Bee Kim
Razi Mahmood
Michael Aydin

# Data 100 Final Report: Contraceptive Dataset

**Abstract:**

There are various cultural, demographic, socioeconomic factors that contribute to whether one chooses to take contraceptives or not. This project attempts to explore and select the best of these factors to create a model that can accurately predict contraceptive methods used given a dataset consisting of nine different characteristics of a couple and their circumstances. We conducted a rigorous procedure including experimentation with different types of models, principal component analysis, and more that resulted in both successes and failures. In the end, we were able to produce a random forest model that can predict with a training accuracy of 0.95 and testing accuracy of 0.541.

## Question Framing:

The dataset we conducted our final project on comes from the 1987 National Indonesia Contraceptive Prevalence Survey in which married, non pregnant women were surveyed. Various characteristics of the husband and wife such as employment and religiosity along with contraceptive types were recorded. The question we aim to answer is which of these variables and/or other factors can best predict the method of contraceptive used. We hypothesize that age, education, number of children, religion, employment, standard of living, and media exposure will all be important features in helping predict contraceptive use. We believe this would be a relevant question to explore because it is valuable to see what specific factors (religious, demographic, economic, etc) hold the most weight in predicting what contraceptive method a woman uses (in 1987 Indonesia). Complications arise as these are many variables in the dataset to consider, and the best combination of them is not immediately obvious. Through the implementation of various analytic methods learnt throughout the course, we hope to find the optimal composition of features that creates the best predictive model to answer our question.

## Data Cleaning/Transformations:

(Note: While normally this process should occur before the EDA, we decided it would be best to perform it after the fact in order to make the visuals more interpretable to humans ie. labels of 0, 1, 2... vs. -0.5, 0.01, 0.23… children). We first wanted to address the categorical variables in the data set via one hot encoding which converts them into a binary representation. This was done so that categorical variables can be passed into models without different weights being placed on nominal integers, and the overall data would be easier to understand. We also standardized the numerical variables so that all the values are on the same scale by taking the difference between the mean and overall column then dividing it by min and max values of each numerical column. There were no null values present in any columns of the dataset before or after doing the train-test-split, so no data cleaning was necessary there.

The main forms of transformation that we did was one hot encoding the set and then grouping portions of the set into smaller data frames for us to perform our EDA on. These data frames

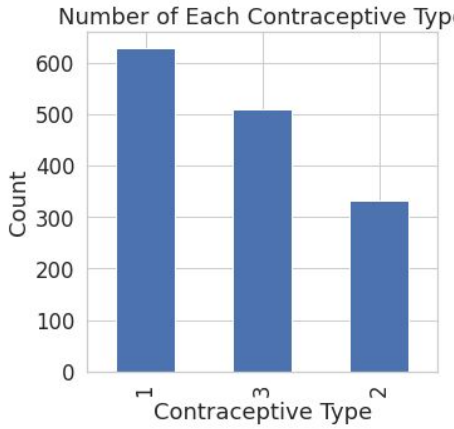
included grouping by the 3 contraceptive types to check for the correlations of husband and wife education levels, the number of children of the mothers, and the wife's age.

Additionally, after doing PCA on the original and encoded set, we selected features that appeared to contribute to the first two principal components in a meaningful way and refitted our models on these new columns within this transformation to check for improvements (more on this later).

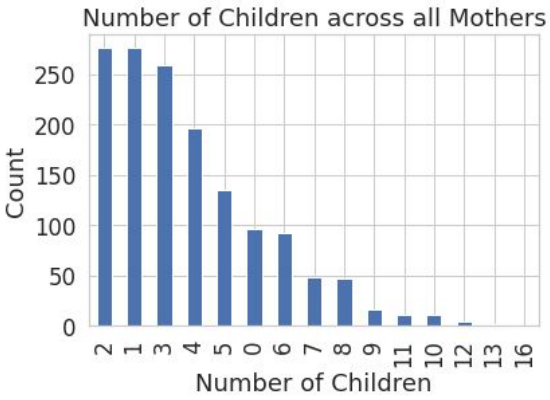
**Exploratory Data Analysis & Visualizations:**

First, we looked at the counts of contraceptive types and number of children in the dataset as these were potential variables our model would predict. We see that the dataset has mostly 'no contraceptive', then 'short term' and lastly 'long term' (See Image 1). The number of children is skewed towards a few children with some outliers of ten or more kids (See Image 2).

