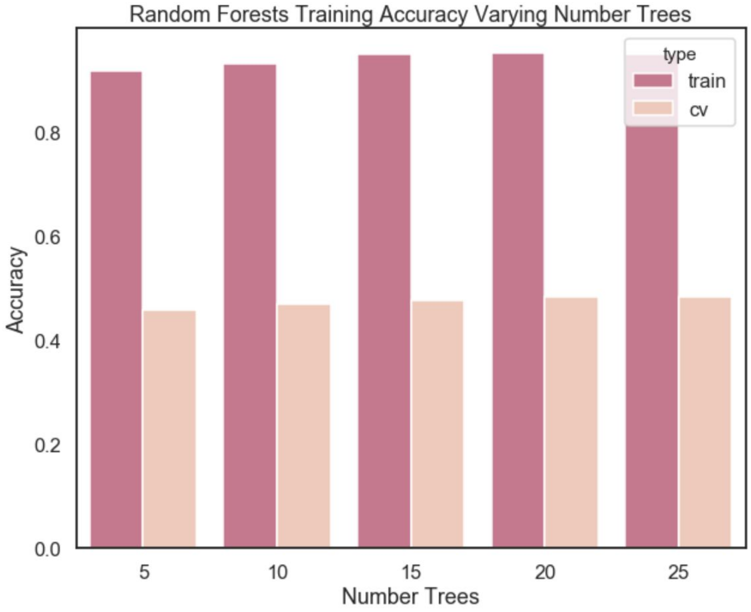

Figure 9

Discussion

Through our “leave-one-out” method of feature selection, we determined that median age married and kids per married year (KPMY) were consistently features that improved our accuracy score. Notably, these covariates were features we engineered in order to reflect generational and cultural gaps between different married couples that were not captured in the original features (e.g. older women on average married at a younger age compared to younger women in our data). Given public access to the Survey of 1987 Summary Report (3), we employed trends captured in the summary statistics, allowing us to estimate the number of years married per couple and initialize when women began having children (assuming age at marriage and age at first birth were correlated, as suggested in (3)). As described in the Feature Engineering Section, capturing cultural changes in birth rates was key to identifying features that held distinct distributions when stratified across contraceptive use. Contextually, KPMY may be more reflective of socioeconomic differences rather than the number of children a woman has.

While our novel features served to capture the variability in outcome variables, the exclusion of wife age and inclusion of wife religion was surprising. As with the number of children, we initially assumed that wife age and religion may play a role in predicting contraceptive use, particularly since opinion on contraceptives varies across generations. However, since we introduced KPMY and median age at marriage into our dataset, wife age did not serve to be efficient as a standalone feature. This result may arise from the redundancy of wife age with KPMY and median age at marriage. Secondly, while we assumed the skewed representation of Muslim women in our data would bring down predictive power, the inclusion of wife religion in combination with other features consistently improved our models' accuracies.

Although we were able to increase the logistic regression's predictive power through our two-step binarization approach, undoubtedly the most restricting limitation in our efforts was the sample size and unbalanced dataset. We can examine that the unbalanced dataset affects our predictive ability by comparing the discrepancy in our recall (85.22%) and precision (67.32%) rates for the binary model. Since our dataset has a 20% higher proportion of women on contraceptive use, we are able to identify women who are on a contraceptive method, but not detect false positives (i.e. women who are not using a form of contraceptive). Moreover, our dataset did not provide an even representation of various socioeconomic features. For example, most of the women in our dataset were Muslim (reflecting Indonesian culture), unemployed, had good media exposure, and had at least completed primary school. This unbalanced representation prevents us from fully capturing the relationship between socioeconomic factors and contraceptive use.

Acknowledging the impact a skewed dataset would have on predictive ability and the limited number of original features, we addressed these issues by consolidating our data to restore an even distribution and created novel features that captured cultural changes. In consolidating our data, we assumed that women using contraceptives shared similar features. Rather than accounting for the skewness through consolidation, moving forward we can apply sampling techniques, such as undersampling and oversampling to reweight the dataset and improve model performance. Additionally, to estimate the number of years each subject has been married, we removed women who were younger than 20 years of age resulting in a loss of information. We could however make an assumption about the age of marriage for women under 20 based on further literature and keep these records in our data. As a future work, we can also explore different modeling techniques such as k-nearest neighbors (KNN), support vector machines (SVM), and linear discriminant analysis (LDA) or ensemble learners such as SuperLearner. These models can all classify categorical outcomes and do not require distributional assumptions. Thus, the use of these models may be more beneficial for our analysis since we did not guarantee all assumptions were met for the logistic regression model. Additionally, each model has its own benefit in reducing error. It would have been helpful to run more models and compare across the results.

While the dataset is successful in maintaining the anonymity of the survey subjects, there are undoubtedly ethical dilemmas concerning bias and stereotyping when including socioeconomic factors as features in a predictive model. Although domain knowledge can be helpful in feature selection, the inferred relationship of these features and the outcome variable must be carefully interpreted to avoid demonizing and ostracizing groups (e.g. people belonging to a specific socioeconomic bracket or religious group). Acknowledging these issues, we aim to prevent bias in our feature selection by applying an objective method to assess which features contributed to accuracy rather than relying on biased preconceptions. Moreover, in assessing the features associated with contraceptive use we communicate our findings in a way that emphasizes the key demographic factors with the intent of equalizing access to contraceptive use. Ultimately, we strive to keep our study objective and centered on encouraging equity among the Indonesian population.

Altogether we obtained informational features from the National Survey, but lacked geospatial residency data. Notably, the survey mentioned that the type of contraceptives were skewed, with community posts and pharmacies most likely supplying short-term contraceptive use (4). Data on the women's residency, particularly in regards to rural versus urban classification, would allow us to test the relationship between contraceptive use and the type of contraceptive use with accessibility. These analyses would complement our current findings by identifying specific geographic regions that need access to contraceptive measures and determining the types of contraceptives that are lacking across regions.

Concluding Remarks

Ultimately, we trained three models that demonstrated varying levels of accuracy on our dataset. Random forests also provide the highest accuracy scores despite their overfitting. With regards to our logistic regression modeling efforts, our two-step binary approach demonstrated consistent accuracy scores and higher accuracy than the multinomial logistic regression. This two-step binary approach is designed to identify socioeconomic features that influence use of contraception, while still understanding factors that influence the distribution of contraception types. Moreover, through the use of a regression we identified features (husband education, education gap, and estimated years married) that determined the likelihood of using contraceptive measures. By including more features, such as geospatial information and the distribution of community posts in Indonesia, we can build on our conclusions to improve the equity and accessibility to contraceptive use.

References

1) Reese, T., Suyono, S., & Suyono, H. (2006, May 23). The Indonesian National Family Planning Program. Retrieved May 13, 2020, from https://www.tandfonline.com/doi/abs/10.1080/00074917512331332792
2) Reese, T. (1975, November). The Indonesian national family planning program. Retrieved May 13, 2020, from https://www.ncbi.nlm.nih.gov/pubmed/12334277
3) National Indonesia Contraceptive Prevalence Survey 1987 ... (n.d.). Retrieved May 13, 2020, from https://www.dhsprogram.com/pubs/pdf/SR9/SR9.pdf