A Study on Indonesian Contraceptive Data

Authors: Alexandra Novales, Richa Bhattacharya, and Sophia Sousa

May 13, 2020

Abstract

The purpose of this analysis was to see if the number of children a woman has or her contraception use can be modeled. By training linear, logistic, Decision Tree, and Random Forest models, as well as using L1 (Lasso) and L2 (Ridge) regularization methods, we explored the effectiveness of each model in predicting the number of children a woman has or the type of contraceptive used. On each model, we used cross validation to better understand how the model might work on new data of the same type. After utilizing the data analytic techniques we learned in class, we found that the data provided is not enough to accurately model the number of children a woman has or the type of contraceptive used. The data does, however, show interesting trends that may be useful for future analysis, such as the relationship between a woman's age, number of children, and the type of contraceptive used. Ethical issues that might arise in exploration of this problem include invasion of women's privacy and/or inequitable access to family planning tools such as education and/or a variety of contraceptives.

Introduction

This report provides an analysis on a subset of data from the 1987 National Socioeconomic Survey in Indonesia, which looks at information pertaining to each household surveyed including, but not limited to, household census, income, education status, and occupation status. The sample we analyzed was called the 1987 National Indonesia Contraceptive Prevalence Survey and focused more on contraceptive use and fertility rates. The survey was performed in order to help expand the country's current databases, as information regarding family planning and fertility behavior in Indonesia was lacking. After reviewing the data, we proposed the following questions to guide our analysis:

1) To what extent do the demographics presented in the National Indonesia Contraceptive Prevalence Survey serve as an accurate measure to predict the number of children a woman has?

2) To what extent do the demographics presented in the National Indonesia Contraceptive Prevalence Survey serve as an accurate measure to predict their choice of contraceptive?

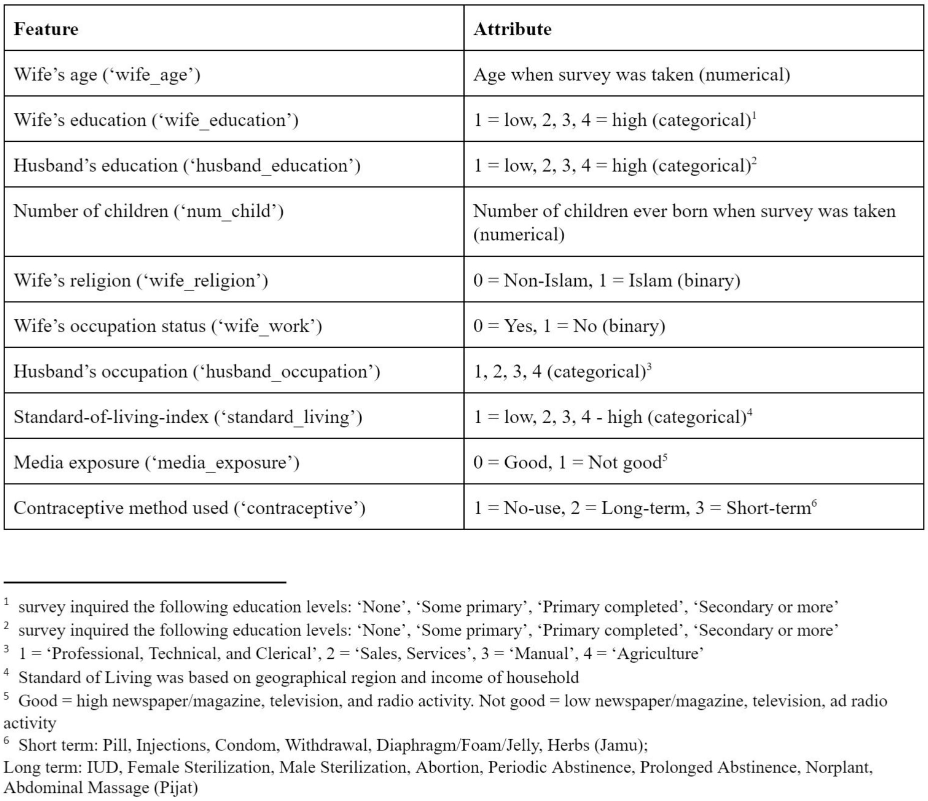
The following sections will cover our data, methods, and results in detail.

Description of Data

As stated in the Introduction, this sample was a subset of Indonesia's 1987 National Socioeconomic Survey. The unit of analysis was per household, children under five years, all-ever married women ages 15-49, and men.

