Data 630 9040

Machine Learning 2215

Professor Bati Firdu

Melissa Hunfalvay

Date: 7-6-2021

Assignment 3

Introduction

Objective

The dataset used for this project was the Contraceptive Method Choice which is a subset of data from the 1987 National Indonesia Contraceptive Prevalence Survey (https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice).

The objective of the analysis was to explore the factors that impact a woman's choice of contraception method. The methods include no use of contraception, long-term methods, or short-term methods. More specifically, using a conditional inference tree (ctree) method of supervised learning, the objective is to categorize the choice of contraception method using input variables of demographic and socio-economic characteristics. Understanding what factors affect a woman's contraceptive choice would be helpful information for the National Family Planning Program as they work to achieve the goals set out by the National Health System in Indonesia.

Supervised learning refers to giving the dataset a predetermined, already known "right answer" or outcome (Ng, 2021). This dataset contained women, in which, for every woman there was a known outcome; the contraception method chosen. Methods of contraception form the dependent variable, and the input variables of demographic and socio-economic characteristics are the independent variables in the dataset. Some examples of the independent variables include, education levels, occupation, standard of living and media exposure.

Problem Domain

In the last part of the 20th century, a National Health System (NHS) was developed in Indonesia. The goals of the program according to the Broad Guidelines for State Policy were to reduce the birth rate, establish a small family norm, and improve the health of mothers and

children. As part of the NHS, family planning services were implemented (NHS Report, 1987).

The National Family Planning Program (NFPP) was established with five principles or targets (Gertler & Molyneaux, 1994):

"First, women under the age of 30 with fewer than two children should plan a maximum of two children; women should delay their first birth to age 20 by postponing marriage and planning births.

Second, women over age 30 and those with three or more children should plan to have no more children and should be offered the most effective means of fertility regulation.

Third, young people should be encouraged to postpone marriage and childbearing through the creation of programs that deemphasize marriage and children as the only means of providing recognition and personal security.

Fourth, in areas with higher rates of contraceptive use, education, basic health services and income generating activities are needed to institutionalize the social benefits of family planning.

Fifth, communities should be assisted in assuming responsibility for care of the aged, to reduce the desire for many children for security in old age."

To achieve these targets, it was important to understand the socio-economic factors and demographic characteristics that impacted a woman's choice of contraceptive method. Therefore, the National Indonesian Contraceptive Prevalence Survey (NICPS) was conducted.

The purpose of the NICPS survey was two-fold. First, to provide data on family planning and fertility behavior of the Indonesian population. With the follow-on objective of evaluating and enhancing the NFPP. The second objective was to measure changes in fertility and

contraceptive prevalence rates and study factors that affect the changes.

Therefore, the purpose of this analysis is to categorize the choice of contraception method using demographic and socio-economic characteristics to inform and evaluate the effectiveness of the NFPP.

Method Rationale

The main methodology chosen is a conditional inference tree (ctree), which is a statistical classification techniques and form of supervised learning.

The rationale for *supervised learning* methodology includes:

a) A known outcome of contraceptive method chosen

b) Variables within the dataset that may be used to evaluate against the known outcomes

The rationale for using the ctree includes:

a) The target/dependent variable (contraceptive method) that forms a multiclassification problem

b) The non-parametric nature of the dataset. Specifically, the data is ordinal and interval. Ordinal data includes the four levels of education and the standard of living index which have values starting at 1 (low) and going to 4 (high). Age, measured in years, is an example of interval data within the dataset.

c) The dependent variable and the independent variables are continuous or can be changed to factors, both of which are well suited to the ctree classification method.

d) The nature of the problem is to inform and help the NFPP with decision making and improvement of services. The if-then statements allow the organization to direct training and resources to appropriate nodes or outcomes found in this type of model.

The NFPP wants to ask questions of a dataset like this to include:

1. Have the training and education objectives impacted beliefs of women regarding their choice of contraception?
2. Are there specific factors (input variables) within the data that inform the National Family Planning Program in ways that may direct future efforts? For example, is the choice of no contraception for a group where contraception is recommended still choosing not to use contraception?
3. Are there specific factors (input variables) that have been impacted by the National Family Planning Program? In other words, are there specific groups of women who have changed their beliefs and now use contraception?
4. Finally, when viewing the NICPS within the context of other efforts within Indonesia, including the National Socio-Economic Survey, has change occurred? One measure to evaluate this outside the NICPS dataset but within the National Socio-Economic Survey may be an examination of birth rates over past decades. A reduction in birth rate may be, at least in part, due to the efforts of the National Family Planning Program.

Analysis

Data

This data set was collected in 1987 as part of National Indonesia Contraceptive Prevalence Survey (NICPS). The data is a subset of the survey containing contraceptive method choice of women in Indonesia. All the women in the dataset were married at the time, were either not pregnant or do not know if they were pregnant at the time. The goal of the data was to understand women's contraceptive choices in Indonesia in 1987, and by doing so evaluate the

NFPP objectives and effectiveness.

The NICPS was administered in 20 of Indonesia's 27 providences. Omitted providences were either logistically difficult or low in population and totaled only 7% of the Indonesian population. The total coverage area where the survey was conducted was 93% of where the population resides. According to the Indonesian Department of Health Services (DHS) report, 27.5% of the women surveyed were urban dwellers the remainder were in more rural areas (DHS report, 1987).

It should be noted that the survey did not include health-related questions because this information was collected in the National Socio-Economic Survey (SUSENAS) in more detail and with wider geographic coverage. Interestingly, the SUSENAS reported a dramatic decrease in fertility levels in Indonesia over the preceding two decades. In 1987, at the time of the NICPS survey the average number of children per family was 3.3 versus 5.5 in the latter half of the 1960's (DHS report, 1987).

The number of observations in this dataset was 1473. Each observation was unique. There were ten variables in the data. The dependent variable chosen was contraception method used which had three levels: no use of contraception, long-term and short-term contraception.

Socio-economic variables included are woman's education from low (1) to high (4), which is a categorical variable. Husbands' education was categorized the same as women's education. Husbands' occupation was also measured on a scale from 1 to 4. Wife's current work status was a binary yes (0) and no (1) variable. A standard of living index was measured from low (1) to high (4) and is a categorical variable. Media exposure was a binary variable measured as good (0) or not good (1).

Demographic variables included the woman's (wife's) age, which is a numeric variable. The number of children ever born, also a numerical variable. Wife's religion was a binary variable, 0 for non-Islamic and 1 for Islamic.

Exploratory Analysis

There are 10 variables in the data set, as shown by the str function (Figure 1). All variables were initially categorized by R as integers (int). There are 1473 rows (or observations) of data. Each row represents one unique woman.

```
> str(cmc)
'data.frame':   1473 obs. of  10 variables:
 $ WifeAge            : int  24 45 43 42 36 19 38 21 27 45 ...
 $ WifeEducation      : int  2 1 2 3 3 4 2 3 2 1 ...
 $ HusbandEducation   : int  3 3 3 2 3 4 3 3 3 1 ...
 $ NumChildren        : int  3 10 7 9 8 0 6 1 3 8 ...
 $ WifeReligion       : int  1 1 1 1 1 1 1 1 1 1 ...
 $ WifeWorking        : int  1 1 1 1 1 1 1 0 1 1 ...
 $ HusbandOccupation  : int  2 3 3 3 3 3 3 3 3 2 ...
 $ LivingStandardIndex: int  3 4 4 3 2 3 2 2 4 2 ...
 $ MediaExposure      : int  0 0 0 0 0 0 0 0 0 1 ...
 $ ContraceptiveMethod: int  1 1 1 1 1 1 1 1 1 1 ...
> |
```

Figure 1: str function output without any pre-processing

An initial review of the data revealed no variables needed to be removed.

The levels within each factor were changed to words for easier analysis and understanding. For instance, as 1 for wife or husband's education was categorized as low, the number 1 was replaced with the word "low". Number 2 was "low-Mod", number three was "mod-high", and number 4 was "high". The same logic was applied throughout the factor variables within the dataset. The titles of each variable were well understood and therefore the variable names remained the same.

