Data 100 Final Project Report

Andrew Liu, Terrance Wang - Spring 2020

Part I: Question/Framing

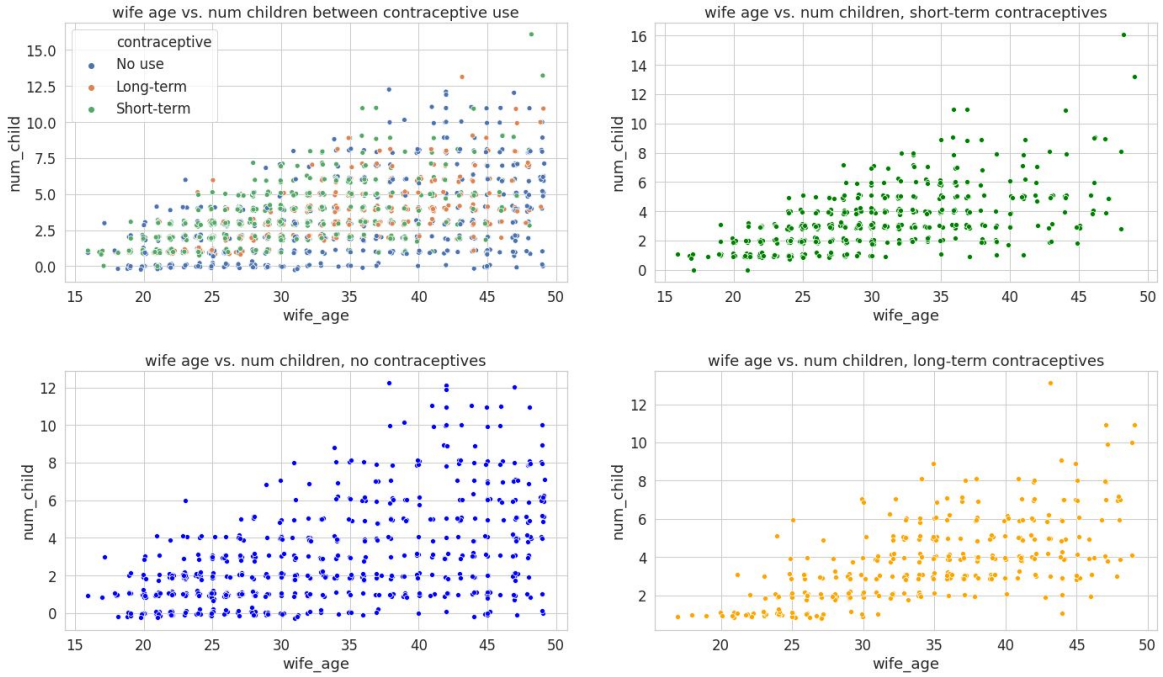
For our final project, we decided to use the Contraceptive Method Choice data set from 1987 National Indonesia Contraceptive Prevalence Survey. The question we wanted our model to address was: “What family socioeconomic factors and models can be used to best predict contraceptive method choice?”. We wanted to answer this question because by better understanding contraceptive method choice, government policy makers can focus policy solutions directed at addressing the determinants that play the largest factor in contraceptive solutions in order to prevent unwanted pregnancies and improve maternal and child health.