1

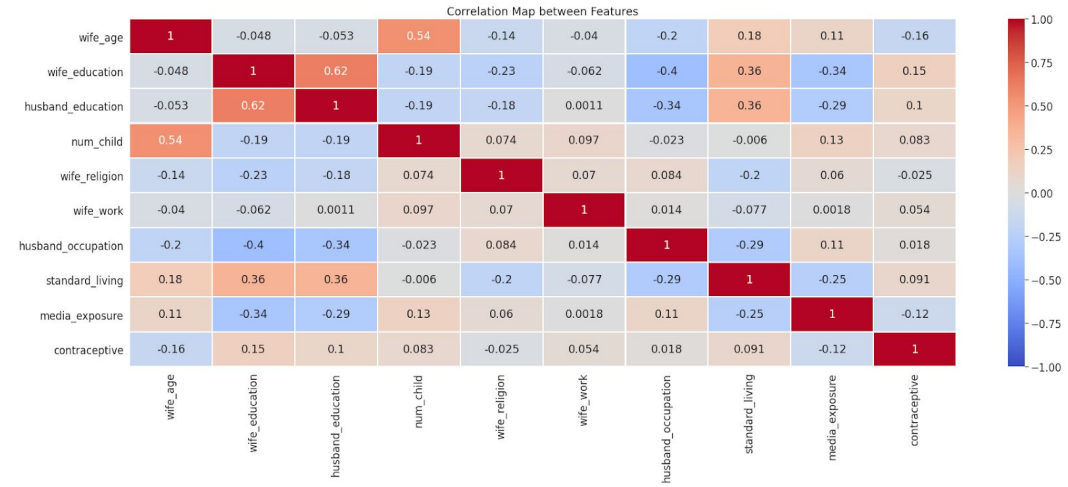



2




We also created a correlation heat map between all the columns of the dataset to get a sense of how some features could be connected, especially in relation to the contraceptive column. This initial visualization showed promise to standard of living, wife education, and husband education as features in our model (See Image 3).

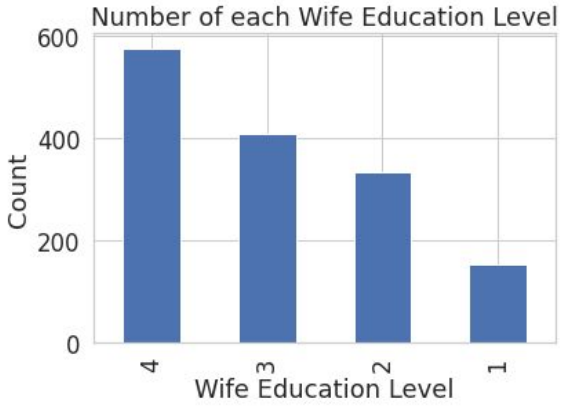
3



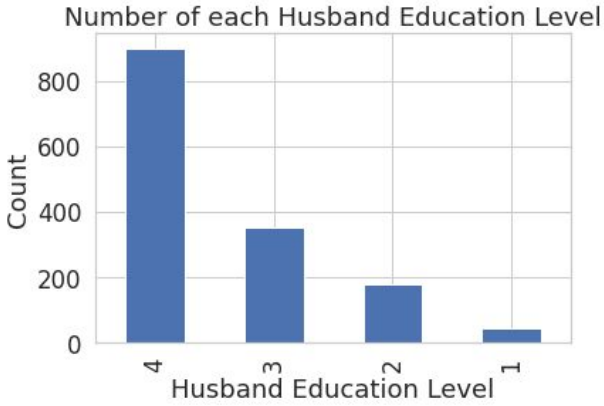

Next, we created some visualizations to investigate wife and husband education. We plotted the counts of each education level for both genders (Images 4 & 5). We found we are dealing with a generally more educated group of people, with far more in the highest education bracket than the