According to the survey design and its source, about 93% of Indonesia's total population was represented in the account. Multiple reasons as to why the survey was taken include, but are not limited to, providing Indonesia with a database and analysis useful for informed choices, to expand Indonesia's international population and health database, and to provide data on the family and fertility behavior of the Indonesian population necessary for program organizers and policymakers to estimate, measure, and study factors that account for changes in fertility and contraceptive rates.

The sample dataset we studied included these select features: 'wife_age', 'wife_education', 'husband_education', 'num_child', 'wife_religion', 'wife_work', 'husband_occupation', 'standard_living', 'media_exposure', 'contraceptive'. A legend of the feature and its attributes are described below. with additional information in the appendix.



Feature	**Attribute**
Wife's age ('wife_age')	Age when survey was taken (numerical)
Wife's education ('wife_education')	1 = low, 2, 3, 4 = high (categorical)[1]
Husband's education ('husband_education')	1 = low, 2, 3, 4 = high (categorical)[2]
Number of children ('num_child')	Number of children ever born when survey was taken (numerical)
Wife's religion ('wife_religion')	0 = Non-Islam, 1 = Islam (binary)
Wife's occupation status ('wife_work')	0 = Yes, 1 = No (binary)
Husband's occupation ('husband_occupation')	1, 2, 3, 4 (categorical)[3]
Standard-of-living-index ('standard_living')	1 = low, 2, 3, 4 - high (categorical)[4]
Media exposure ('media_exposure')	0 = Good, 1 = Not good[5]
Contraceptive method used ('contraceptive')	1 = No-use, 2 = Long-term, 3 = Short-term[6]

[1] survey inquired the following education levels: 'None', 'Some primary', 'Primary completed', 'Secondary or more'

[2] survey inquired the following education levels: 'None', 'Some primary', 'Primary completed', 'Secondary or more'

[3] 1 = 'Professional, Technical, and Clerical', 2 = 'Sales, Services', 3 = 'Manual', 4 = 'Agriculture'

[4] Standard of Living was based on geographical region and income of household

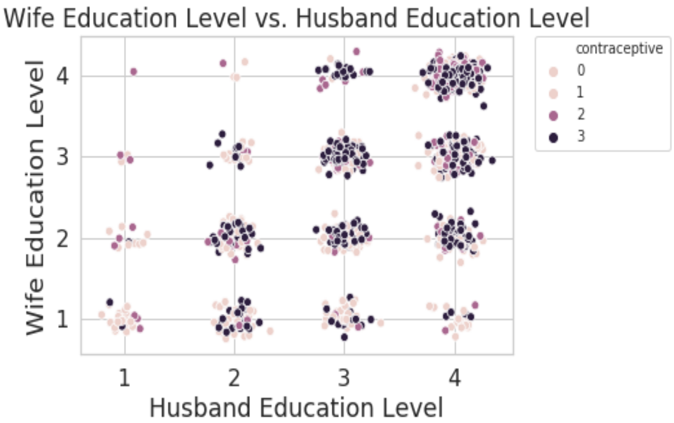
[5] Good = high newspaper/magazine, television, and radio activity. Not good = low newspaper/magazine, television, ad radio activity

[6] Short term: Pill, Injections, Condom, Withdrawal, Diaphragm/Foam/Jelly, Herbs (Jamu);
Long term: IUD, Female Sterilization, Male Sterilization, Abortion, Periodic Abstinence, Prolonged Abstinence, Norplant, Abdominal Massage (Pijat)