Part II: Data Cleaning

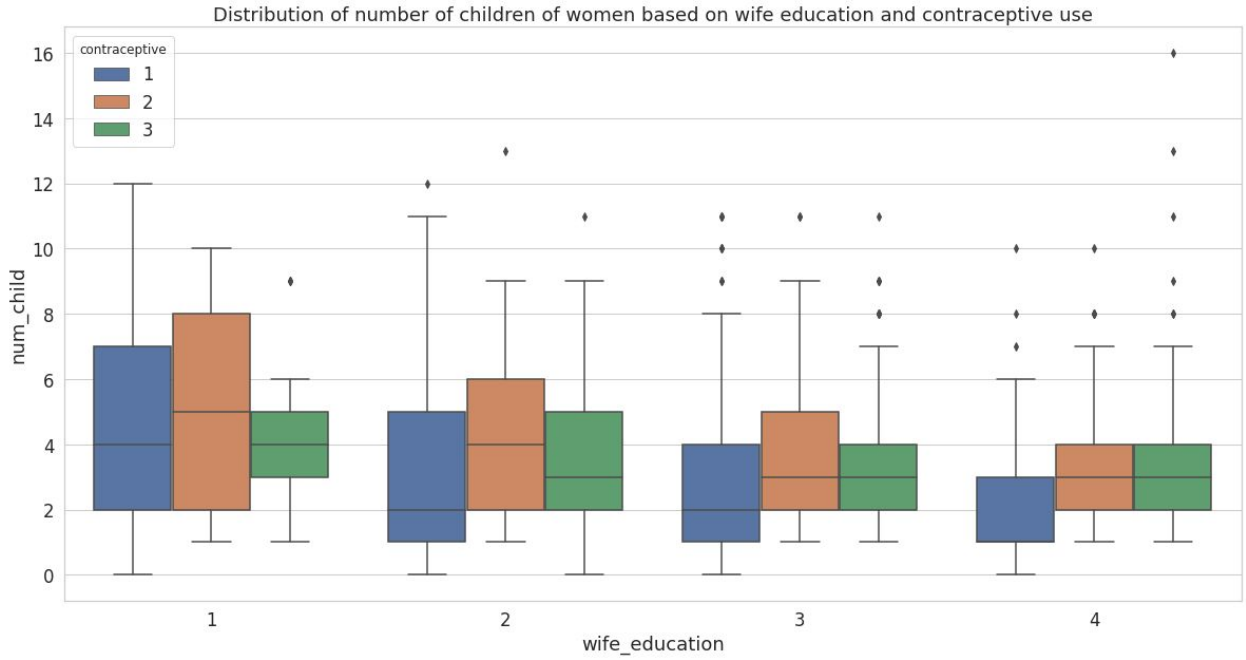
We examined our data using pandas and determined that the data does not have any NaN and null values that required filling. We first decided to replace the values in the contraceptive column with their real values, 1=No-use, 2=Long-term, 3=Short-term. We used the pandas “get_dummies” function to one-hot encode the categorical variables in our data set. Additionally, we added to our data by squaring our num_child and wife_age data in order to try and add features that would separate our data more. Based on our EDA, we saw that the data tended not to be clearly separated between the different contraceptive methods, so we hoped that by adding more features we could improve our model prediction accuracy. Lastly, we standardized our quantitative variables, wife_age, num_child, wife_age^2, and num_child^2, by using the sklearn preprocessing tool StandardScaler.

Part III: Data Visualizations

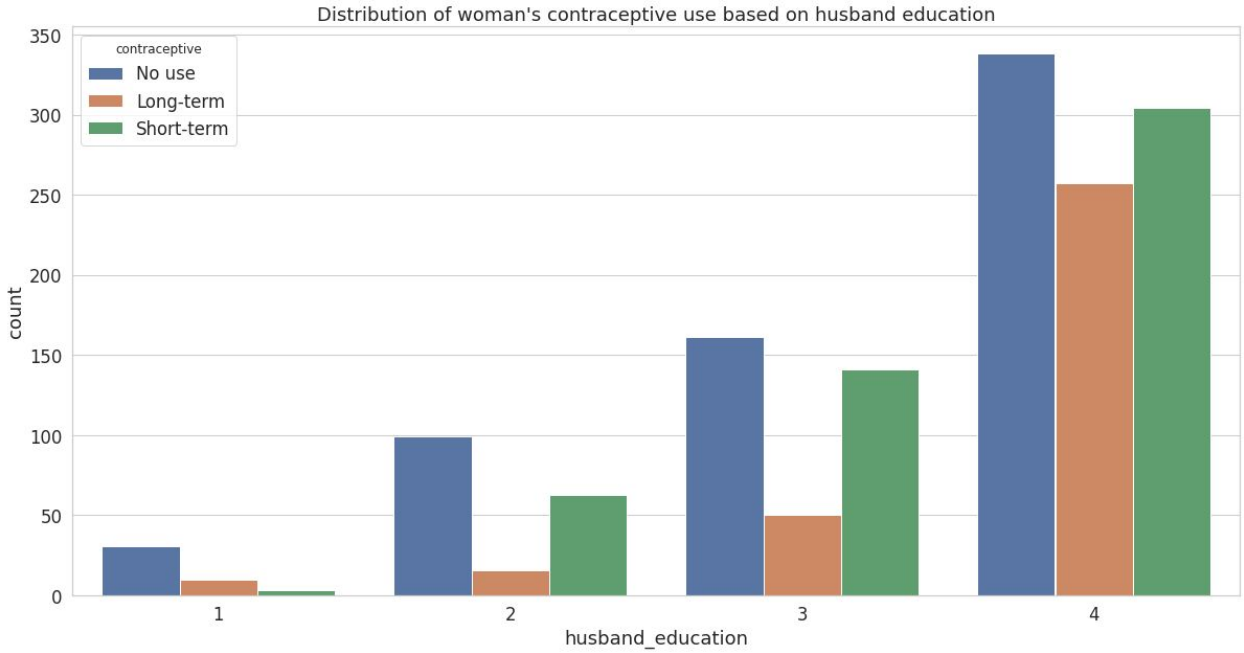
For our visualizations, we wanted to see if there were any features that we could use to separate the population into the different contraceptive method choice. Thus, we plotted pairs of features and colored each contraceptive method group differently.