Missing data values were checked. None were found in this data set (Figure 2).

```
> # List variables have missing values
> # Check for all variables
> apply(cmc, 2, function (cmc) sum(is.na(cmc)))
            WifeAge       WifeEducation    HusbandEducation          NumChildren        WifeReligion         WifeWorking   HusbandOccupation
                  0                   0                   0                    0                   0                   0                   0
LivingStandardIndex       MediaExposure ContraceptiveMethod
                  0                   0                   0
```

Figure 2: apply command to check for the missing values within each variable in the data set

To further explore the data, the summary command was run for all variables (Figure 3). The summary command displays descriptive statistics that assist in determining distribution of variables and implied skewness. This command is used as a first step in further decision making and understanding of the dataset.

```
> summary(cmc)
    WifeAge      WifeEducation   HusbandEducation  NumChildren      WifeReligion     WifeWorking     HusbandOccupation LivingStandardIndex
 Min.   :16.00   Min.   :1.000   Min.   :1.00     Min.   : 0.000   Min.   :0.0000   Min.   :0.0000   Min.   :1.000     Min.   :1.000
 1st Qu.:26.00   1st Qu.:2.000   1st Qu.:3.00     1st Qu.: 1.000   1st Qu.:1.0000   1st Qu.:0.0000   1st Qu.:1.000     1st Qu.:3.000
 Median :32.00   Median :3.000   Median :4.00     Median : 3.000   Median :1.0000   Median :1.0000   Median :2.000     Median :3.000
 Mean   :32.54   Mean   :2.959   Mean   :3.43     Mean   : 3.261   Mean   :0.8506   Mean   :0.7495   Mean   :2.138     Mean   :3.134
 3rd Qu.:39.00   3rd Qu.:4.000   3rd Qu.:4.00     3rd Qu.: 4.000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:3.000     3rd Qu.:4.000
 Max.   :49.00   Max.   :4.000   Max.   :4.00     Max.   :16.000   Max.   :1.0000   Max.   :1.0000   Max.   :4.000     Max.   :4.000
 MediaExposure   ContraceptiveMethod
 Min.   :0.000   Min.   :1.00
 1st Qu.:0.000   1st Qu.:1.00
 Median :0.000   Median :2.00
 Mean   :0.074   Mean   :1.92
 3rd Qu.:0.000   3rd Qu.:3.00
 Max.   :1.000   Max.   :3.00
>
```

Figure 3: Summary Command showing descriptive statistics for all variables in the data set.

For each factor variable additional exploration was conducted to include using the head command for showing the first 100 rows of data to explore frequency of outputs; a count of the levels within the factors to determine distribution across the levels within the variable; and a percentage of each level within the factor (Figure 4).

```
> # Exploratory analysis - Wifes' Religion, Variable type = factor
> # Change 0:1 levels to descriptive terms for easier analysis
> cmc$WifeReligion<-factor(cmc$WifeReligion, levels = 0:1, labels = c("Non-Islamic", "Islamic"))
> # show first 100 raw variables
> head(cmc$WifeReligion, 100)
  [1] Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic
 [14] Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Non-Islamic Islamic     Islamic     Islamic
 [27] Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic
 [40] Islamic     Islamic     Islamic     Islamic     Non-Islamic Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Non-Islamic
 [53] Islamic     Non-Islamic Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic
 [66] Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic
 [79] Islamic     Islamic     Non-Islamic Non-Islamic Islamic     Non-Islamic Islamic     Islamic     Islamic     Non-Islamic Islamic     Islamic     Islamic
 [92] Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic     Islamic
Levels: Non-Islamic Islamic
> # Count the number of values within each category
> table(cmc$WifeReligion)

Non-Islamic     Islamic
        220        1253
> # Percentage
> table(cmc$WifeReligion)/length(cmc$WifeReligion)

Non-Islamic     Islamic
  0.1493551   0.8506449
```

Figure 4: Additional exploratory analysis for the Wife's religion variable

Factor variables were also visualized using a bar chart (Figure 5). These visualizations helped see the data and assist with further processing and decision making. Furthermore, it is very important in socioeconomic research to view these variables in context with population normative values at the time. Ideally, the survey is a microcosm of the same proportion of values in each variable. For example, if education levels were high in Indonesia in 1985, then the NICPS survey, as a microcosm of the society should also reflect high education levels. If that is not the case, and sampling does not reflect the same proportion of the societies normative values at the time, then the results of the survey may be difficult to generalize (Sauver, Grossardt, et al., 2012).

Figure 5: Standard of Living bar plot

It can be observed via the visualizations and descriptive statistics that there were more males that had a “high” level of education (61%) compared to females (39%). Interestingly, the percentage of males in the high education category was more than any other category (Figure 6). Indicating that the male group in the survey were highly educated. However, when examining

the husband's occupation, the smallest percentage was in the "high" category (1.8%; Figure 7). This is a little perplexing as usually a high education leads to a high occupation. This may be explained by the age of the husband; however, this variable is not in the data set. It could also be explained by the improvements in the economic welfare in Indonesia at the time, which resulted in expansion of the education system (Gertler & Molyneaux, 1994). Over the few decades prior to the survey education rates rose in Indonesia from 50% in 1961 to 60% in 1971 and to 94% in 1985 (Biro Pusat Statistick, 1975, 1983, 1987). Conceivably the education level of the males could be high, but if they were young, they may just be starting their careers hence the very low level of "high" in the occupation status.

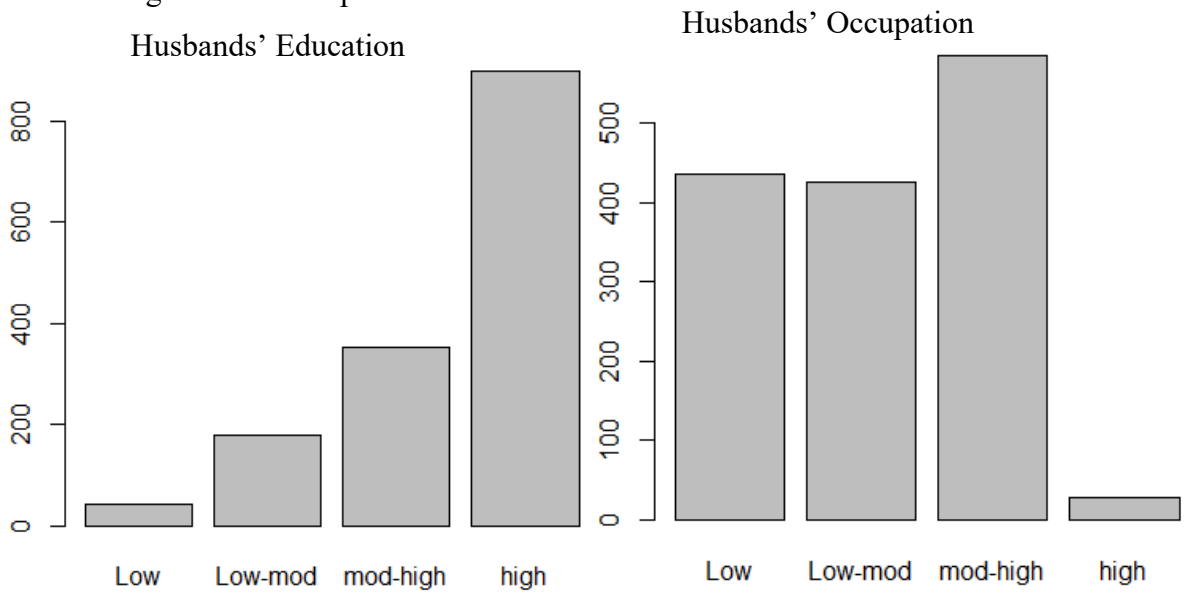

Figure 6: Husbands' Education is high; Figure 7: Husband's occupation is low.

Wife's education and working status were also examined. The woman in the survey, when combining both moderate to high and high levels of education were predominantly well educated (28% and 39% respectively, total = 67%, see Figure 7). However, at the time many of the women were not working (75%, Figure 8). In 1987, women in the workforce in the United States was 64.6% (Toossi & Morisi, 2017). However, unlike in Westerns societies, a combination of custom, religion and societal norms shows women in Indonesia in the workplace

at the same time was 33% (Wright & Tellei, 1993). Therefore, the NICPS survey is reflective of women in the workplace in 1987.

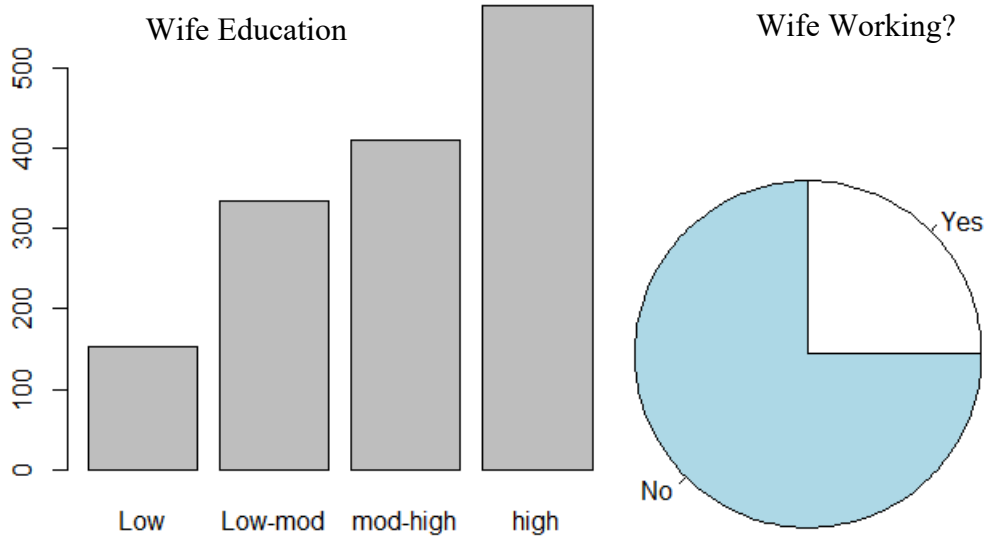

Figure 7: Womans level of education; Figure 8: womans working status at the time of the NICPS survey

The Standard of Living Index revealed close to half the people in the survey had a high standard of living (46%). When combined with the moderate to high level (29%) this included a large majority of the people surveyed (75%). During the 1960's to mid-1990's massive improvements occurred in many dimensions of standards of living in Indonesia and the larger regions of East and Southeast Asia (RAND, 1997). In Indonesia, real per capita Gross Domestic Profit (GDP) rose four-fold between 1965 and 1995 (RAND, 1997). In the same period, poverty rates declined from over 40% to just under 18% by 1996 (RAND, 1997). Therefore, the NICPS survey showing rates of 75% from moderately high to high standards of living seem like an accurate microcosm reflection of the status of Indonesian society in 1985.