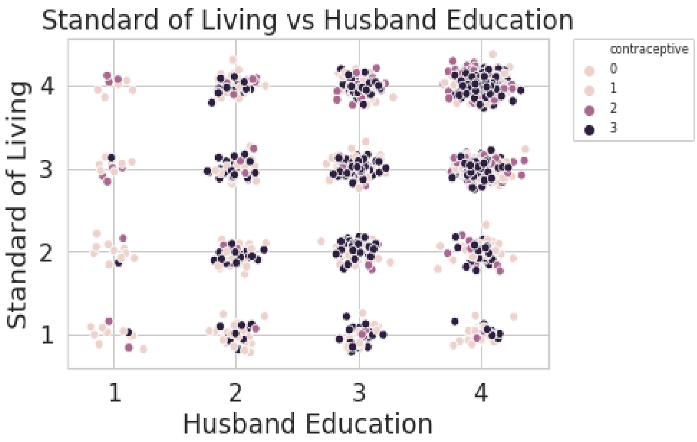
Description of Methods

Data Cleaning and Exploratory Data Analysis

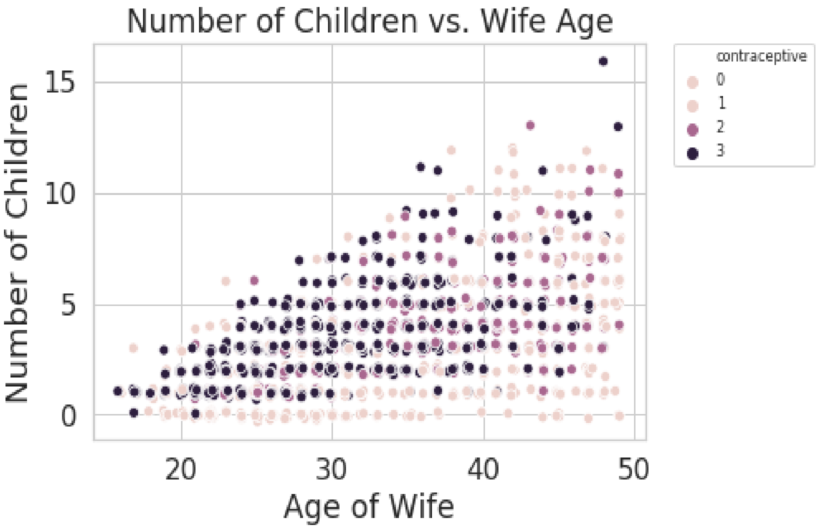
First, we began by cleaning our data. This included checking for null values, outliers, incorrect types, and any value that was not within its respective range (i.e. all values for binary categorical data were either or 1). We then utilized exploratory data analysis by creating visualizations relevant to our questions and given features.



We see the lower-right diagonal half is more populated with points than the upper-left half. This could mean that in a marriage the husband's education is more important compared to the wife's. We also see that contraceptive use is more common as both education levels increase, with some exceptions.



Here we see that despite the standard of living, if a husband's education is lower, the use of contraceptives is less likely. The visualization also shows that if the husband's education is either a 3 or 4, the standard of living is likely to be higher as well.



The graph to the left shows that women with no children tend not to use any contraceptive. It also shows that women with more children tend to use some type of contraceptive. Also, older women tend to use short-term or no contraceptives. This might be because older women have gone through menopause, erasing the need for long-term control.

******more visualizations are available in our Jupyter Notebook******

Based on these visualizations, we finalized the data and moved on to feature creation and selection. We created one additional feature based on the ceiling of the average of two other features. Because many of the features were categorical, we used one hot encoding to encode categorical data as real numbers. This way, the magnitude of each dimension is meaningful.

Feature Selection

The feature selection process involved discussing which features could be related and/or useful, based on our own background knowledge, visualisations created during EDA, and curiosity. We created a new data frame including a new feature that averages the education of the household (‘avg_household_education’), a one-hot-encoded version of the categorical features ('wife_education', 'husband_education', 'husband_occupation', 'standard_living', 'contraceptive', 'avg_household_education'), the original non-categorical features, and the original binary categorical features. We used this data frame to choose which feature combinations might work best with each model. After testing various combinations using different models with different features, we found that using all the features from our new data frame yielded the highest training accuracy across models.

Creating Models

Because we posed questions that involved predicting two different variables--contraceptive type and number of children--we explored a variety of models. In order to keep our features organized, we created two new data frames: one to predict the number of children (data_linear), and another to predict the type of contraceptive used (data_logistic).

First, we trained a linear regression model, imported from scikit-learn, in order to predict the number of children a wife in Indonesia might have. This model is suitable, since our features are linearly independent, and we are predicting numerical data--number of children. Then, we computed the root mean squared error (RMSE) of our model to evaluate how far our predictions were from the original data. We trained several more linear models with different combinations of features; however, the lowest root mean squared error of the linear models was the model trained with all features including average education in place of husband and wife education. We also found the average validation RMSE of each model; again, the lowest validation RMSE belonged to the same model with the lowest train RMSE.

Second, we trained a logistic regression model, also imported from scikit-learn, in order to predict the type of contraceptive a woman in Indonesia might use. We chose this model because we hope to predict a multiclass feature--the different types of contraceptive. We found the training accuracy for different logistic models trained with different feature combinations. The model trained with all the features yielded the highest training accuracy; however, the training accuracy of this model did not satisfy our standards for a *good* model. In fact, the validation accuracies were close to the training accuracies, further emphasizing the ineffectiveness of these models.

