DS100 Final Project Report

Contraceptive Data Set

Authors: Callie Nguyen, Cynthia Lai, Xuanyi Zhang

Spring 2020

I. Abstract

The Contraceptive Method Choice Dataset compiles data collected from the 1987 National Indonesia Contraceptive Prevalence Survey. Information on the 1473 women include their choice of contraception of either no usage, short-term methods or long-term alternatives, as well as socio-economic and demographic attributes such as education level, working status and standard of living. This report investigates data modelling methods to predict a woman's contraceptive usage based on socio-economic attributes.

II. Question of interest

Given these attributes, our data exploration is focused on the relationship between socio-economic conditions and contraceptive methods. Specifically, the study was interested in the impact of a woman's education on her choice of family planning devices, or, in other terms, **how accurately we could predict a woman's choice of contraceptive between no-use, long-term and short-term, based on her education level**. Indeed, our study found that many other attributes such as the husband's education are also key determinants of contraceptive method. However, we found a positive correlation between these other determinants and the wife's education and as such, wanted to delve into the direct and indirect capacity of a woman's education level to determine contraceptive methods.

III. *Data Cleaning / Feature Engineering*

At first glance, the dataset is fairly straightforward without any missing values.

When we plotted out the distribution of number of children for women who work vs women who do not work, an interesting data point was the instance of 16 children. Amongst all women who had a high degree of education, the next highest number of children instance had 13 children. Even when compared against a broader group of instances with other similar attributes (high aggregated education, high standard living and short-term contraceptive use), 16 children was still considered an outlier as shown on the boxplots, so the instance with 16 children was dropped from the dataset used in the models.

In our EDA, it was also clear from the graphs that the use of numbers as nominal variables were misleading, suggesting an inaccurate ordinal relationship between different levels of an attribute. Some attributes are not labeled conventionally (ie 0 = Yes and 1 = No), which we briefly changed back for the ease of interpreting. Although some attributes such as wife's education or standard of living are labelled in order (i.e. 1-4 representing lowest to highest), the value of the numbers themselves in relation to each other are not representative of the relationship between different levels. In order to ensure that this nominal relationship is not mis-interpreted in our models, we hot-encoded our categorical variables. This binary format will allow the model to equally weigh each of the features.