Consistent with the religious make-up of the country, 85% of the survey population identified themselves as Muslims. Muslims are the predominant religion in Indonesia, at 87% of people reporting this religious preference (Statistica, 2010.)

For the two numeric variables (int) the summary command and standard deviation and mode were calculated (Figure 9). As it is important to check for "data sanity" among the variables and therefore the additional calculations were deemed prudent and appropriate to include in the exploratory analysis.

```
> # Exploratory analysis - Wifes' Age Variable type = int
> # Descriptive Statistics
> summary (cmc$WifeAge)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16.00   26.00   32.00   32.54   39.00   49.00
> # Standard deviation
> sd(cmc$WifeAge)
[1] 8.227245
> # Mode
> Data_WifeAge <- cmc
> names(sort(-table(Data_WifeAge$WifeAge)))[1]
[1] "25"
```

Figure 9: Summary command, standard deviation, and calculation of the mode for woman's age

Visualizations for numeric variables included histograms (Figure 10) and boxplots. These were considered appropriate for the continuous nature of the numeric variables to show both distribution and outliers. These visualizations along with the descriptive calculations were used to make further decisions regarding data cleaning and processing.

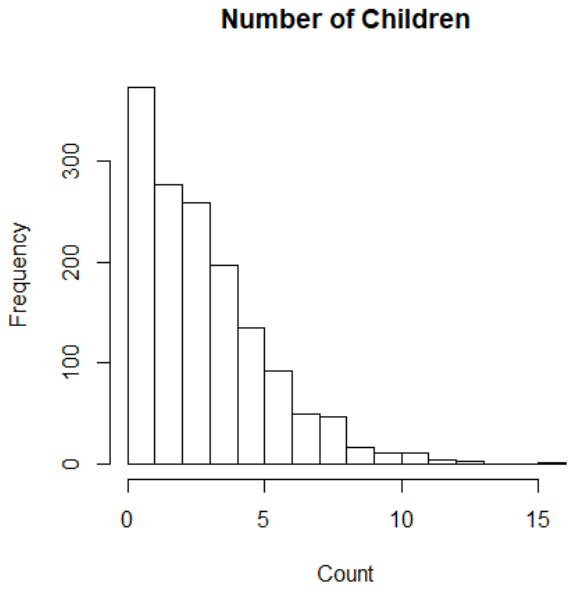

Figure 10: Histogram of number of children

From the exploratory analysis e average age of the woman in the data set was 32.54 years (*SD* = 8.23). Given these statistics and the histogram (Figure 11) the woman's age data was reasonably distributed throughout the range from 16 to 49 years of age.

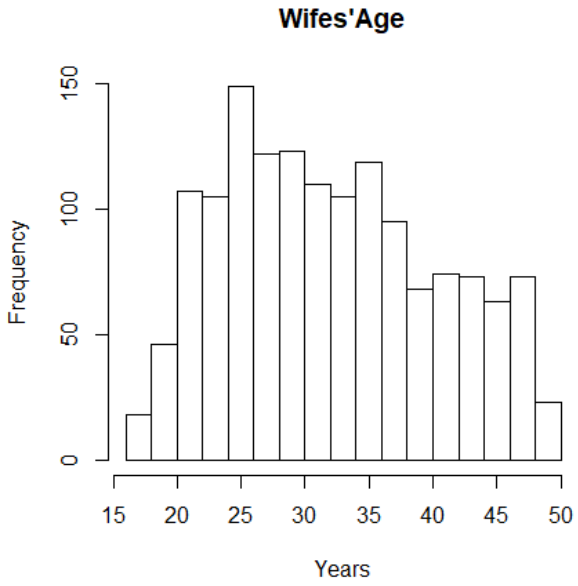

Figure 11: Woman's Age

The second integer variable was number of children. Exploratory analysis showed the number of children per household ranged from none to 16 children. The average number of children was 3.3 and is consistent with the official DHS report (DHS, 1987).

Preprocessing

Armed with the objective and exploratory analysis, as well as the model type preprocessing was conducted. Preprocessing for this analysis was very limited as there was no missing values and no variables needed removing. There were two pre-processing steps however, first, int were changed to factors. Second, data sanity checks were conducted were needed.

Changing integers to factors: Some of the numeric "int" variables needed to be transformed to factors for the analysis as they would make more sense when evaluating the ctree model. This included all categorical and binary variables. This was important as without the change to a factor the tree may have displayed values in between the categories, such as 0.5 for religion which would have been a nonsensical value.

After the integers were changed the str function was then re-run showing the changes in the dataset (Figure 12). At this point in the exploratory analysis there were ten variables, two were numeric (2 = int) and eight factors. Factors are discrete, categorical variables with pre-determined levels such as non-Islamic and Islamic.

```
> # 3. Data Pre-processing
>
> # a) change categorical and binary variable to factors
> # done above as part of the labeling changes
> str(cmc)
'data.frame':   1473 obs. of  10 variables:
 $ WifeAge            : int  24 45 43 42 36 19 38 21 27 45 ...
 $ WifeEducation      : Factor w/ 4 levels "Low","Low-mod",..: 2 1 2 3 3 4 2 3 2 1 ...
 $ HusbandEducation   : Factor w/ 4 levels "Low","Low-mod",..: 3 3 3 2 3 4 3 3 3 1 ...
 $ NumChildren        : int  3 10 7 9 8 0 6 1 3 8 ...
 $ WifeReligion       : Factor w/ 2 levels "Non-Islamic",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ WifeWorking        : Factor w/ 2 levels "Yes","No": 2 2 2 2 2 2 2 1 2 2 ...
 $ HusbandOccupation  : Factor w/ 4 levels "Low","Low-mod",..: 2 3 3 3 3 3 3 3 3 2 ...
 $ LivingStandardIndex: Factor w/ 4 levels "Low","Low-mod",..: 3 4 4 3 2 3 2 2 4 2 ...
 $ MediaExposure      : Factor w/ 2 levels "Good","Not Good": 1 1 1 1 1 1 1 1 1 2 ...
 $ ContraceptiveMethod: Factor w/ 3 levels "No Use","Long Term",..: 1 1 1 1 1 1 1 1 1 1 ..
```

Figure 12: str function output after renaming and transforming variables

Data sanity checks: For the numeric variables it can be observed that one woman had 16 children (Figure 11)! Is this correct? A sanity check was conducted on this to determine if the data was legitimate or an outlier. The following steps were taken to help make this decision; first, a summary function was used on the women's age. This revealed an age range from 16-48 years (Figure 13). These are legitimate age ranges for giving birth to children (Healthline, 2020).

Next, the summary command was used to check the number of children born to determine the range and compare it to the histogram. Indeed, the largest number of children born by one woman was 16. Finally, the age of the women who had 16 children was checked. The woman was 48 years old. As it is possible for a woman of 48 to have 16 children then the data remained legitimate and was not cut from the data set.

```
> # b) Explore numeric variables for data sanity check
> # Wife's age (numerical) - check range
> summary(cmc$WifeAge)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  16.00   26.00   32.00   32.54   39.00   49.00
> hist((cmc$WifeAge))
> # Number of children ever born (numerical) - check range
> summary(cmc$NumChildren)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  0.000   1.000   3.000   3.261   4.000  16.000
> hist((cmc$NumChildren))
> boxplot(cmc$NumChildren) # shows outliers, explore more
> max(cmc$NumChildren)
[1] 16
> cmc[which(cmc$NumChildren>15),"WifeAge"] # 48 year old, legitimate
[1] 48
```

Figure 13: Sanity data check on number of children

Algorithm Intuition

The Conditional Inference Tree (ctree) is a non-parametric class of regression trees applicable for numeric (int) and factor variables among other variables (Horthorn, Hornik, Zeileis, 2006). The ctree method creates an inverted decision tree based on statistical rules such

as a p value of 0.05. The tree begins with a "root" at the top, which is the first variable, and then splits based on decision rules and if-then-else logic that creates branches from the root to the "nodes" or "leaves" at the bottom of the tree. A "splitting node" refers to a leaf or node where further decisions need to be made resulting in a further split of the data.