lowest. The men are more educated than the women on average. Then we used the one hot encoded data to look at the counts of education levels after grouping by contraceptive type used (See Image 6 and 7). For women, both short and long term contraceptive is dominated by the highest education bracket and slowly decreases as the education level decreases. The no contraceptive group is much more evenly distributed between education levels of 2, 3, and 4. It should be kept in mind that there are overall way more women with higher education so this is not too surprising. There is a similar story with the men except the highest education bracket dominates all contraceptive groups more heavily. Again this is due to the high number of educated men and lack of uneducated men.

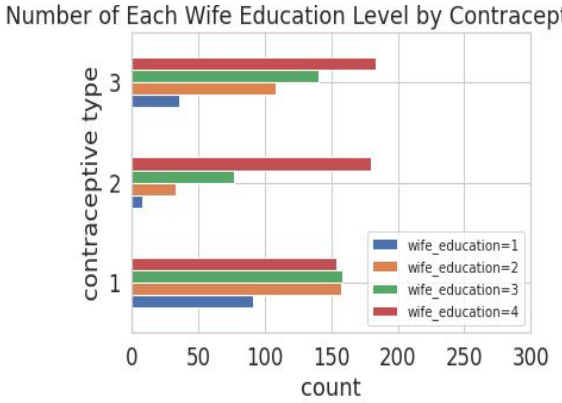
4



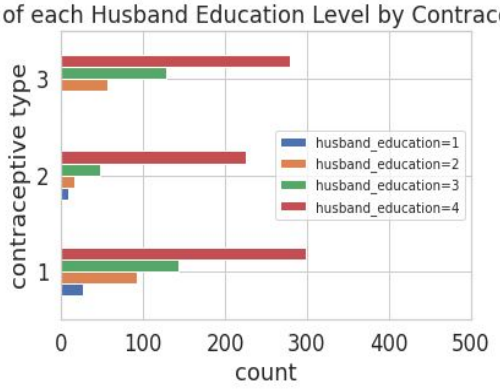

5




6

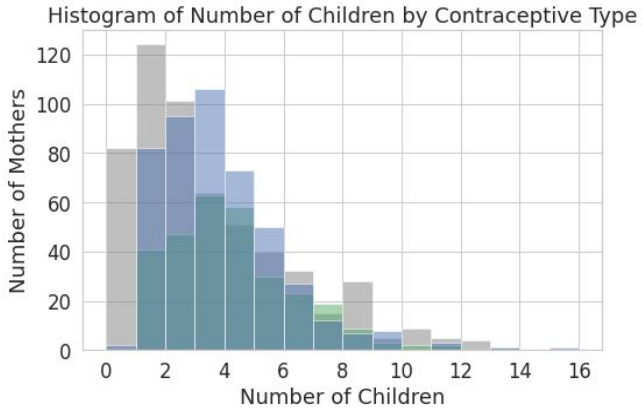



7

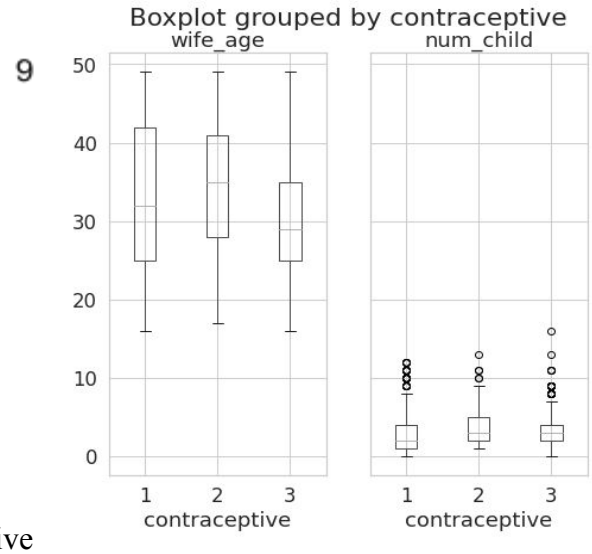


Next, we looked at the relationship between the number of children and contraceptive type using a histogram (See Image 8). The histogram does not skew towards fewer children for any of the contraceptive methods. To our surprise, long and short term contraceptive types skew towards more children.