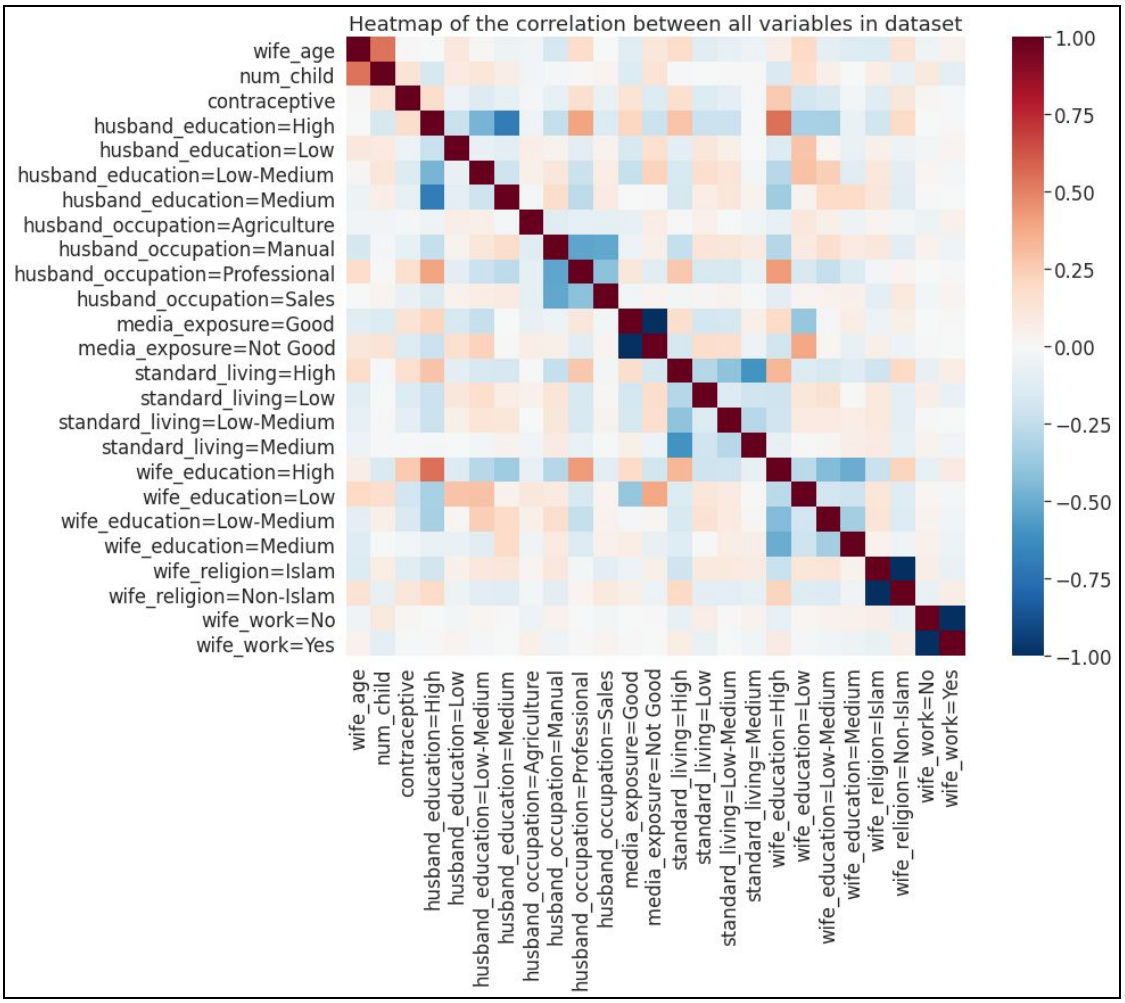
On the same note, since the numerical variables are on different scales, we also normalize them in order for the models to weigh them equally. However, we normalized the quantitative attributes after EDA for simple interpretation, for instance, a family with 3 children is more understandable than that with -0.11 standard units.

Similarly, we found during EDA that it was difficult to ascertain what numerical labels meant for qualitative categorical features such as husband occupation. With further research into the context of the study, we replaced the labels of husband occupation with the appropriate categories from the survey

(agriculture, manual, sales and professional). Ideally, a more quantitative label, such as average income for these industries, would have supported the model. Nonetheless, changing these labels assisted with interpretation especially while conducting EDA.

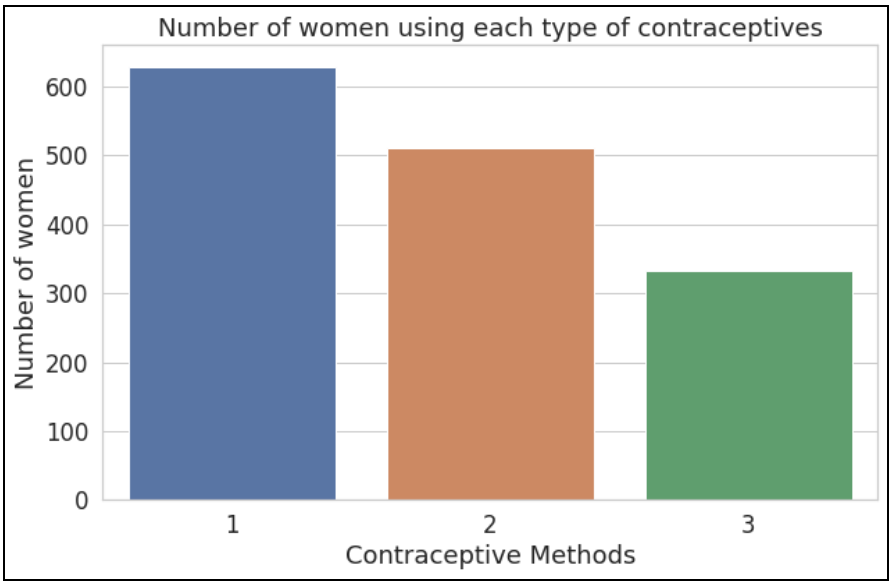
IV. EDA

Firstly, we wanted to determine which attributes were most correlated with the **wife's education** and **contraceptive choice**. We created a **heat map**, using both the depth of color and a numerical value of correlation to determine that husband occupation, husband education and media exposure were most highly correlated with wife's education, suggesting some correlative relationship.