The dependent variable in this dataset is nominal and has 3 levels (no use, short-term and long-term contraceptive choice). Therefore, the ctree will be a multi-way (3-way) split as there are three partitions as distinct categories.

Classification decision trees are used to predict qualitative responses. Each unique observation is used to predict the most common class of training observations from within the region to which it belongs (James, Whitten, Hastie, Tibshirani, 2013). When interpreting the classification tree, it is often of interest to determine not only the class prediction corresponding to the terminal leaf, but also the proportion of each class among the training data observations as this assist in understanding the broader context of the qualitative nature of the dataset.

The ctree method is built in two steps, the first is the initial construction, and the second is pruning. In initial construction, the first question asked of the data is what is the variable that optimizes the given criteria? Then, does the data fit a threshold? If it does, then continue doing further splitting. This *recursive splitting* continues incrementally, through a "divide and conquer" approach until either there are no more observations in the same class, and/or, no more attributes remain and/or the specified stopping criteria is met. The goal of this process is to increase the sameness (homogeneity) of each partition with respect to the dependent variable.

At each splitting decision the model checks against the thresholds for that specific splitting node. In other words, the decisions are made at the local or node level making them optimal for that node, even if they are not optimal for the entire, global model.

The principle algorithm used for the ctree is the calculation of the prediction accuracy or the proportion of each class (Figure 14).

$$\hat{p}_k = \frac{1}{N_m} \sum_{x_i \epsilon R_m} I(y_i = k)$$

Figure 14: Prediction accuracy algorithm for ctree classifications

Where $I(y_i = k)$ is the indicator function. This function maps the input into a class. For instance, non-Islamic is 0 so it would map to a 0 class and Islamic is 1 which would map to the 1 class. Where $\sum_{x_i \epsilon R_m}$ sums the observations and shows that the X should fall within the region of observation. Finally, $\frac{1}{N_m}$ divides divide by the total number of observations.

To apply this formula, if an observation falls within the accuracy level obtained by the formula, then you will predict it as that category of interest. For instance, if the accuracy of prediction for women's religion is .9 and a future observation falls within that region then it is categorized as that region of interest.

For growing ctrees there are two different key algorithms in the initial construction phase. These include the ID3 and C4.5 formulas which calculate *information gain* and the CART algorithm which forms the *Gini Index.*

Information gain uses entropy, which is a lack of order or predictability, to measure the uncertainly in each variable. If all observations fall into one class within the variable, then the

entropy is zero, and that would refer to a pure data set where there is no uncertainty. An entropy of 1 is maximum lack of order. The formula for entropy of a binary category (Figure 15) is taking the proportion of observations, divided by the number of total observations from the class, multiplied by the log of that proportion and then sum over the different classes. In the case of this dataset, we would use 3 k's and sum over three classes as we have three possible outcomes for the dependent variable.

$$H(D) = -\sum_{k=1}^{k} p_k log p_k$$

Figure 15: Entropy algorithm for information gain in ctree split measures

The entropy value for the entire data set is calculated. Then the entropy for each split is calculated along with the factors as a proportion of each split (i.e., number of observations from a given split divided by total number of observations). These values are then added for each split to find the total gain value. Whereby the difference in the marginal entropy (i.e., entropy before the split) and then the entropy after the split, the difference between these two values is called *information gain*. The split with the largest information gain, that is the largest reduction in entropy, (uncertainty) is selected.

The *Gini Index* (Figure 16) counts the classification error and is generally less used than the information gain metrics. Nevertheless, the Gini Index takes the compliment of the proportion, and it is summed over the number of classes. The Gini Index and Information Gain produce the same result.

$$\mathrm{G} = -\sum_{k=1}^{k} p_k(1 - p_k)$$

Figure 16: Gini Index algorithm

The second step in the ctree method is *pruning*. Trees may be pre-pruned or post-pruned. Pre-pruning occurs as part of the learning process and tree construction, whereas post-pruning occurs after the tree is constructed to remove branches. Pruning occurs when the tree has too many branches which may cause *overfitting.* Overfitting occurs when there are too many branches (or splitting variables) within the tree making it difficult for the tree to generalize. Poor generalization results in poor accuracy of new observations. This can also be seen when comparing the training data set to the test data set. If the accuracy between these two data sets is very different then overfitting may be the cause.

There are several strategies to avoid overfitting. One example of pre-pruning includes building stopping conditions, whereby if a condition is met, the tree stops building. An example of post-pruning to avoid overfitting may be to trim nodes on the tree, starting at the bottom of the tree and moving up.

Model Fitting

The key steps used to fit the model were:

Step 1: To make sure the results were reproducible by using the set.seed command

Step 2: To split the data into 70% training data and 30% test data. Inspect the results via the str. Command (Figure 17).

```
> str(train.data)                          > str(test.data)
'data.frame':   1044 obs. of  10 variables 'data.frame':   429 obs. of  10 variables
```

Figure 17: Results of splitting data into training data (left hand side) and test data (right hand side)

Step 3: Build the model using the dependent (target) variable all the independent variables (Figure 18). Contraceptive Method variable includes three levels: no-use, short-term, long-term.

```
model<-ctree(ContraceptiveMethod~., train.data)
```

Figure 18: shows the dependent variable as Contraceptive Method and all independent variables

Step 4: Print and inspect the model from the first iteration of training set data

Step 5: Create a tree on the training data using the simple method to visualize the results

Step 6: Print and inspect the confusion matrix

Step 7: Print and inspect addition statistics to include sensitivity, specificity, positive predictive values, negative predictive values, prevalence, detection rate, detection prevalence and balanced accuracy.

Step 8: Print and inspect the training data classification accuracy

Step 9: As accuracy was not as high as anticipated then to check for overfitting compare training classification accuracy to test classification accuracy

Step 10: As both test and training classification accuracy was only moderate then iterate on the tree by pre and post pruning to include:

a) Pre-prune by setting the maximum tree depth

b) Post-prune by examining the nodes from various positions within the tree

Step 11: Experiment and iterate over the variables. Use decision making logic as it pertains to the objective or question being asked of the data.

Step 13: Do a final summary and inspection of the model in relation to the stated objectives

Results

Output

The default parameters of the model included all independent variables, and the target (dependent) variable was contraception method (no-use, short-term, long-term). Results of the textual output for the summary model can be seen in Figure 19.

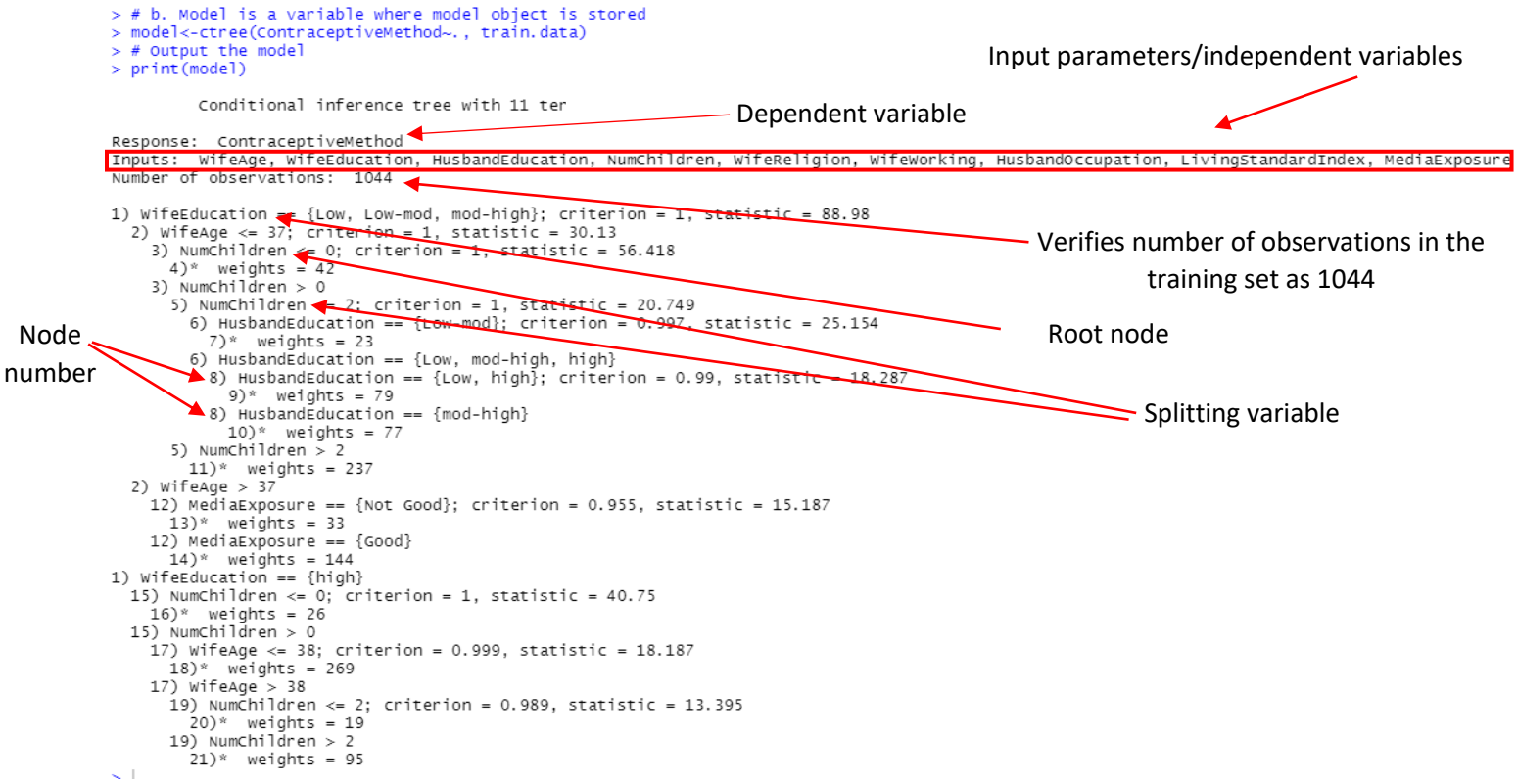

Figure 19: Shows the textual structure of the model including node splitting and leaf nodes.