8

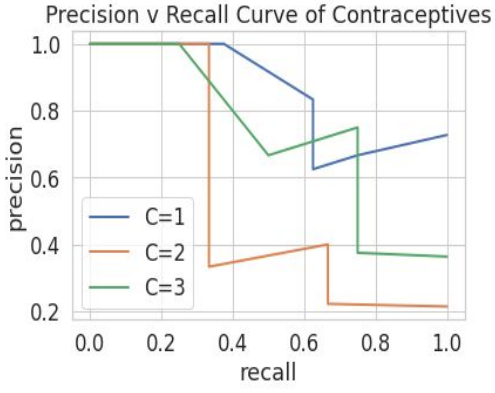



9




(Image 9) shows two boxplots grouped on the 3 contraceptive types that assess the distributions of wife_age and num_child. The first boxplot regarding wife_age shows that for each method of contraceptive, the range of ages is fairly equal- varying from around 16 to 50 years. The stretch of 50% of the data is longest for women taking no contraceptive and shortest for those taking long-term. Additionally, long-term contraceptive use seems to lean younger than the other two groups. The second boxplot regarding num_child shows some interesting trends as well. Group 3 (long-term) has the shortest range in regards to its quartiles but has the farthest stretch in regards to outliers. This is partially expected since we hypothesized that women using long-term contraceptive methods would have less children, which is generally true when disregarding the outliers. However, it's interesting that this same group has the farthest outreach when taking into account outliers. Regardless, all three groups appear to skew to fewer children, with Group 2 (short-term) tending to have more children relative to the other two. This is also another surprising development since our initial thought would be the group taking no contraceptive would have the most children. Hopefully, in the upcoming modelling, we will see how these contradictions in the num_child distribution comes into play in our feature selection.