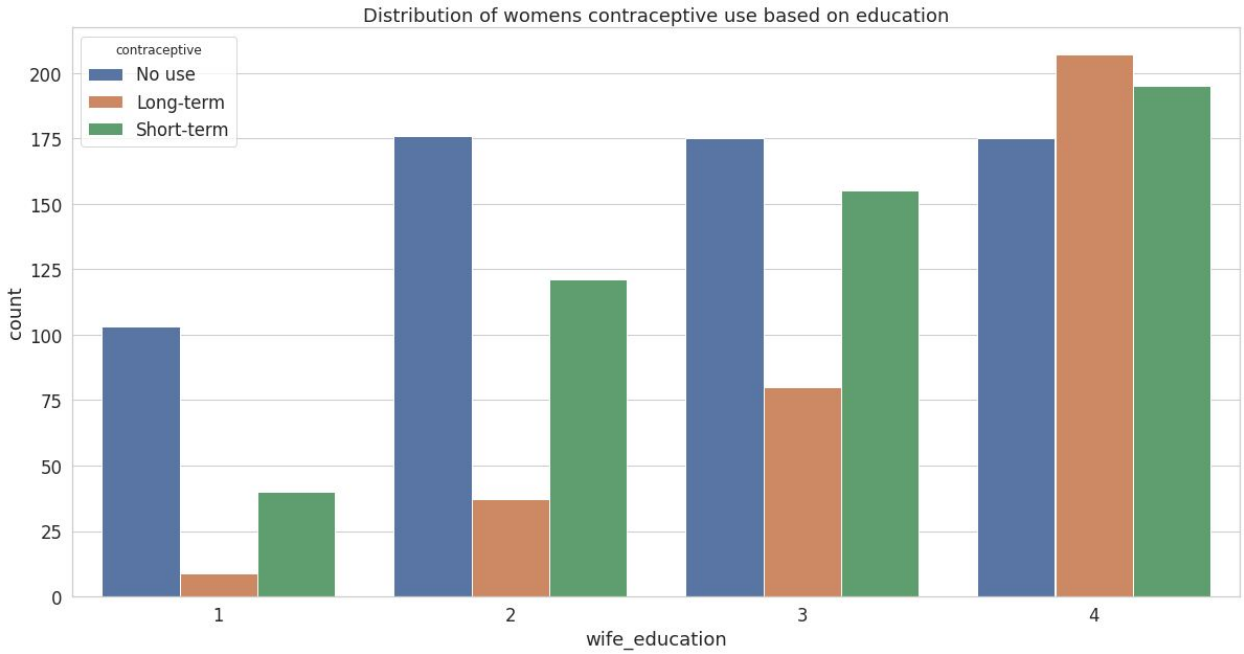
There seems to be a trend of older women having more children, but the relation does not seem very strong, nor does it change much between contraceptive use. There seems to be fewer children on average among people who use no contraceptives, but this is not a large difference. This is interesting because one would think that the number of children someone has would be related to if they use contraceptives or not. However, perhaps women start using contraceptives after they have had enough children, and women looking to have more children do not use contraceptives.