Results of the first version of the ctree (Figure 20) and related statistics (Figure 21) resulted in a less than optimal result, including an overall model accuracy of 58%. When examined against test data the test accuracy revealed just 53%. This is not a satisfactory level of prediction classification accuracy. Furthermore, as there is not much disparity between the training and test accuracy and therefore, it is unlikely that overfitting caused the lack of

accuracy. It is possible that the limited number of observations is impacting the accuracy along with the three levels of the target variable. Therefore, the stated objective and analytics question was reviewed and revisited. A slightly different angle was examined that still fit the stated objective.

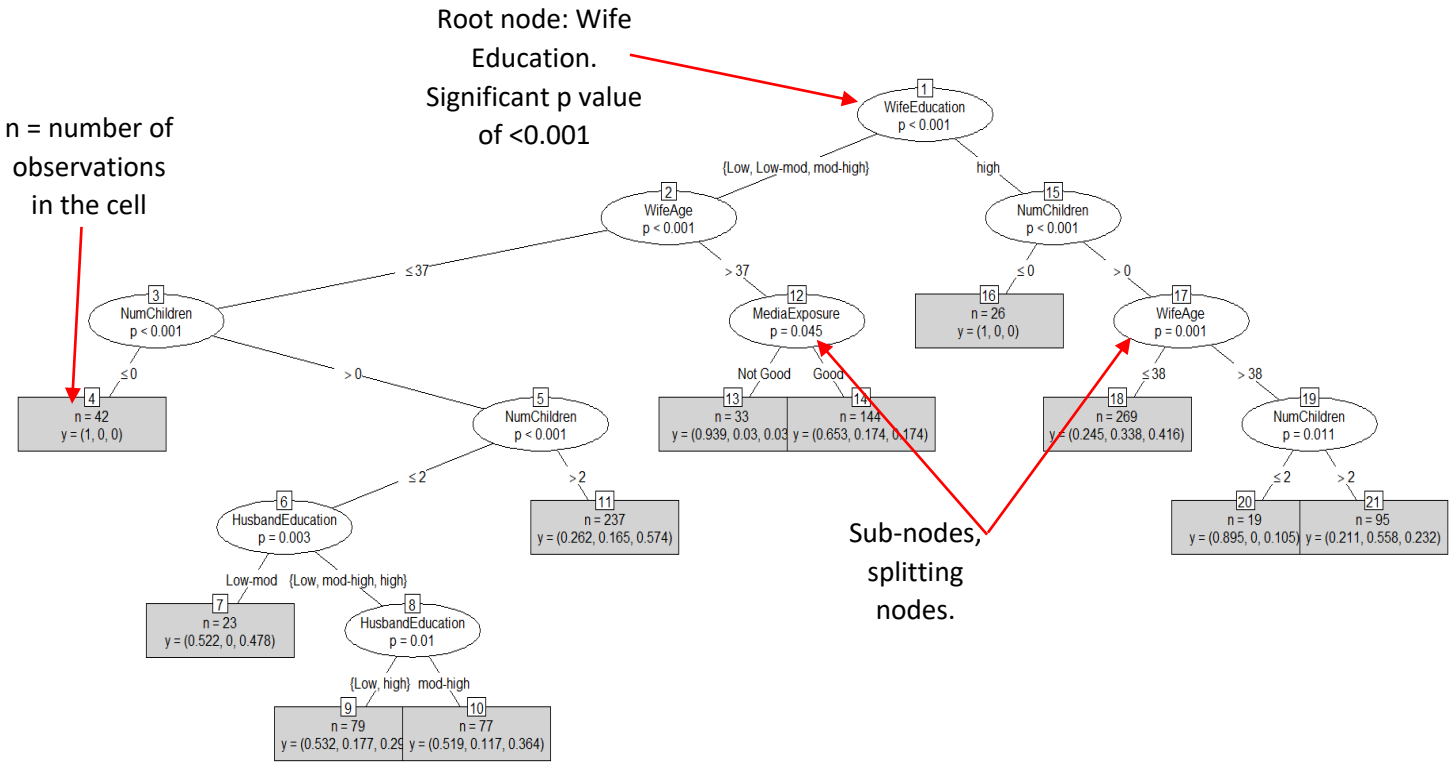

Figure 20: First version of the ctree

```
> trainPred <- predict(model, newdata = train.data, method="ContraceptiveMethod")  # predictions for the train data
> prop.table(table(predict(model), train.data$ContraceptiveMethod))

                No Use  Long Term Short Term
  No Use     0.29118774 0.04693487 0.08620690
  Long Term  0.01915709 0.05076628 0.02107280
  Short Term 0.12260536 0.12452107 0.23754789
> confusionMatrix(trainPred, train.data$ContraceptiveMethod, dnn=c("predicted", "actual")) # additional statistics
Confusion Matrix and Statistics

           actual
predicted    No Use Long Term Short Term
  No Use        304        49         90
  Long Term      20        53         22
  Short Term    128       130        248

Overall Statistics

               Accuracy : 0.5795
                 95% CI : (0.5489, 0.6097)
    No Information Rate : 0.433
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3314

 Mcnemar's Test P-Value : < 2.2e-16

Statistics by Class:

                     Class: No Use Class: Long Term Class: Short Term
Sensitivity                 0.6726          0.22845             0.6889
Specificity                 0.7652          0.94828             0.6228
Pos Pred Value              0.6862          0.55789             0.4901
Neg Pred Value              0.7537          0.81138             0.7918
Prevalence                  0.4330          0.22222             0.3448
Detection Rate              0.2912          0.05077             0.2375
Detection Prevalence        0.4243          0.09100             0.4847
Balanced Accuracy           0.7189          0.58836             0.6558
> sum(predict(model)== train.data$ContraceptiveMethod)/length(train.data$ContraceptiveMethod)  # classification accuracy
[1] 0.5795019
```

Model classification accuracy

Figure 21: Statistical results of the first model generated

The stated objective of the analysis was to explore the factors that impact a woman's choice of contraception method. The methods include no use of contraception, long-term methods, or short-term methods. Understanding what factors affect a woman's contraceptive choice would be helpful information for the National Family Planning Program. The objective remained the same, and the goal to help inform the National Family Planning Program remains. However, instead of viewing the dependent target variable in three levels based on type of contraceptive choice the two positive contraceptive choices (short-term and long-term) were collapsed forming two factors – no use of contraception (1) and use of contraception (0). Use of contraception whether short- or long-term were combined to form one level or factor and renamed as use of contraception.

Results of the two-factor dependent variable significantly improved the model. Accuracy of the overall model on the training data set was 73% (Figure 22). If the level of accuracy is between 70-80% the model is categorized as a "good" model (Vallantin, 2020). Other variables including a 77% sensitivity rate and a 68% specificity rate provides further evidence that the model is sound. Sensitivity, specificity, and accuracy on the test data was also examined at this time and was within range of the training data results.

The ctree model revealed 19 nodes with significant levels of greater the 0.05. The root node was wife's' education level, which was split between high and low, low-mod and mod-high. When following the high level of wife's education, the number of children become the first leaf variable which was split of whether the woman had children (>0) or did not have children (<=0). When following the branch in the other direction, where the wife's education was high and low, low-mod and mod-high the wife's age became the first leaf (or sub node) on the tree. Wife's age was split at less than or equal to 36 years of age or greater than 36 years of age. Other factors within the tree included Living standard Index media exposure and wife's religion.

In summary, these results are informative for the National Health System in Indonesia in that they can evaluate the effectiveness of the National Family Planning Program's goals with a good level of accuracy. More specifically, both organizations can view the woman's choice of using contraception versus not using contraception in terms of measurable predictive variables that the organizations can work to influence. Most encouragingly would be the root node variable of wife's education level. Therefore, the stated objective to explore the factors that impact a woman's choice of contraception method was achieved by this analysis.