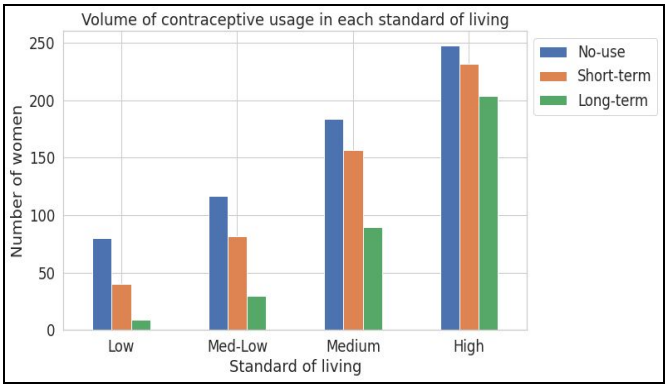


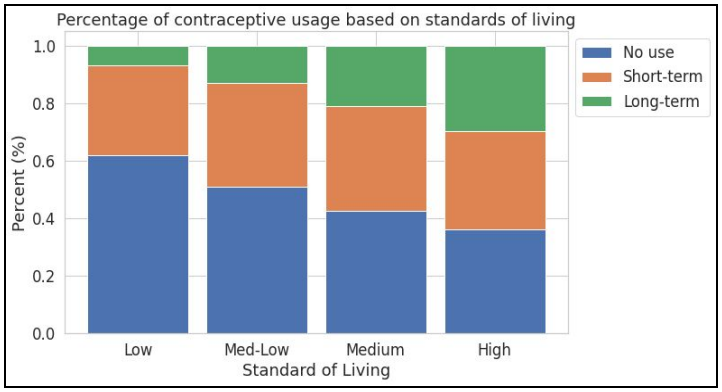
Contraceptive was also correlated with a wife's education and standard of living. To verify, we also used **seaborn's pairplot on the original dataset**. While it could not provide specific information, by creating some jitter to separate the scattered points, we were able to affirm our heat map observations and observe a few general trends noted in the notebook.

To gain a sense of our data, we initially calculated the numbers of instances using each type of birth control. As expected, the most popular form was **non-usage** with 629 instances, followed by 511 women using **short-term** options while 333 women used **long-term** contraceptives.

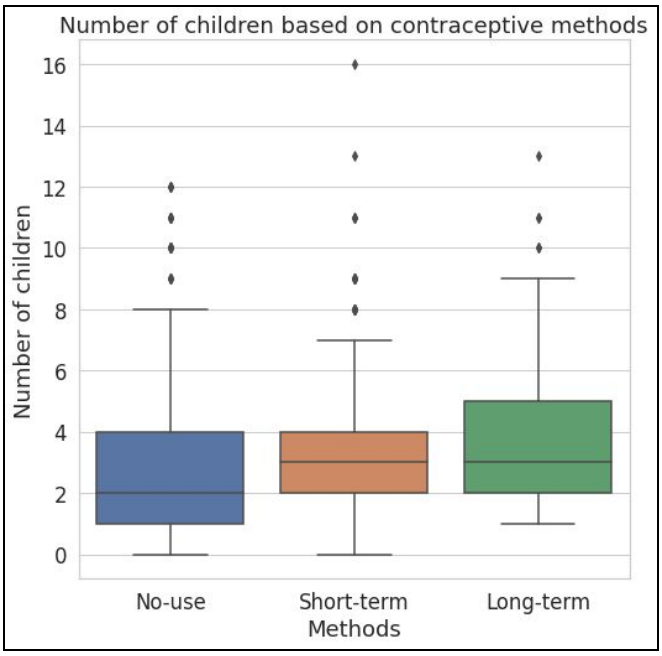


Our side-by-side bar charts visualize this trend in each of the **standard of living levels**. Since this would partly be reflected by the larger overall proportion of women in this standard of living level, we also include the proportional distribution of **contraceptive choice based on standard of living**. In following our assumption that cost might be a factor, we used standard of living as an indicator of socio-economic status. Since long-term contraceptive options such as IUDs tend to be more expensive, our hypothesis was that women with higher socio-economic statuses would prefer long-term contraceptives. Our pivot table analysis suggested that indeed, more than 60% of these women using long-term methods had a standard of living at the highest level.