Initially, both types of regression models were not accurate in their predictions of the number of children a wife might have or the type of contraceptive a wife might use. As a result, we decided to incorporate regularization.

The two regularization techniques used were Lasso and Ridge Regression. Each of these are useful in strengthening our linear models. For both, we used cross validation to find the best tuning parameter for each model. As we learned in lecture, we used L1 regularization to find the most useful features. This technique deems unnecessary features as “unuseful” by setting them equal to 0. As a result, we dropped those features from our training data, and used the more refined training set to create a new linear model. Unexpectedly, L1 regression did not reduce our accuracy by a significant amount.

Finally, we tried Decision Trees and Random Forests. Decision trees are suitable in classifying multi-class features, and Random Forest models use Decision Trees. Both were used to predict the type of contraceptive used or number of children a woman had. Both had extremely high training accuracies but with low cross validation accuracies. This is likely due to the fact that Decision Trees and Random Forests are notorious for overfitting.

those features from our training data, and used the more refined training set to create a new linear model. Unexpectedly, L1 regression did not reduce our accuracy by a significant amount.

Finally, we tried Decision Trees and Random Forests. Decision trees are suitable in classifying multi-class features, and Random Forest models use Decision Trees. Both were used to predict the type of contraceptive used or number of children a woman had. Both had extremely high training accuracies but with low cross validation accuracies. This is likely due to the fact that Decision Trees and Random Forests are notorious for overfitting.

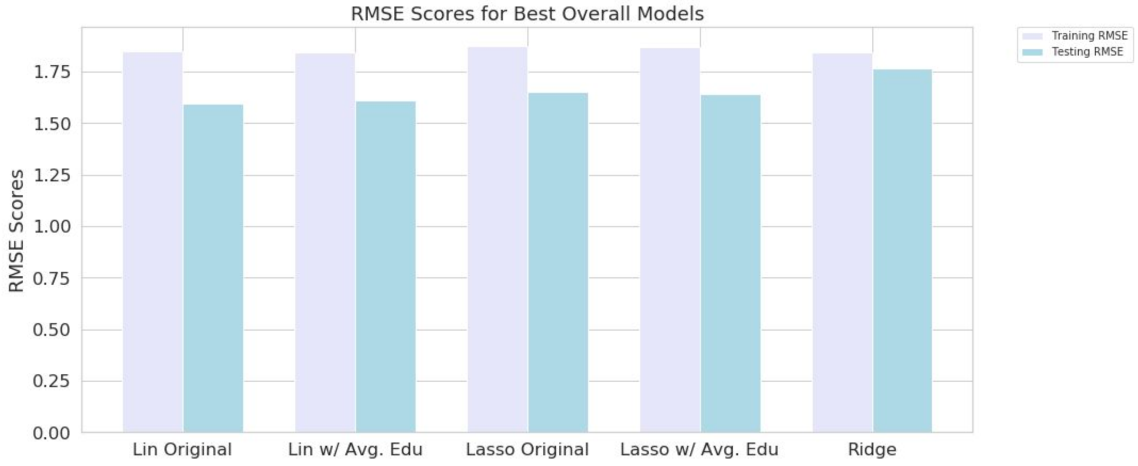
Summary of Results

NOTE: The test error was not computed until a "best" model was found for each prediction; the test error was calculated solely for visualization purposes.

When predicting the number of children, we attempted to use OLS and regularization. L1 regularization was used in order to find the "best" features for our linear model. Contrary to what we expected, the RMSE of the linear model trained with only the "best" features was higher (by 0.01) than the linear model with all the features. It's possible Lasso didn't work in the way we hoped because we have limited features. Lasso regularization works best when there are many features that can be narrowed down. We would need more features to effectively use L1 regularization.

