Assignment 3 - Decision Tree Analysis on Contraceptive Method Choice

Tanushree Kumar

DATA 630 - Summer 2024

Professor Ami Gates

University of Maryland Global Campus

Due: July 2, 2024

Introduction

The objective of this decision tree analysis is to accurately predict the target class, and the current method of contraception, of women from Indonesia based on different demographic and socio-economic factors. The objective of this analysis is to build a decision tree model to classify and predict the contraceptive method choice among women based on various demographic and socio-economic factors. Specifically, we aim to answer the question: "What factors influence the choice of contraceptive methods among women?" and "How accurately can these factors predict the contraceptive method chosen?".

This dataset is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey facilitated by the Demographic and Health Survey (DHS). The survey was aimed at married women and collected information on various aspects of their lives such as their background; reproduction; knowledge and practice of family planning; their husband's background; etc (*Indonesia - National Contraceptive Prevalence Survey 1987 | Study Description*, 2013).

Contraceptive method choice is a significant aspect of public health, family planning, and women's health. Understanding the factors that influence women's choices can help in designing better family planning programs, policies, and educational campaigns. In many countries, a variety of contraceptive methods are available, including long-term methods (such as IUDs and sterilization) and short-term methods (such as pills and condoms). The choice of method can be influenced by multiple factors, including age, education, number of children, and socioeconomic status. Effective family planning can lead to improved health outcomes, economic benefits, and overall societal well-being.

The chosen method for this analysis is a decision tree analysis. A decision tree generates a model in a tree-like manner that shows all of the decisions and their possible outcomes. The decision tree has multiple nodes with each node representing a test applied on an attribute and each branch representing an outcome of that test. As such, the decision tree algorithm is a powerful and intuitive classification tool that can handle complex datasets with multiple variables. It is particularly suitable for this analysis because it can easily interpret the relationships between various predictors, demographic and socio-economic factors, and the target variable, contraceptive method choice. Decision trees are non-parametric, meaning they do not assume any specific distribution of the data, making them flexible and widely applicable.

Analysis

The dataset used in this analysis is the Contraceptive Method Choice dataset available from the UCI Machine Learning Repository (Lim, 1997). The dataset is comprised of 1473 observations spread across 10 variables including the target variable:

- WifeAge: Wife's age; numerical
- WifeEducation: Wife's education; categorical; 1 = low, 2, 3, 4 = high
- HusbandEducation: Husband's education; categorical; 1 = low, 2, 3, 4 = high
- NumChildren: Number of children ever born; numerical
- WifeReligion: Wife's religion; binary; 0 = Non-Islam, 1 = Islam
- WifeWorking: Is wife currently working; binary; 0 = Yes, 1 = No
- HusbandOccupation: Husband's occupation; categorical; 1, 2, 3, 4
- LivingStandardIndex: Standard of living index; categorical; 1 = low, 2, 3, 4 = high
- MediaExposure: Exposure to media; binary; 0 = Good, 1 = Not Good
- ContraceptiveMethod: Contraceptive method used; class attribute; 1 = No-use, 2 = Long-term, 3 = Short-term

The original questionnaire provided information on what the numbers for the categorical variables meant (*Indonesia - National Contraceptive Prevalence Survey 1987 | Documentation*, 2013). For both, ‘WifeEducation’ and ‘HusbandEducation’, the numbers correspond to the same schooling options: 1 = Primary School; 2 = Junior High; 3 = Senior High School; 4 = Academy. The numbers for ‘HusbandOccupation’ correspond to: 1 = Professional, Technical, and Clerical; 2 = Sales, Services; 3 = Manual; 4 = Agriculture. The ‘LivingStandardIndex’ was likely evaluated on several factors such as the source of drinking water; source of water for other tasks such as laundry; access to what kind of toilet facility; access to electricity; access to a vehicle; and the material of the floor of their house. The ‘MediaExposure’ was likely evaluated on being able to read a newspaper and the frequency of reading a newspaper, watching television, and listening to a radio.