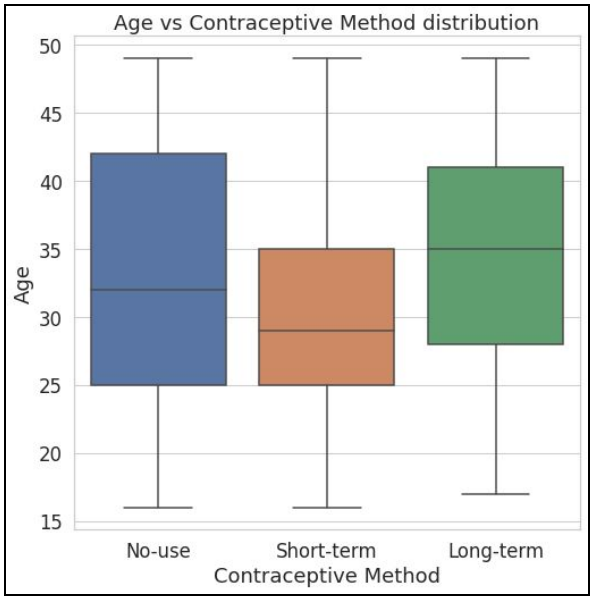




Initially, we also assumed that the **contraceptive method** would have an impact on the **number of children**. Women who either were not using contraceptives were more likely to want more and have more children. However, surprisingly, we found that women only short term and long-term contraceptives on average have more children.



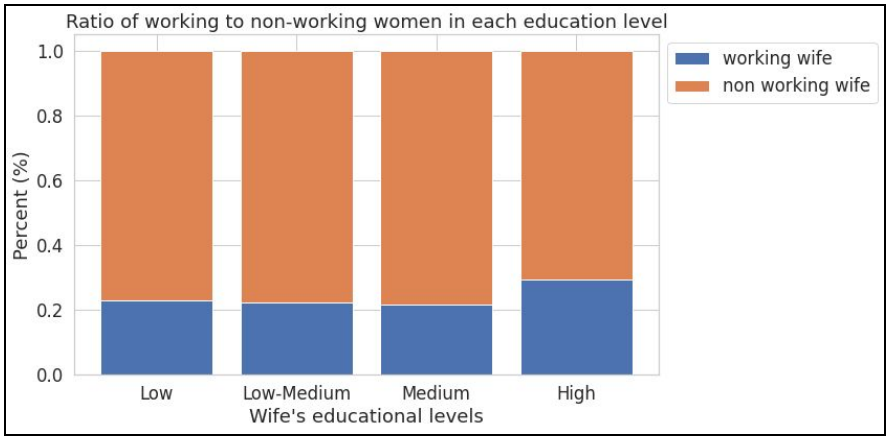
We also recognized that the **age of a woman** may impact **contraceptives** used and wanted to quantify this confounding factor. Indeed our boxplot distributions showed that long term contraceptive users tend to be older whereas short term contraceptive users are younger.