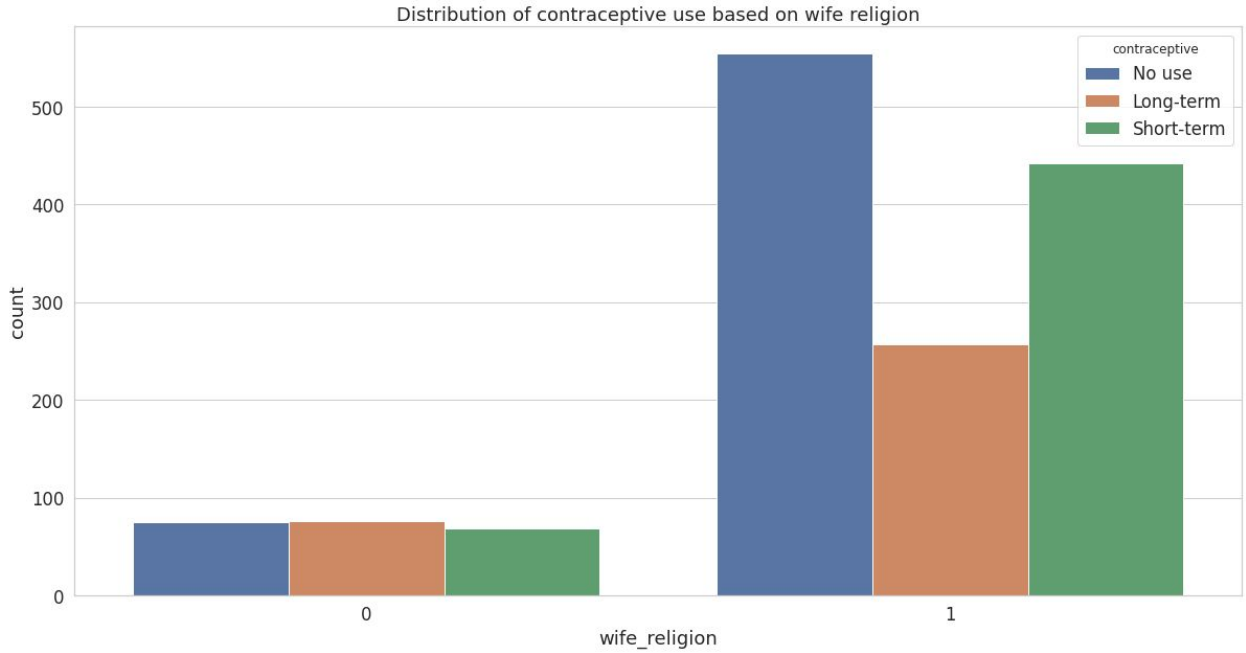
This visualization is interesting because it better shows how women who do not use contraceptives have fewer children on average than women who use other contraceptives method choice. It also shows that on average, women with lower education have slightly more children.



Then, we looked at the husband's education to see if that had any association with the contraceptive method of choice. There seems to be much fewer women who use long term contraceptives when the husband's education level was either 1, 2, or 3, which were the lower education levels.



This visualization above clearly shows that there is a correlation between wife education and contraceptive method choice, with more lower educated women using no contraceptives. This may be because education level is correlated with being more well off, and being able to afford more contraceptive methods.



This last visualization shows that there are many more women surveyed who are Islamic than who are not. Also, among the Islamic women, fewer women use long-term contraceptives than the other two methods.

Part IV: Methods and Experiments

We realized quickly that this was a classification question, as our response variable contraceptive method choice is a nominal categorical variable with values 1=No-use, 2=Long-term, 3=Short-term. With this, we decided to approach this classification question with several different combinations of models and hyperparameters, and then see which model would be the most useful for prediction.

We split our training and test data into a 75/25 split using the sklearn function train_test_split. In order to speed up our process, we wrote a pipeline function: evaluate features. This function took in a model, features, and a y column, and would calculate the training and validation scores of the model on the data with five fold cross validation.