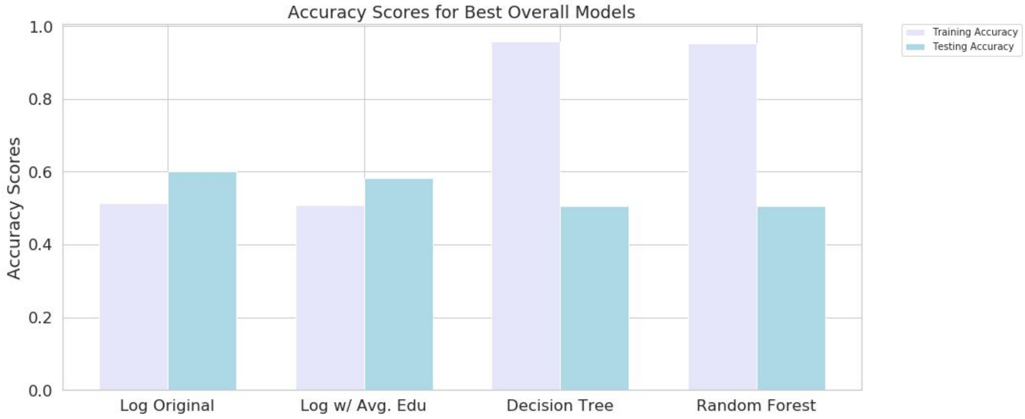
The lowest RMSE of the linear models was ~1.8456. This model included all of the original features and average education, except for husband and wife education. Using Ridge regularization, the RMSE lowered to ~1.8433. Again, this regularization technique did little to reduce our RMSE; regardless, the Ridge model was the best predictor of how many children a woman has. This RMSE means that for every prediction our model creates, it is about 2 children off of the actual number of children. While an RMSE of 1.8433-1.8456 may initially seem low, we decided 2 children was too significant of an error, and therefore the model is not a decent predictor.

The bar chart below shows the different RMSE scores across each model. The lavender bar on the left represents the train RMSE, and the blue bar on the right represents the test RMSE. The test RMSE of each model was taken after choosing Ridge as the best model for predicting the number of children, and are only shown to demonstrate the true accuracy of our models. As expected, the test RMSE is lower than the train RMSE for each model.



When predicting the type of contraceptive used, the logistic regression models created had a low training accuracy, ranging from 41.9% to 51.3%. Again, the highest accuracy was a result of using all features. Overall, the highest training accuracies came from using Decision Trees and Random Forests, which yielded accuracies of 95.8% and 95.2%, respectively. We deemed Random Forests as our “best model” for predicting contraceptive type used, because of its high training accuracy and its utilization of many Decision Trees. The cross validation and test accuracies of these models, though, tell a more realistic story of the accuracy of our model. The average cross validation scores were 48.8% for the Decision Tree and 53.6% the Random Forest, and the test scores were both 50.5%. It is interesting to see that despite the Decision Tree and Random Forests’ high training accuracies, their test accuracies are similar to that of the training accuracies of our logistic models.

Similar to the bar chart above comparing RMSEs, the following bar chart compares the train and test accuracy of the models used to predict contraceptive type. Again the test accuracies were computed after deciding on a model, and are only shown to visualize the accuracy of our models on new, similar data. It’s interesting to see that the logistic models have higher test accuracies than those of our Decision Tree and Random Forest models. This highlights the tendency of Decision Trees and Random Forests to over fit data--the training accuracies may be high but the testing accuracies are significantly lower.



Despite trying a variety of techniques to create a decent model, it seems that our data is restrictive. Initially, we hoped to create two models: one that would predict the number of children a woman has, and another that would predict the type of contraceptive a woman uses. None of the models we created met our standards of a “good model”--the RMSE scores were too high and the accuracies were too low. When the accuracies were high, we questioned their validity. This led us to believe that maybe the issue did not lie within our models, but within the features to which we had access. Perhaps if we had access to more information that could be translated into additional features, our models would have made better predictions. In addition, we acknowledge the fact that our data relies on the responses and decisions of real life women who are entitled to their own opinion and are not confined to the trends of their demographics. Also, while studying the survey, the questionnaire was taken in person -- affecting the way women would answer. As a result, predicting the number of children and types of contraceptives is too difficult given our data and the methods to combat this are beyond our knowledge. Future analysis can better answer the original questions posed.

Discussion

During our process of analyzing and modeling the contraceptive data provided, we took note of certain trends and features we found particularly interesting. For example, the feature ‘media_exposure’ stood out in that it seemed vague and relatively unimportant. After studying the official survey more, however, we found that the inquiry of media exposure helped show which households had more access to or the desire to expose themselves to magazines/newspapers, television, or radio. When accounting for database expansion and family planning predictions, this survey question makes sense, because at the time of the survey, the media was beginning to play a larger role in everyday life and entertainment. Unfortunately, our Lasso Regression technique classified ‘media_exposure’ as unnecessary in predicting the number of children a woman might have.