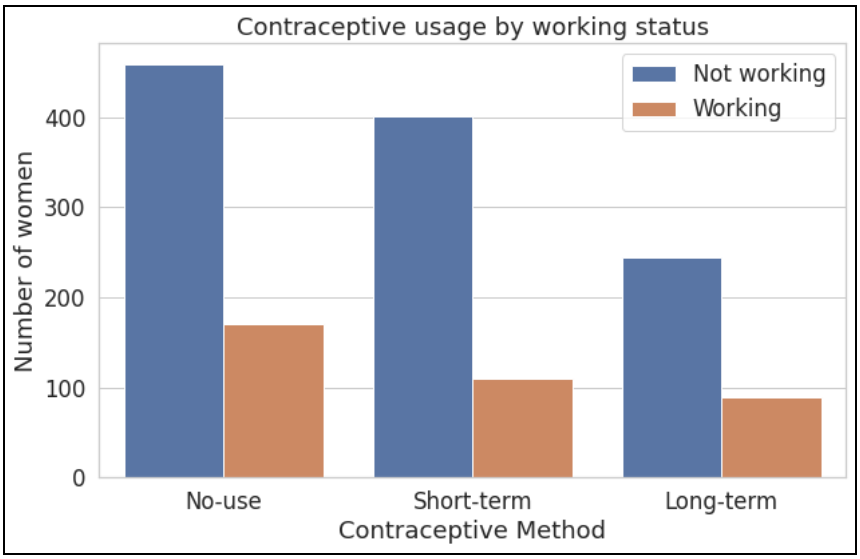


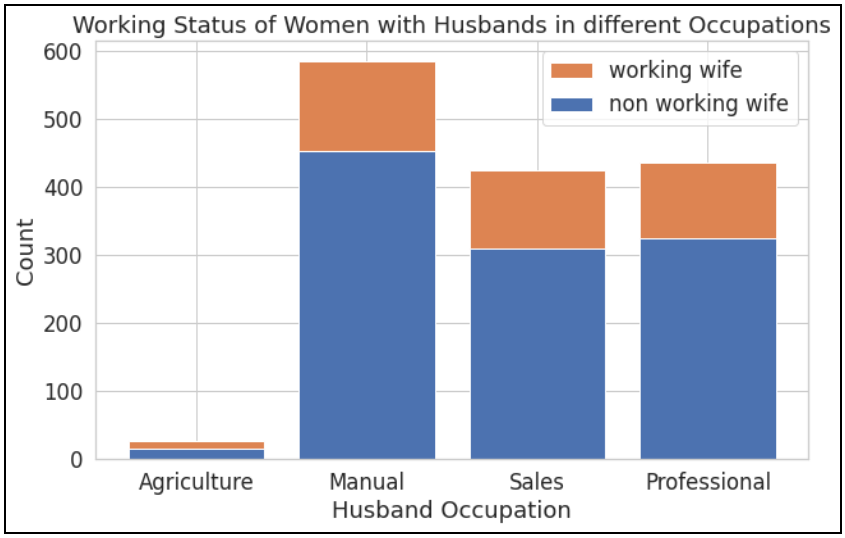
Now that there was a clear relationship between socioeconomic status and contraceptive method, we wanted to further delve into whether there was a relationship between the **wife's education** and **socioeconomic status** that would enhance our prediction of contraceptive.

Firstly, we wanted to quantify the effect of the **wife's working status**. Two working adults in a household are more likely to have a higher level of income, thus affecting contraceptive usage. Our EDA found that a higher proportion of women using long-term contraceptives tend to be working, and similarly, highly educated women are more likely to be working. This could suggest that highly educated

women are more likely to work, and thus would opt for long-term contraceptives that would not distract them from work.