We first used a logistic regression model using one vs. rest to predict the contraceptive type. We tried .1, 1, and 10 for the hyperparameter C, as well as tried to remove some variables from the features. Varying the hyperparameter C tells the model how to choose between correctly classifying each training point vs. having a large separation margin between the two groups at the cost of

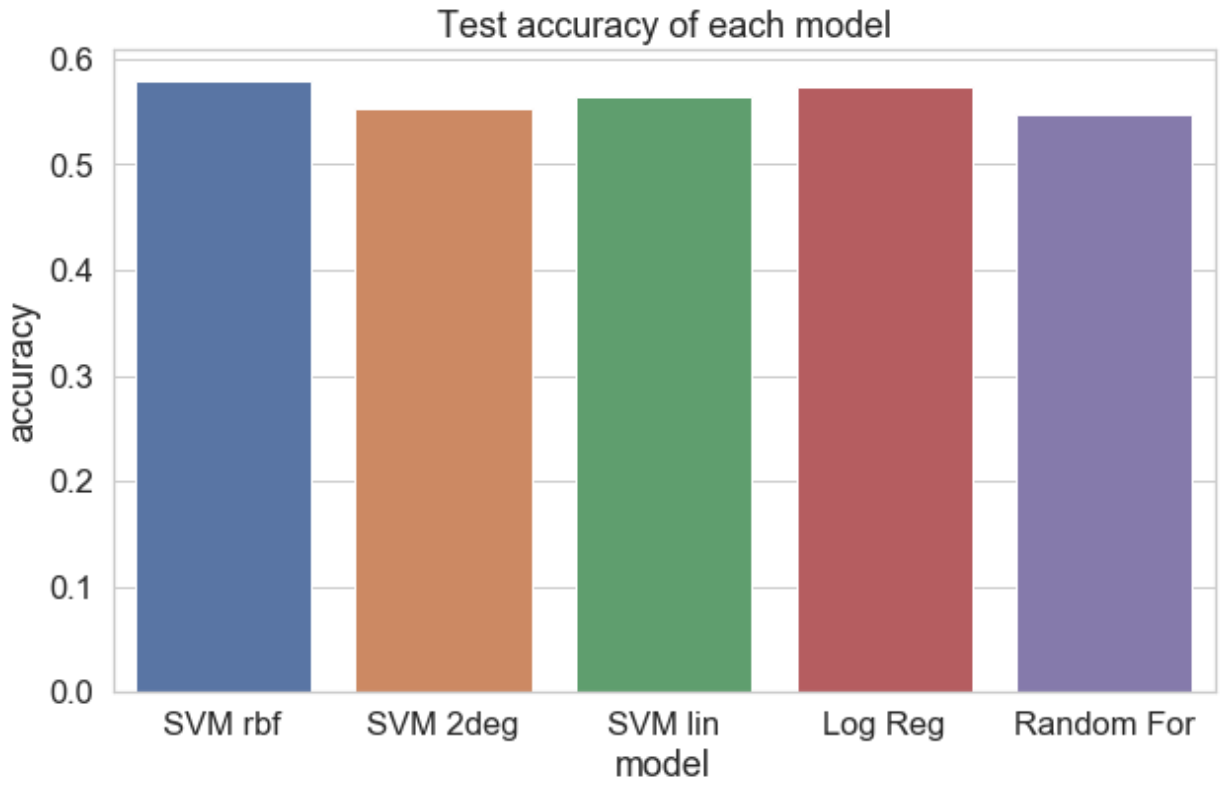
misclassifying some points. We chose logistic regression as our first model because it works well for classifying a point into a categorical group. We use one vs. rest to adapt the logistic model to a multi-class classification problem, by creating multiple models that predict one of the categories vs. the rest of the categories.

Afterwards, we can use the coefficients of the logistic regression model to get a rough idea of how important each variable is. Each coefficient gives a relation between a feature and the log odds of a data point belonging to a certain class. Because of this, we can hypothesize that coefficients with smaller absolute values are less important to predicting the class, and removing these features may allow our model to generalize to unseen data better. We ended up removing only two features, wife_work and media_exposure, which had lower absolute coefficient values.

We continue to try other classification models, this time using SVMs. A support vector machine is another model used in classification that tries to draw the best hyperplane that separates the data. It does this by identifying support vectors in the data, or data points that are the most similar to other classes. We try SVMs because from our visualizations and our logistic regression validation accuracy, we hypothesize that the data is not easily separable. Therefore, by using the kernels of an SVM, we may be able to project the data into another dimension that allows the points to be separated more easily. We try SVMs using a linear, 2 degree polynomial, and a radial basis function kernel. All of these SVMs achieve about 55% validation accuracy, so the kernels do not help classify the data much.

Finally, we also use a random forest model to try to predict contraceptive method choice as well. Because of the nature of random forests, we were able to get very high training accuracy, about 95%. However, due to this high number, we predicted that the random forest greatly overfit our training data.