10



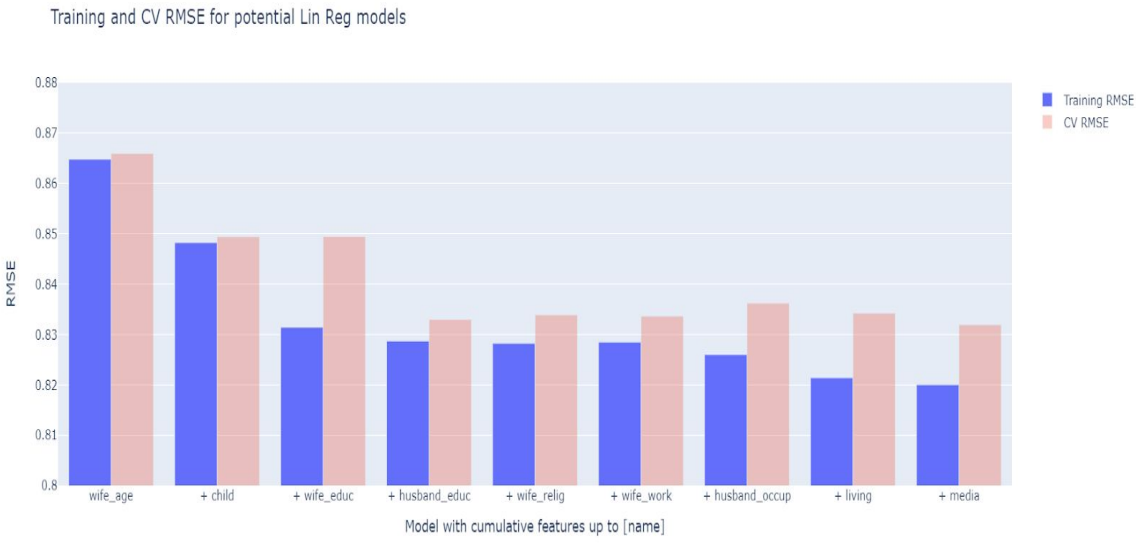

Also, when looking at the overall precision-recall curve of the testing data (in Image 10) to check the precision-recall with the appropriate hyperparameters of our final model (which was the random forest), we can see that precision for the contraceptive 1 starts at 1.0 but then fluctuates down between 0.8 and 0.6 as the recall

increases, which can imply that there are more false positives when trying to accurately predict the women who have taken contraceptive 1. For predicting contraceptive 2, there's a massive drop in precision from 1.0 to 0.4 after the recall goes roughly above 0.3, which continues to fluctuate between 0.4 and 0.2. This implies that there are even more false predictions of women taking contraceptive 2, due to potential overfitting on our prediction model or improper classification of other features that are more commonly associated with women taking contraceptive 2. For contraceptive 3, there seems to be a similar case as with predicting for contraceptive 1 since the precision steadily drops from 1.0 to around 0.7 and later fluctuates from there and 0.8 before dropping to 0.4 as the recall increases. There aren't as many false positives as with contraceptive 2 since we end up with a slightly higher 0.4 precision value, but the similar pattern of the model misidentifying the contraceptive is still present. The curve, in general, seems to tell us that the performance of the binary classification of the contraceptives does not entirely match what is originally given to us, which correlates to the overall middling accuracies and cross-validation scores we got in our other models.

**Method/Experiments:**

- Linear Regression:

We started off with linear regression models as they are simple to create and less computationally demanding, allowing us to create many different models using an increasing number of features to see what the effect of adding more features is on training/CV RMSE.




(Image 11) We found that as the model accumulates more features, the training RMSE decreases. The CV RMSE generally follows the same course, except it increases when we add wife_education, wife_religion, and husband_occupation albeit by a very small amount. It is still something to keep in mind as we proceed with our modelling. Doing this linear regression as a starter provided good visualization as to the general trend of decreasing RMSE as we add more features. We found there was no dramatic jump in training or CV RMSE that

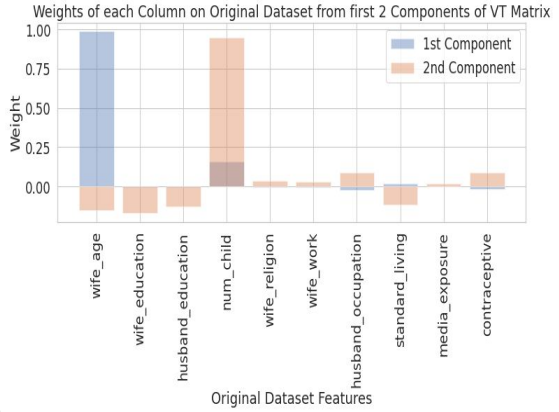
greatly signifies overfitting and warrant regularization. Regardless, we want to move on to performing more intricate modelling and feature selecting for our dataset.

- Logistic Regression, LogisticRegressionCV Models: (LR & LRCV)