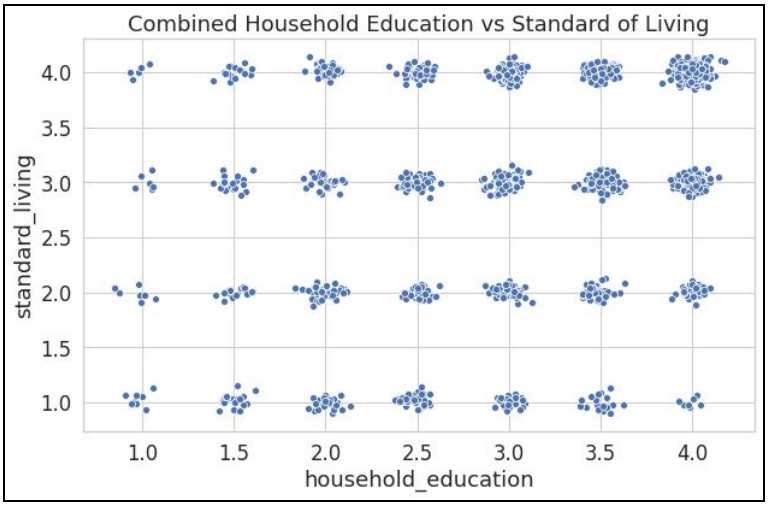






After dividing husband occupation into agriculture, manual, sales and professional, we compared the ratio of **working to non-working women with husbands in each of these occupations** as well as a wife's education level distribution in these occupations. It is assumed that men in agricultural occupations are less likely to have a high salary and household standard of living whereas men in professional occupations are more likely to have a high salary and high standard of living. Our graphs indicated that while there wasn't a clear distinction of working/non-working ratios in different occupations (the higher agriculture ratio may be skewed due to its small sample size), wives with higher education levels make up a larger proportion of husband's occupations with higher assumed salaries (professional and sales). Conversely, wives with lower education levels tend to have husbands working in agriculture or manual occupations.

We plotted and found similar results comparing husband and wife education levels so we decided to aggregate their **education levels** with a mean to compare against **standard of living** in a scatter plot. As expected, households with higher aggregate education are more likely to result in a high standard of living, which would affect contraceptive methods of choice.



Finally, to reduce the dimensionality of our data and gauge which features may be strong indicators of contraceptive use, we conducted **Principal Component Analysis** on our original non-hot encoded data set for clarity purposes. It appeared that the first three principal components generated a majority of the variance. After plotting out the first three sets of feature weightings, we discovered that the features which will most likely provide insights into contraceptive methods in our models are **wife age, wife education, husband education and husband occupation.**

V. Modelling

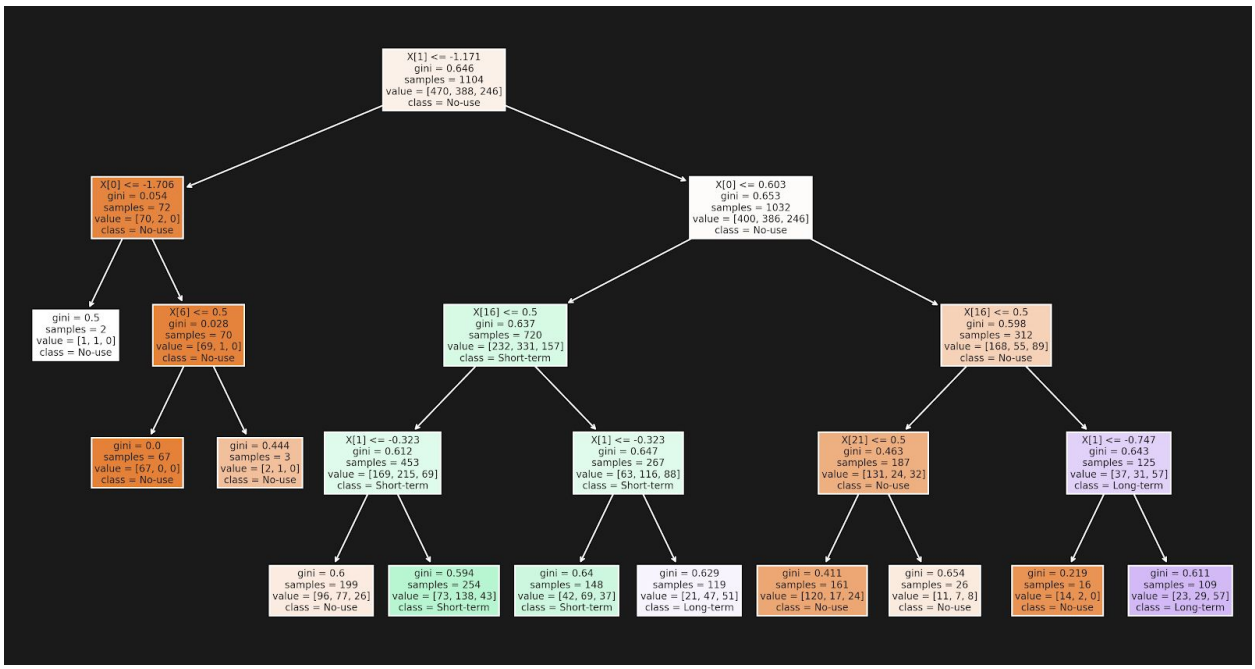
Since our response variable is categorical, we find it appropriate to use **Logistic Regression**, **Decision Tree** and **Random Forest** as classifiers for our models.