Model Properties

After iterating with various pruning strategies and collapsing the dependent variable into two levels, results revealed the best ctree model. The textual output of the training model is in Appendix A. The ctree included a maximum tree depth of five levels (Figure 22).

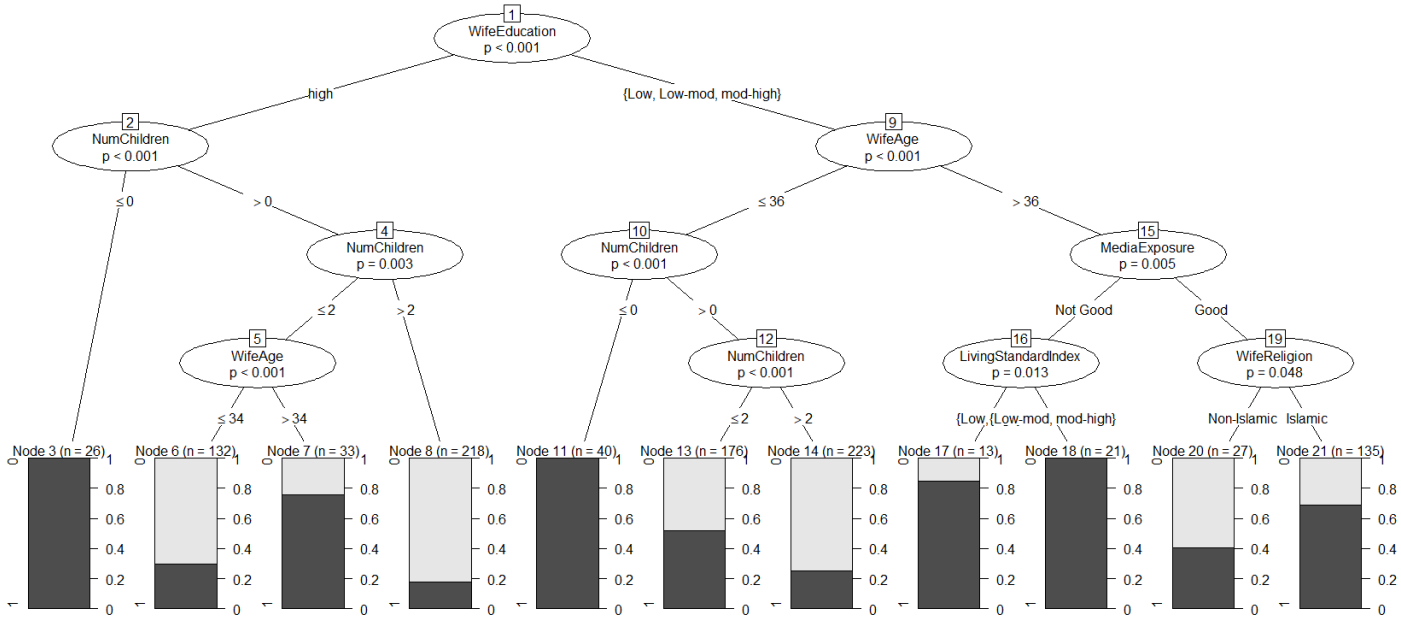

Figure 22: ctree output from no-use to use of contraception. 1(No-use), 0(use - short or long term)

The root node is wife's education. This splits at a p value less than 0.001. Moving towards Node 8, the question to be asked of the data is: Is wife's education level high? If yes, traverse the tree to the next child node which is number of children. Again, the p value less than 0.001.

Next question to ask of the data is: Does the woman have children? If the woman has children, then traverse the tree to the number of children again. Ask the following question of the data: Does the women have more than two children? If the answer is yes, traverse the tree to node 8, which as a total of 218 women from the entire dataset who fall in this node. The dark area of the box relates to no-use of contraception. The light area relates to the use of some form

of contraception (short or long-term). Therefore, in interpreting node 8; women who are highly educated with more than two children have a probability of 80% in using contraception and a probability pf 20% in not using contraception. Table 1 shows a summary of the interpretation of each node within the tree.

Table 1: Summary interpretation of each node

Node	**Interpretation**
3 N =24	Highly educated women with no children have a 100% probability of not using contraception.
6 N = 132	Highly educated women with less than 2 children who are 34 years of age or less are predicted to use contraception at a 70% probability and not use contraception at a 30% probability.
7 N = 33	Women who are highly educated with less than 2 children and are older than 34 years of age have approximately 70% probability of not using contraception. Their probability of using contraception is 30%,
8 N = 218	Women who are highly educated with more than two children have an 80% probability of using contraception. Their probability of not using contraception is 30%
11 N = 40	Women who have low, low to moderate or moderate to high levels of education, who are also 36 years of age or younger, with no children are 100% probability of not using contraception.
13 N = 176	Women who have low, low to moderate or moderate to high levels of education, who are also 36 years of age or younger, who have less than or equal to 2 children have a 50% probability of using or not using contraception.
14 N = 223	Women who have low, low to moderate or moderate to high levels of education, who are also 36 years of age or younger, who more than 2 children have a 75% probability of using contraception.

17 N = 13	Women who have low, low to moderate or moderate to high levels of education, who are also older than 36 years of age, and whose media exposure is not good and whose standard of living is low have an 80% probability of not using contraception. They have a 20% probability of using contraception.
18 N = 21	Women who have low, low to moderate or moderate to high levels of education, who are also older than 36 years of age, and whose media exposure is not good and whose standard of living is low to moderate or moderate to high have a 100% probability of not using contraception.
20 N = 27	Women who have low, low to moderate or moderate to high levels of education, who are also older than 36 years of age, and whose media exposure is good and whose wife's religion is not Islamic have 60% probability of using contraception and 40% probability of not using contraception.
21 N = 135	Women who have low, low to moderate or moderate to high levels of education, who are also older than 36 years of age, and whose media exposure is good and whose wife's religion is Islamic have a 70% probability of not using contraception and a 30% probability of using contraception.

When reviewing the model, the overall predictive accuracy was 73% on the training data. Sensitivity was 77%, specificity was 68%, positive predictive value was 76%, negative predictive value was 69%. These results indicate that the model on training data was good and should then be evaluated on the test data.

Evaluation

Evaluate the model on the test data: The model accuracy on the test data was found to be 67% (Figure 23). A 6% discrepancy between the training and test data was deemed acceptable and not indicative of overfitting.

```
> # f.  Evaluate the model on test data
> confusionMatrix(testPred, test.data$ContraceptiveMethod, dnn=c("predicted", "actual")) # additional statistics
Confusion Matrix and Statistics

         actual
predicted   0   1
        0 177  68
        1  75 109

               Accuracy : 0.6667
                 95% CI : (0.6199, 0.7112)
    No Information Rate : 0.5874
    P-Value [Acc > NIR] : 0.0004488

                  Kappa : 0.3163

 Mcnemar's Test P-Value : 0.6158468

            Sensitivity : 0.7024
            Specificity : 0.6158
         Pos Pred Value : 0.7224
         Neg Pred Value : 0.5924
             Prevalence : 0.5874
         Detection Rate : 0.4126
   Detection Prevalence : 0.5711
      Balanced Accuracy : 0.6591

       'Positive' Class : 0

> sum(testPred== test.data$ContraceptiveMethod)/length(test.data$ContraceptiveMethod)          # classification accuracy
[1] 0.6666667
> |
```

Figure 23: Model output from test data no-use to use of contraception

Confusion Matrix (Figure 23): results from the confusion matrix command show how many records in the test data have each predicted no use or use of contraception. The number of correctly classified instances in the test data set = 177 + 109 = 286. The number of misclassified instances in the test data set = 75 + 68 = 125. The total number of instances in the test dataset = 198 + 125 = 429 (which corresponds to the str command output in Figure 16 of test data observations). The classification accuracy is the sum of numbers on diagonal/sum of all numbers = 286/429 = ***67%*** *classification accuracy.*

Sensitivity Rate /True Positives: these are rates that correctly classify the contraception method (use or no-use) class belonging to the negative class value. The number of true positives in the dataset is 177 and is found at the top left cell of the confusion matrix (Figure 23).

Formula: true positive/total actual positive

Calculation: 177/(177 + 109) = 0.7024 or **70%**

Specificity Rate /True Negative: these are the rates that correctly identify contraception method (use or no-use) class belonging to the positive class value. The number of true negatives in the dataset is 109 and is found at the bottom right of the confusion matrix (Figure 23).

Formula: true negative/total actual negative

Calculation:109/(177 + 109) = 0.6158 or **62%**

Positive Predictive Value: is the true positives divided by the predicted positives. It is found by using the top row of the confusion matrix (Figure 23).

Formula: True positive/predicted as positive

Calculation: 177/(177 + 68) = 0.7224 or **72%**

Negative Predictive value: is the true negatives divided by the predicted negatives. It is found by using the bottom row of the confusion matrix (Figure 23).