Evaluating our logistic model and three SVMs on our test set, we 58.8% test accuracy for the SVM + rbf kernel, 57.2% test accuracy on the SVM + 2 degree polynomial kernel, 56.3% test accuracy on the SVM + linear kernel, 55% on random forest and 57.4% test accuracy on the logistic regression model. It seems like logistic regression and SVMs with non-linear kernels are not bad for classifying the data. The random forest model had the highest accuracy from our cross-validation studies but the lowest test value, showing that the random forest model did indeed overfit based on training data. All of our models performed about the same though, and the approximately .55% test accuracy means that none of the models performed exceptionally well on the data, and these models might not be useful for any real world applications because of the low accuracy.



Part V: Analysis and Conclusion

Interestingly enough, the women who used no contraceptives also had the fewest number of children, which was the opposite relationship we thought they'd have. We also thought that the wife's education level would be a good feature to predict contraceptive methods, and we were right. There was actually a clear relationship between higher education and higher long term contraceptive use.

One feature we thought would be useful was the number of children a woman had, but this was relatively similar across the three contraceptive methods, and did not seem to be useful in predicting contraceptive method choice.

This data did not have any NaN or missing values, so there was little data cleaning that actually had to be done. However, it was difficult to find what the categorical variables actually meant, because the variable meanings were not provided in the site where we got the data. Instead, we had to dig through the original survey, which was an extremely long pdf. We were not able to find meanings for all the variables, which hindered us from gaining a complete understanding of the dataset. For example, we could not find whether or not a 4 in occupation level meant, if occupation level was based on income bracket, or type of work. This would have been useful information to know, because we can hypothesize that a husband with a lower income level job would be associated with greater long-term contraceptive use due to the family not being able to support additional children. Another example of this ambiguity in the data came with media exposure. We were confused over what "good"

and “not good” media exposure meant, and since we did not have the relevant domain knowledge to apply it to our model, we largely kept this column untouched.

A limitation of our analysis is we never conducted specific analysis into which particular features should be engineered or correlated best with contraceptive method choice. We didn’t conduct principal component analysis to identify clusters of similar data. Additionally, our model is limited by the fact that we were not allowed to implement a neural network, which might result in a higher test accuracy. A last limitation of our model is with regards to us not understanding certain attributes of the dataset (as described previously). Having full domain knowledge about the dataset would allow us to use it to its full potential, by better understanding if certain features might be correlated and removing them.

Assumptions we made to this dataset included the following:

- That people accurately filled out the survey
- That the survey of people is representative of the target population of Indonesia
- That the way the survey was conducted was did not contain systemic biases

With regards to ethical dilemmas, one particular dilemma we see is that this study was conducted in 1987, which is over thirty years ago. Thus, if we were presenting a classification model to the government of Indonesia, our model would be based on outdated data and likely biased in some way since times have changed. One likely possibility that our data would have changed in 30 years is with the standard of living index. According to the World Bank, Indonesia’s poverty level has dropped dramatically every year since 1998. This could change our values in the standard of living index column drastically.

An ethical concern we have studying this data is surrounding the religion column. The only choices are Islam vs Non-Islam. While this breakdown of the data makes sense since Indonesia is ~87% Muslim, it also has a large percentage of Christians (~9%), whose religious views may also impact contraceptive method choice. Another ethical concern is surrounding whether or not survey results were collected in person. A married pregnant woman may feel uncomfortable discussing her contraceptives choice and give answers that she feels the surveyor wants if they are discussing her contraceptive choice face to face.

If we had a more recent dataset, we would be able to construct a model that better reflects the population of Indonesia than our data from over thirty years ago. Additionally, if we had more data points, such as family annual income and location of the family, perhaps we would be able to create a model that could better predict a woman’s contraceptive method choice. Data breaking down the family’s hopeful size would also be helpful. For example, a family that had no preference or wanted a large family size would likely have different contraceptive method choice than a family who wished for a small size.

Avenues for further research include better fine-tuning each model. For example, due to overfitting from our random forest model, we could try implementing pruning or a max depth. We could also implement regularization to better improve our test accuracy on our logistic regression models, or conduct PCA to better choose features. Lastly, we could attempt to implement more

complex models involving reinforcement learning in order to finally decide on the best models/features to predict contraceptive method choice.