Using Logistic Regression to predict contraception using only numerical features, our base model yields a mediocre CV 10-Fold accuracy score of 47%, after which we realized that there were 3 classes of birth control methods (no-use / short-term / long-term) to classify; this is a **multiclass classification** problem. We are better off using the ***one-versus-rest*** approach with **Multiclass Logistic Regression** where the multiclass prediction problem is divided into separate binary prediction problems. One thing to note is that, while the training accuracy and the CV accuracy varies for different iterations of our logistic regression models, they all appear to be close to each other. So conclusion can be made that the logistic regression models do good jobs on not underfitting/overfitting our data. On top of that, we also have a list of highly correlated or heavily weighted features to combine with MLR in an attempt to better our base model. As a result, we found that the model using **MLR on all features** yields the best result among all LR models. The **model improvement score** is **3%**.

With **Decision Tree**, the base model to predict contraception based on all given features introduces 2 alarming problems: One is, as expected, fairly low 46% accuracy score, and two is a high training accuracy of 97% - which is an indicator of overfitting. From what we know about decision trees, it is likely that the overfitting problem comes from the fact that the tree grows too deep with too many branches in an attempt to purify the attributes. The simplest solution to this is to stop the tree from growing by setting the tree's maximum depth. Since we are using the GridSearchCV library to pick the best max_depth, we also tested on other features such as criterion and min_samples_leaf. As expected, changing the ***depth of the tree*** not only helped us improve the CV accuracy score of the models, but also took care of the overfitting problem that we have in the base model. 2 **hyperparameter tuned models** both yield better results, with the first one having only slightly better score of 55.97% since we also tuned the min_samples_leaf parameter. The **model improvement score** is **9.9%.**

Similar to the challenge we faced with Decision Tree, **Random Forest Classifier** also yields a low CV 10-Fold prediction score of 50% with a high training score of 94%. This is expected because Random Forest is built from multiple Decision Trees. Our approach stayed the same where we used GridSearchCV to optimize the hyperparameters and train our models. The 2 parameters of interest are ***depth of trees*** - max_depth and the ***number of trees in the forest*** - n_estimators to tackle overfitting without sacrificing the speed of the model. As such, fine-tuning these 2 parameters resulted in better Random Forest models and higher CV scores, and while other parameters also helped improving scores, the change was not noticeable. However, the training accuracy is still a little high, indicating that we might have slightly overfit the models. The **model improvement score** is **5%**.

From this, we have acquired 10 different models to test and compare with metrics of success being accuracy %. After creating a table compiling all their training accuracy scores and CV 10-fold scores, **Hyperparameter Tuned Tree 1** seemed to generalize the best among all models. We decided to move forward with it as our final model because it has proven to give a good CV score without the cost of overfitting.



This model resulted in the test accuracy score of **56.25%**. It means among 1473 women, based on the information we have, we expect to correctly classify 825 contraceptive methods that they choose. This result is not quite as high as we want our predictions to be, but by far is still the highest score we have obtained.

VI. Discussion

i. What were two or three of the most interesting features you came across for your particular question?

An interesting relationship we observed was the strong relationship between husband and wife's education, specifically if they are both highly educated. According to the heatmap and EDA, their correlation is significantly higher than any other relationships relating to education levels. At the same time, women with higher education also have the highest correlation with the type of contraceptive they

choose. This might lead to a more meaningful connection between the household's education levels and the choice of family planning devices.

Another interesting heatmap observation was that a high level of women's education was more likely to generate high correlations with factors such as husband education, husband occupation, standard of living, working status and media exposure. While there was a general matching trend where men and women have similar levels of education and occupation for all women, other women education types do not have such high correlations, suggesting that highly educated women have a greater impact on household decisions, including contraception methods.

ii. Describe one feature you thought would be useful, but turned out to be ineffective.