Another interesting feature was ‘husband_education’. In our Wife Education vs Husband Education plot, we found that the husband’s education was more important than the wife’s education. In other words, there were more husbands with level 3 or 4 education than wives were. We also found that the higher the husband’s education level is, the more likely a woman is to use contraceptives in general. This makes sense seeing that more educated couples probably have a better background in safe family planning.

During our modeling process, we tried different feature combinations. We were hopeful that ‘wife_education’ would play an important role--surprisingly, we were wrong. As seen in our code, no combination of features seemed to have any shocking effects on our models’ accuracies, including the feature combination only using attributes related to women. Bold of us to think that women's education could have played a significant role in the late 80s. Given the vast amount of factors that could go into predicting the number of children or contraceptive use, it was difficult for us to find additional data or create new features. In the end, the only new feature added was the average education between the husband and wife of a household. As seen in our notebook, the addition of this feature only decreased our RMSE by ~0.04. Another issue we encountered was choosing the right model to begin with. Because the initial accuracies and RMSEs of our linear and logistic models were so low, we decided to create and evaluate every model we could. With this being our first fully independent analysis, it was difficult to interpret where our models went wrong or see what data we were missing due to lack of experience.

Our data analysis was limited to the features provided and the models we learned in class. It was difficult to create new interesting features based on the given features, as most of the attributes were categorical. Perhaps other models would have suited our data better, for example, parabolic/quadratic models might have been useful. During our exploratory data analysis, we noticed that the older a woman was, the more likely she was to use long-term contraceptives; however, after around the age of 45, we noticed that many women reverted back to using no protection. This could potentially be indicative of menopause or a similar natural phenomenon that decreases the need for contraceptives-- affecting our model. In addition, women who didn’t have any children were more likely to not use contraceptives. This relationship among women who don’t use contraceptives can be seen in our Number of Children vs Wife Age graph, in a half-U shape. As for assumptions made, we assumed that the features were linearly independent from one another and confirmed this by taking the rank of our covariate matrix. We also assumed that if any of the women were pregnant at the time of being surveyed, the children in her womb were not included in the summation of the total number of children she had. In addition, we assumed that every woman had equal access to the different kinds of contraceptives.

Moving forward, one ethical dilemma we faced is the data is reflective of real women. Oftentimes, data analyses restrict its subjects to numbers, columns and rows. Our data, however, has a story behind it. Each row is a person and each column describes an aspect of their life. People are entitled to their own choice of contraceptive and the number of children they have, and creating a model that predicts this based on a few features should not disregard one's personal "choice". It felt wrong to assume that, based on certain demographics, a woman might want *x* type contraceptive or *x* number of children, because everyone is different and cannot be restricted to a model. Contraceptive use is extremely sensitive information and some women may not want their information studied in this context. Another factor to consider, is that not all women can physically have children, and this can result in the exclusion of women in our data. Although the survey took place 30+ years ago, one way to address this, should future studies occur, is to ensure confidentiality (given that the data was anonymous to begin with) and that analyses (including ours) are purely for scientific research. We also realized that not all women may have had equitable access to safe contraceptives, despite wanting or needing it. Through our data and further analysis, we can see who is more likely in need and provide those demographics with useful information and resources regarding safe family planning.

Ideally, more data would help strengthen our analysis. For example, the age of women may have an underlying effect on our models. It seemed that women younger than 20 and older than 40 make major differences in our predictions. If more women were surveyed, we could group the results by age and make our predictions on those separate groups. Our data is also from 1987 Indonesia--it might be more relevant to have recent data in our analysis. This way, we can evaluate trends over time. In addition, knowing the fertility rates of Indonesia over time could help us understand why women were having a certain number of children or used contraceptives. Other potentially useful information could be whether the women planned to have the number of children they had or whether they had access to contraception.

Due to the structure of the class, each assignment had some sort of guidance on what to analyze and how to analyze it. In this project, however, we were especially challenged on how/where to start. As a group, our knowledge was limited to what was taught in lecture and our own experiences, so it was difficult to find solutions to our growing problems--we probably hadn't learned how to solve them yet! Regardless, this project helped challenge our current critical thinking skills and gave us practice in completing the workflow used by real data scientists: evaluating whether or not models are "good" based on the original data, finding accuracies <90% is acceptable, and, most important, always searching for ways to improve a model, even though it adds more challenges and time spent on the analysis.