Formula: True negative/predicted as negative

Calculation: 109/(75 + 109) = 0.5924 or **59%**

False Positive: also known as a *Type 1 error* occur when the null hypothesis is incorrectly rejected. The creates a “false positive” that leads to a conclusion that the alternate

hypothesis is true when it is not. The number of false positives in the dataset is 75 and is found in the bottom left of the confusion matrix (Figure 23). Therefore, **75** people in this dataset may be misclassified as having a significant difference when there is not one.

False Negative: also known as a *type II error* is the non-rejection of a false null hypothesis. Whereby a true difference is not found. The number of false negatives in the dataset is **68** and is found at the top right of the confusion matrix (Figure 23).

Prevalence: is the number of true positives over the total number of instances. In other words, how often does the "yes" condition occur.

Formula: True positive/total number of instances

Calculation: 177/(177+68+75+109) = 0.5874 or **59%**

Precision: is the number of true positives over the number of predicted positives (Figure 23). In other words, when the model predicts yes, how often is it correct?

Formula: true positive/number of predicted positives

Calculation: 109/(109 + 68) = 0.6158 or **62%**

Kappa statistic: Is the agreement between predicted and observed considering the accuracy by chance (Figure 23). Values range from -1 to +1 which indicate perfect agreement. Values closer to zero indicate no agreement. This data set has a Kappa value of 0.3163, which is low.

Formula: (Bati, 2021):

$$Kappa = \frac{n_a - n_c}{n - n_c} = \frac{p_a - p_c}{1 - p_c}$$

$n: number\ of\ cases$

$n_a: number\ of\ agreement$

$n_c: number\ of\ agreement\ due\ to\ chance$

$$p_a: proporiotn\ of\ observation\ in\ agreement,$$

$$p_c: proportion\ of\ agreement\ due\ to\ chance$$

Conclusion

Summary

The objective of the analysis was to explore the demographic and socio-economic characteristics factors that impact a woman's choice of contraception method (use or no use).

The purpose of such an analysis was designed to inform the National Family Planning Program as they work to achieve the goals. The National Family Planning Program (NFPP) was established with five goals or targets (Gertler & Molyneaux, 1994). Each goal is evaluated in relation to each terminal node of the ctree analysis (Table 2). Furthermore, the outcome or interpretation of that nodes predictability is evaluated to determine if the results are in line with the goal (green) or disagreement with the target goals (red, Table 2). Although not a perfect comparison, as predications are not perfect and there may be other factors to consider, the table is intended to be a guide to assist the NFPP in determining where their programs are working and what areas they need to focus on.

Table 2: Linking survey results to NFPP goals

Node	**Interpretation**	**Related Goal/Target for National Family Planning Program**
3 N =24	Highly educated women with no children have a 100% probability of not using contraception.	*Target 1: women should delay their first birth to age 20 by postponing marriage and planning births.*
6 N = 132	Highly educated women with less than 2 children who are 34 years of	*Target 2: women over age 30 and those with three or more children should plan to have no more children*

	age or less are predicted to use contraception at a 70% probability and not use contraception at a 30% probability.	*and should be offered the most effective means of fertility regulation.*
7 N = 33	Women who are highly educated with less than 2 children and are older than 34 years of age have approximately 70% probability of not using contraception. Their probability of using contraception is 30%,	*Target 2: women over age 30 and those with three or more children should plan to have no more children and should be offered the most effective means of fertility regulation.* *Target 4: in areas with higher rates of contraceptive use, education, basic health services and income generating activities are needed to institutionalize the social benefits of family planning.*
8 N = 218	Women who are highly educated with more than two children have an 80% probability of using contraception. Their probability of not using contraception is 30%	*Target 2: women over age 30 and those with three or more children should plan to have no more children and should be offered the most effective means of fertility regulation.* *Target 4: in areas with higher rates of contraceptive use, education, basic health services and income generating activities are needed to institutionalize the social benefits of family planning.*
11 N = 40	Women who have low, low to moderate or moderate to high levels	*Target 1: women should delay their first birth to age 20 by postponing marriage and planning births.*

	of education, who are also 36 years of age or younger, with no children are 100% probability of not using contraception.	*Target 2: Second, women over age 30 and those with three or more children should plan to have no more children and should be offered the most effective means of fertility regulation.*
13 N = 176	Women who have low, low to moderate or moderate to high levels of education, who are also 36 years of age or younger, who have less than or equal to 2 children have a 50% probability of using or not using contraception.	*Target 2: Second, women over age 30 and those with three or more children should plan to have no more children and should be offered the most effective means of fertility regulation.*
14 N = 223	Women who have low, low to moderate or moderate to high levels of education, who are also 36 years of age or younger, who more than 2 children have a 75% probability of using contraception.	*Target 2: women over age 30 and those with three or more children should plan to have no more children and should be offered the most effective means of fertility regulation.* *Target 4: in areas with higher rates of contraceptive use, education, basic health services and income generating activities are needed to institutionalize the social benefits of family planning.*
17 N = 13	Women who have low, low to moderate or moderate to high levels of education, who are also older	*Target 2: Second, women over age 30 and those with three or more children should plan to have no more*

	than 36 years of age, and whose media exposure is not good and whose standard of living is low have an 80% probability of not using contraception. They have a 20% probability of using contraception.	*children and should be offered the most effective means of fertility regulation.*
18 N = 21	Women who have low, low to moderate or moderate to high levels of education, who are also older than 36 years of age, and whose media exposure is not good and whose standard of living is low to moderate or moderate to high have a 100% probability of not using contraception.	*Target 2: women over age 30 and those with three or more children should plan to have no more children and should be offered the most effective means of fertility regulation.* *Target 5: communities should be assisted in assuming responsibility for care of the aged, to reduce the desire for many children for security in old age.*
20 N = 27	Women who have low, low to moderate or moderate to high levels of education, who are also older than 36 years of age, and whose media exposure is good and whose wife's religion is not Islamic have 60% probability of using	*Target 4: in areas with higher rates of contraceptive use, education, basic health services and income generating activities are needed to institutionalize the social benefits of family planning.*

	contraception and 40% probability of not using contraception.	
21 N = 135	Women who have low, low to moderate or moderate to high levels of education, who are also older than 36 years of age, and whose media exposure is good and whose wife's religion is Islamic have a 70% probability of not using contraception and a 30% probability of using contraception.	*Target 2: Second, women over age 30 and those with three or more children should plan to have no more children and should be offered the most effective means of fertility regulation.* *Target 5: communities should be assisted in assuming responsibility for care of the aged, to reduce the desire for many children for security in old age.*

In conclusion, the NFPP has made positive changes towards their goals in several areas, specifically with offering means of fertility regulation (Target 2) and the social benefits of family planning (Target 4) specially related to women with lower levels of education who are older. However, there is still work to be done in educating (even the highly educated women) on birth planning (Target 1) and to older women of Islamic faith who need assistance in reducing the desire to have children for security in old age.

Limitations

Several limitations were found while conducting the analysis. Related to the data; the target variable using three levels of contraception choice produced an unsatisfactory model. It is also possible that other categorical variables within the dataset could have been categorized in a more meaningful manner. Specifically, levels of education or husband's occupation which all

have 4 levels. A related limitation to the categorization of variables, was the lack of information on some of the variables making it hard to interpret. For example, extensive research was done to try and understand the 1 to 4 levels of education yet, there was only national statistics on a scale that did not relate to the 1-4 scale in the dataset.

Limitations were also found in the results of the model. For instance, Node 3 node 11 and Node 18 all had perfect predictions for the class of no use of contraception. However, these results make little sense especially as in node 3 and 11 the women had no children! Results would certainly have made sense should they have revealed that they *did* use contraception. It should be noted that all three of these nodes also had the lowest number of instances per node. Node 3 had 26, node 11 had 40 and node 18 included just 21 women from the dataset. Limitation of the size of the data may have affected the model results.

As aspects of the model do not make practical sense then should this also limits the advantage of conducting a classification tree. One of the advantages of ctrees are the ability for them to be easily understood, especially for those without an analytical background (e.g. other executives). If a model does not make intuitive sense in one area it could call into question the entire model. Others may ask: "If it does not make sense that women without children also do not use contraception then how can we believe anything else you are telling us?" As data analyst, it is critically important to be able to effectively communicate our results, and unfortunately, I would have a hard time feeling confident in this model due to the impractical nature of some of the outcomes. In my opinion therefore, this is the biggest limitation of this model.

There are, of course, possible reasons why women who fit this node of no children and no use of contraception may be legitimate. For example, they may be waiting to get married to have children, or may be very young, or may be infertile or may choose to abstain from sex. However,

these reasons can be partially eliminated as all the women in the dataset were at an age where they could have become pregnant and therefore, it seems legitimate that there would be at least some of the women who would choose contraception, especially in light of the NFPP education programs.