We initially believed that the number of children in each family would be a useful feature with significant weighting in determining contraceptive use. Our hypothesis was that women who want children are more likely to have a higher number of children and would not be using contraceptives. Similarly, women who do not wish to have more children or many children would be on some form of contraceptive. Long-term contraceptives are usually used for women who do not wish to have children in the long-term so we assumed that this was likely used by women who either did not want to have children at all or only 1-2 children. However, our EDA found the converse result. On average, non-users had 2 children with 50% of the group having 1-4 children whereas 50% of long-term users had between 2-5 children with an average of 3 children. 50% of short-term birth control users generally had 2-4 children with an average of 3.

This inconsistency with our assumptions is likely due to the survey being a snapshot in time. It was a better reflection of women who either, had not had their desired amount of children, thus not using contraception and have a lower number of children, or women who, already had the desired amount of children and were thus on long-term/short-term contraceptives to avoid having more. Since the relative desire to have more or less children is a personal preference that is highly variable for each woman and not included as a data point, the number of children itself was not effective in our model.

iii. What challenges did you find with your data? Where did you get stuck?

The dataset contains multiple inconsistencies in the way the information is conveyed. While some features were ranked 1-4 from highest to lowest, others were ranked in the opposite direction. Moreover, we did not have many numerical values to work with, leading to some limitations in computation and model selection. While the categorical values are ranked with integers, their ordinal nature did not provide adequate indication of how the different labels (1-4) are related.

The data is also heavily dominated with data of women with higher education levels and families with, generally, higher living standards. While this was beneficial in improving our accuracy for women with such attributes, it skewed our model predictions for other women. Initially, we hoped to outsource our data to find more useful information about confounding variables such as average incomes/health insurance/cost of methods to improve the models and consequently prediction scores. However, given that this dataset was conducted in Indonesia in 1987, at which universal healthcare did not exist, we could not find any meaningful facts to supplement the original set.

iv. What are some limitations of the analysis that you did? What assumptions did you make that could prove to be incorrect?

The biggest limitation we faced was that the data is heavily focused on the socio-economic aspects of families and lacking information about other factors such as geographic locations or price

points of contraception, which might be considered as confounding variables. Thus, we were unable to come up with a better generalized model to predict, which in turn is reflected by the mediocre accuracy of the final model. Specifically, with the Logistic Regression, we train one of our models by using carefully chosen education levels and standard of living features under the assumption that they were good predictors of the contraceptive methods, which in turn proved to not yield any improvements. It proved that all features as a whole are better indicators than specific attributes despite them having high correlation and are heavily weighted.

v. What ethical dilemmas did you face with this data?

When we were looking into the relationship between a household's socio-economy and wife's education level, taking into account the context of the dataset, we assumed that a highly educated wife tended to marry a highly educated husband with a well-paying job and vice versa. This is a sticky assumption because it reflected the gender-role stereotype between men and women, especially in traditional Asian countries like Indonesia where women are expected to rely on their husband for financial support.

vi. What additional data, if available, would strengthen your analysis, or allow you test some other hypotheses?

An hypothesis in which we first were interested was that whether the family economy plays a big role in the wife's choice of family planning device. To be able to test this theory, we would need more numerical data, mainly relating to incomes and costs of methods, in order to find a relationship between our response variable and the figures. At the same time, this additional information might also be able to help strengthen our analysis, since we assumed through EDAs that education levels affect the standard of living of a couple.

vii. What ethical concerns might you encounter in studying this problem? How might you address those concerns?

In a more holistic picture, the socioeconomic attributes might be too shallow to be used as main indicators of the choice of contraception. There are many factors that might affect a woman's choice of usage, including but not limited to her accessibility to healthcare, the affordability of the methods in certain countries (Indonesia in this case), or simply her personal preferences or beliefs. These are the unknown variables that we, unfortunately, are not well informed, and hence cannot make a good judgement about. Using a household's financial health alone might lead us to draw biased causal conclusions against low-income people, which as a result deepens the healthcare-derived stereotypes upon poor families. We might address this concern by understanding and acknowledging our limitations with the dataset, and potentially acquire more data to specifically tackle said unspecified attributes to give an objective understanding of the matter at hand.