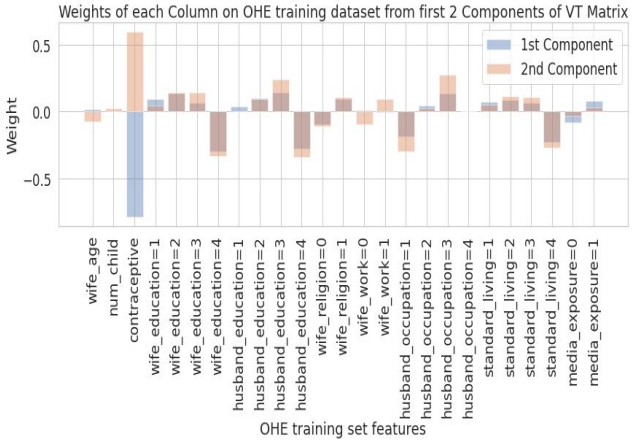
Next, we worked with a LogisticRegression (LR) and LogisticRegressionCV (LRCV) model. We felt this would make sense since we were dealing with a multiclass set of categorical columns and we wanted to handle the binary classification of these variables that were one-hot-encoded. After fitting the models on the encoded categoricals and the contraceptive column, we tried getting the training accuracy of both models on a 5 K-fold approximation with roughly similar values of 0.515 and 0.526. Later, we tried doing hyperparameter tuning by experimenting with different values of 'c' and 'cs' values ranging from 2 to 20 in a for-loop to see which value would yield a higher accuracy and cross-validation scores. From this experiment, we found the highest training accuracies of both models to be 0.515 and 0.516 and the highest cross-validation value of 0.55. The lower accuracy for the LRCV score could be attributed to how a higher Cs value of 7 reduces the strength of regularization compared to a value of 4 or 5.

We attempted to implement LASSO regularization to improve these LR(CV) models by selecting the features that had non-zero parameters. In both the LR and LRCV models, the training accuracies decreased to 0.518. The testing accuracy increased from 0.534 to 0.54 for the LR model and remained the same for the LRCV model. Ultimately, we concluded that this improvement was not very significant, and we should search for better methods of feature selection.

We found a similar result when attempting a PCA approach of selecting columns that appeared to have bigger weights of contribution by assessing the barplots of the first 2 components of our $V^T$ matrix following SVD decomposition of the original dataset and the one-hot-encoded training set (see Images 12 & 13). After fitting a new LR and LRCV model on these selections to check for improvements, we saw that this approach didn't seem to improve the scores as the training accuracies were both stuck down at around 0.51.




12




13

- Decision Tree:

We tried using a decision tree model to see if this would be a good choice of model. We experimented with some hyperparameters such as max depth of the tree. In order to get a clean graphic for the tree while still getting good training accuracy we settled on a depth of 6 after some experimentation. Selecting higher max depths could result in much higher training accuracy but it overfits the training data and test accuracy did not improve or suffer. We tested various values for max depth by doing a for loop from 1 to 30 and compared training accuracy. The decision tree ended with a test accuracy of 0.588.

- Random Forest:

Naturally, next came random forests. We experimented with different max depths but this was redundant as leaving this parameter blank let the model choose optimal values since each tree in a random forest over fits in a different way. We also experimented with a number of estimators, and found that choosing values that were too high for this caused run time issues. We then ran a for-loop over values of n_estimators from 1 to 30 and settled on 22 because of the best accuracy. Additionally, we applied the same PCA strategy of fitting our model on the features that appeared to have larger weights, and we got our final training accuracy to be 0.95.

**Analysis and Conclusion:**

The final model we're going with is the random forest with PCA features. There were pros and cons to the various models but we ultimately went with this one, mostly since it has a very high training accuracy of 0.95. Although this model appears overfitted to the training data, the nature of random forests is that they make many trees that overfit in different directions so we thought this would end up with a good test accuracy. The test accuracy of our final model was 0.541. We thought this was respectable as it beat the training accuracy of some of our other models.

**(i)** Some features that we found interesting were the different wife/husband education levels. We found based on our bar plots comparing education of the wife and husband to the contraceptive type, along with the PCA plots, that the number of husbands with an education level of 4 have higher counts across each contraceptive type. Also, wives with education levels from 2-4 looked

uniformly distributed for the 1st contraceptive type, although there were more wives for the other contraceptives that had an education level of 4.

**(ii)** Some features that turned out ineffective were the wife_work and media_exposure levels. From the EDA, we saw in the pivot table that the distribution of contraceptive methods used was fairly equal for both working and nonworking women which contrasted with our prior thought that the employment status might indicate one method over another. This was further seen in the PCA weights, where the wife_work levels seemed to have equivalent weights (although in opposite directions). Naturally, the media exposure levels were ineffective from a logical standpoint and did not have rather large weights to help our model predictions, as discovered by PCA.