The limitation of explaining the practical results of the model output is a class predicament between the accurate prediction and the why the prediction was made (Bati, 2021b). In other words, according to the model output we know a perfect prediction, however when we ask "Why?" the answer is unclear and contrary to what is practically known, which is that contraception prevents pregnancies. Logically therefore, women without children would be using contraception.

In a related fashion, the model output showed 40-50% predictability for Node 13 and 20. This range of predictability is low, again making the result less sure should a new woman enter the dataset and fall within those nodes.

These limitations are common in decision tree models that include many factors, as it is difficult to achieve a consistent message that can be delivered in a human-friendly explanation (Bati, 2021b). This problem can quickly escalate into full blown refusal to believe any output from a "machine" and the humans to trust the system is eroded as the truthfulness of the output is questioned.

It is well understood that decision trees are often relatively inaccurate (Deng, Runger & Tuv, 2011). Furthermore, decision trees are unstable, meaning that a small change in the data can lead to a large change in the structure of the optimal decision tree (Utgott, 1989).

For all the reasons listed this brings the generalizability of the model into question. Even

if the model generalizes from training to test data, if the model itself is unstable, or inconsistent, then it can become difficult to generalize should, even the smallest change, need to occur to the model over time. Often, changes to model are inevitable.

Improvement Areas

This data was collected more than thirty years ago, should the survey be redone, adding more input variables and more observations would improve the possibility of a more meaningful outcome. This is especially the case where the nodes predict perfectly yet make little intuitive sense.

Specific recommendations for additional independent variables include combining the data with birth rates and health related variables of the mother, for instance weight, height, body mass index. Similar health variables of the husband would also be helpful.

Geospatial location, even if only rural or urban would be an interesting addition to the dataset. Especially for the NFPP as the access to centers and information on family planning, which may not be available in all parts of Indonesia could influence future decisions about where the centers are located.

Given the data retrieved, and questionable results in some nodes, it may be worth exploring other analytical approaches. One example may be a random forest method for classification.

It is recommended that combining this data with other data from the DHS collected at the time and using a random forest method be used. Random forest method of analysis is another classification technique that constructs multiple decision trees (Piryonesi, Madeh; El-Diraby, Tamer, 2020). Random forests generally perform better that ctree however, their accuracy is lower than gradient boosted trees (Piryonesi, Madeh; El-Diraby, Tamer, 2021). The disadvantage

of this method is that they can be more difficult to interpret, nevertheless, interpretation of difficult material can be assisted by visualizations which can make the message easier for non-analytics persons to understand. In my opinion, this is less problematic than having a decision tree that is easier to understand but does not make intuitive or practical sense.

Appendix A

Textual Output of the Training Data

```
> # b. Model is a variable where model object is stored
> model<-ctree(ContraceptiveMethod~., train.data)
> # Output the model
> print(model)

        Conditional inference tree with 11 terminal nodes

Response:  ContraceptiveMethod
Inputs:  WifeAge, WifeEducation, HusbandEducation, NumChildren, WifeReligion, WifeWorking, HusbandOccupation, LivingStandardIndex, MediaExposure
Number of observations:  1044

1) WifeEducation == {high}; criterion = 1, statistic = 52.881
  2) NumChildren <= 0; criterion = 1, statistic = 36.8
    3)*  weights = 26
  2) NumChildren > 0
    4) NumChildren <= 2; criterion = 0.997, statistic = 12.735
      5) WifeAge <= 34; criterion = 1, statistic = 32.821
        6)*  weights = 132
      5) WifeAge > 34
        7)*  weights = 33
    4) NumChildren > 2
      8)*  weights = 218
1) WifeEducation == {Low, Low-mod, mod-high}
  9) WifeAge <= 36; criterion = 0.999, statistic = 15.777
    10) NumChildren <= 0; criterion = 1, statistic = 60.049
      11)*  weights = 40
    10) NumChildren > 0
      12) NumChildren <= 2; criterion = 1, statistic = 23.648
        13)*  weights = 176
      12) NumChildren > 2
        14)*  weights = 223
  9) WifeAge > 36
    15) MediaExposure == {Not Good}; criterion = 0.995, statistic = 13.395
      16) LivingStandardIndex == {Low, high}; criterion = 0.987, statistic = 15.469
        17)*  weights = 13
      16) LivingStandardIndex == {Low-mod, mod-high}
        18)*  weights = 21
    15) MediaExposure == {Good}
      19) WifeReligion == {Non-Islamic}; criterion = 0.952, statistic = 7.708
        20)*  weights = 27
      19) WifeReligion == {Islamic}
        21)*  weights = 135
>
```

References

Bati, F. (2021). Lecture: *Classification.* University of Maryland University College. Data 630: Module 4.

Bati, F. (2021b). Lecture: *Model Explainability.* University of Maryland University College. Data 630: Module 4.

Biro Pusat Statistik (BPS). (1975). Penduduk Indonesia: Hasil Sensus Penduduk 1971 (*Population of Indonesia: Results of the 1971 Population Census*). Jakarta: Biro Pusat Statistik.

Biro Pusat Statistik (BPS). (1983). Penduduk Indonesia: Hasil Sensus Penduduk 1980 (*Population of Indonesia: Results of the 1980 Population Census*). Jakarta: Biro Pusat Statistik.

Biro Pusat Statistik (BPS). (1987). Penduduk Indonesia: Series SUPAS No. 5 (*Population of Indonesia: Results of the 1985 Intercensal Population Survey*). Jakarta: Biro Pusat Statistik.

Deng, H., Runger, G., Tuv, E. (2011). *Bias of importance measures for multi-valued attributes and solutions*. Proceedings of the 21st International Conference on Artificial Neural Networks (ICANN).

Gertler, P.J. & Molyneaux, J.W. (1994). How economic development and family planning programs combined to reduce Indonesian fertility. *Demography, 31*(1): 33-63.

Healthline (2020). *When can you get Pregnant and What's the Best Age to have a Baby?* Retrieved from: https://www.healthline.com/health/womens-health/childbearing-

age#:~:text=What's%20the%20childbearing%20age%3F,between%20ages%2012%20and%2051

Horthorn, T. Hornik, K. & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical Statistics, 15 (3); p. 651-674. DOI: https://doi.org/10.1198/106186006X133933

Indonesian National Health System Report (1987). *National Indonesia Contraceptive Prevalence Survey. Summary Report.* Retrieved from: https://dhsprogram.com/pubs/pdf/SR9/SR9.pdf

James, G., Whitten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning: With Application in R.* Springer, New York, NY.

Piryonesi S. Madeh; El-Diraby Tamer E. (2020). "Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems". *Journal of Transportation Engineering, Part B: Pavements*. *146* (2) doi:10.1061/JPEODX.0000175

Piryonesi, S. Madeh; El-Diraby, Tamer E. (2021). "Using Machine Learning to Examine Impact of Type of Performance Indicator on Flexible Pavement Deterioration Modeling". *Journal of Infrastructure Systems*. **27** (2) doi:10.1061/(ASCE)IS.1943-555X.0000602

Ng, A. (2021). *Introduction to Supervised Learning.* Retrieved from: https://www.coursera.org/lecture/machine-learning/supervised-learning-1VkCb

RAND Corporation. (1997). *Indonesian Living Standards Before and After the Financial Crisis.* RAND Corporation. Retrieved from: https://www.rand.org/content/dam/rand/pubs/monographs/2004/RAND_MG137.pdf

Sauver, J. L., Grossardt, B. R., Leibson, C. L., Yawn, B. P., Melton, L. J., 3rd, & Rocca, W. A. (2012). Generalizability of epidemiological findings and public health decisions: an illustration from the Rochester Epidemiology Project. *Mayo Clinic proceedings*, *87*(2), 151–160. https://doi.org/10.1016/j.mayocp.2011.11.009

Statistica (2010). Share of Indonesian Population in 2010, by Religion. Retrieved from: https://www.statista.com/statistics/1113891/indonesia-share-of-population-by-religion/

UCI Machine Learning Repository (1997). *Contraceptive Method Choice Data Set*. Retrieved from: https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice

Toossi, M. & Morisi, T.L. (2017). Women in the workforce before, during and after the Great Depression. *U.S. Bureau of Labor Statistics.* Retrieved from: https://www.bls.gov/spotlight/2017/women-in-the-workforce-before-during-and-after-the-great-recession/pdf/women-in-the-workforce-before-during-and-after-the-great-recession.pdf

Vallantin, L. (2020). *Why you Should not Trust only in Accuracy to Measure Machine Learning Performance.* Retrieved from: https://medium.com/@limavallantin/why-you-should-not-trust-only-in-accuracy-to-measure-machine-learning-performance-a72cf00b4516

Utgoff, P. E. (1989). Incremental induction of decision trees. *Machine learning, 4(2),* 161–186. doi:10.1023/A:1022699900025

Wright, L., & Tellei, V. (1993). Women in Management in Indonesia. *International Studies of Management & Organization, 23*(4), 19-45. Retrieved July 5, 2021, from http://www.jstor.org/stable/40397258