**(iii)** One of the main challenges we dealt with early on was the major dominance of categorical variables, and how we only had 2 numerical variables to check for correlations. Typically, it would be easier to check for patterns using histograms and barplots on numerical features. But we mostly had to group our main dataset into smaller data frames to identify patterns between different categories like wife and husband education levels against the contraceptive type.
Also, the fine-tuning of our hyperparameters for the logistic regression models and decision trees was tedious, since it typically affected the runtime of our file and would occasionally error despite carefully creating new models to test out the values. We had experimented with different values that would hopefully improve our training accuracy and CV scores, but the scores across most of the models we tried usually varied between 0.51 and 0.54, confusing us on how much we could significantly improve the behavior of our models with the given data. We also had to be careful with overfitting when choosing our models, such as with the decision tree since we initially had higher training and testing accuracies of 0.58 and 0.6 than the LR(CV) models, but we knew that decision trees tend to overfit on most of the features. Even when selecting the features from the PCA, we had to make sure that we selected ones that had generally higher weights, with some features seeming more important than others in the same category. This can be seen with the wife_education level of 3 and 4 having higher weights than 1 and 2, the husband_occupation levels of 1 and 3, and standard_living value of 4 having a higher weight than their other counterparts within their respective categories.

**(iv)** One of the limitations that we dealt with was the limited number of columns originally given to us in the prime Excel dataset, and the lack of numerical variables aside from just the wife age and number of children. We simply assumed that different education levels were ranked from less to more educated individuals, and that standard living and husband occupation levels were from worst to best. The assumption about the husband occupation or wife education levels could prove to be incorrect as a level 4 education or job may actually dictate a less reputable amount of credentials for the husband or wife instead of a level 1 rank, which may change our

understanding of the patterns between these categories and the contraceptive type. We also had to assume that the higher weights on the PCA bar plots would correspond to larger significance when fitting our models and hopefully improve our performances, which was not always the case as seen with the LR and LRCV models.

**(v)** An ethical concern is an inherent bias due to how this data was collected. The dataset sampled only married women. By not considering unmarried women, there could be a bias in the dataset. For instance, single mothers are not considered which could add more data towards a lack of contraceptive use, along with other implications of being a single mother which could affect other columns of the data. We also take into consideration that this predictive model only works specifically for Indonesia in 1987 and is not indicative of any trends internationally or in the present day. In general, an ethical concern when studying the problem of contraceptive type used is privacy issues and general taboo of the subject due to societal, cultural or religious pressures on the subject matter.

**(vi)** An interesting dataset that could strengthen our predictive look at contraceptive use would be data on childhood education these women received on contraceptive use. For instance, a dataset that recorded middle or high school health classes, sex ed, etc would be useful. Another interesting point of further research would be to look at different types of contraceptives used. Short term and long term are rather vague and each type of contraceptive has varying effectiveness. This combined with data on childhood and adolescent education on contraceptives would be something intriguing to research. Another potential feature we could account for are various cultural factors. We could do this more locally by studying the location of the women surveyed (which may correlate with standard_living), or internationally by performing a similar analysis on a survey that was taken in another country outside of Indonesia.

**(vii)** An ethical concern that may be encountered is inherent biases of the person doing the data analysis. For example, we acknowledge that we had some leaning prior to beginning this project (ie. described in the introduction how a higher educated or employed woman would be more likely to use contraceptives). This may subconsciously affect the analysis one does in which they attempt to find conclusions that confirm their underlying beliefs. Additionally, the data can be manipulated to push a certain agenda. One should make sure to thoroughly look through the methods employed and keep a cynical viewpoint when looking through a report that may seem objective on the surface.

Bee Kim
Razi Mahmood
Michael Aydin

*(The following 2 visuals were reported as blank on Gradescope, so we've attached them here)*