It should also be noted that in the original dataset, the method of contraception was split into modern methods and traditional methods (Central Bureau of Statistics et al., 1998). Although there is no explicit mention of which contraception methods fall under which temporal level, it can be inferred through just reasoning that modern methods such as [birth control] pills, IUDs, and injections would be considered long-term. Whereas the short-term methods would include using withdrawal, condoms, diaphragm, foam, and jelly.

Before developing the model, exploratory data analysis (EDA) was performed to understand the distribution and relationships of the variables. Based on the EDA, data preprocessing was performed such as handling missing values, converting categorical variables into factors, and scaling numerical features ensuring the data is clean and in an appropriate format for decision tree analysis. Using functions such as summary(), str(), and sum(is.na()) gave an overview of the dataset and showed that no missing values were present. The dataset description website also stated no missing values for any columns. However, the target variable 'ContraceptionMethod' was read as numeric instead of a factor. A command was run to convert its attribute type to a factor variable with three levels. The summary command was run again after the conversion to validate that the target variable is indeed a factor variable. All the variables were either numerical or integer variables which meant little data preprocessing had to be performed.

The summary statistics were examined to see the distribution of the data. The wife's age ranged from 16 to 49 years with a mean age of 32.54 years. The wife's education ranged from 1 being low to 4 being high, with a mean of 2.959. The husband's education had the same range but with a mean of 3.43. It was also interesting to observe the quartiles for the wife and husband's education as the husbands seemed more educated than the wives. The quartiles for the husband's education were: 1st (3), Median (4), 3rd (4); whereas the quartiles for the wife's education were: 1st (2), Median (3), 3rd (4). The number of children ranged from 0 to 16 with a mean of 3.26, more appropriately rounded to 3 children. The wife's religion was a binary integer with 0 being Non-Islam and 1 being Islam. The mean was 0.8506 with the quartiles being 1st (1), Median (1), and 3rd (1), indicating that most wives are Muslims. The range for if the wife is working is also binary with 0 being Yes and 1 being No. The mean was 0.7495 with the quartiles being 1st (0), Median (1), and 3rd (1), indicating that a majority of the wives are not working. The standard for living index ranges from 1 being low to 4 being high. It has a mean of 3.134 and quartiles of 1st (3), Median (3), and 3rd (4), indicating that a majority of the people have an adequate standard of living. The range for media exposure is also binary with 0 being good and 1 being not good. The mean was 0.074 with the quartiles being 1st (0), Median (0), and 3rd (0) indicating that most people have good media exposure. The target variable of the contraception method has three levels 1 being no-use, 2 being long-term, and 3 being short-term. The counts were 629 for no-use, 333 for long-term, and 511 for short-term.

Overall, the dataset is well-structured and contains a mix of numerical and categorical variables. The majority of the wives in the dataset are Muslims and not working. The data also shows that most observations have good media exposure. The contraceptive method used is well-distributed among the three levels, with "No-use" being the most frequent.

The decision tree algorithm recursively splits the data based on the feature that provides the highest information gain or the most significant reduction in impurity. The splitting continues until a stopping criterion is met, such as a maximum depth or minimum number of samples per leaf. The key input parameters include: ‘max_depth’ which is the maximum depth of the tree; ‘min_samples_split’ which is the minimum number of samples required to split an internal node; and ‘criterion’ which is the measure used to evaluate the quality of a split. Experimentation can also be done using different values of ‘max_depth’, ‘min_samples_split’, and other parameters to optimize the model’s performance. Cross-validation can also be used to assess the model’s generalizability and avoid overfitting.

To fit the model, the data was split into a training set of 70% and a testing set of 30%. The training set was used to build the model while the test set was used to evaluate the accuracy of the model. A seed value was set which ensured the results would be reproducible when the model is run again. The model was built using two commands with the first being to create the formula and the second being the builder. The formula was made with ‘ContraceptionMethod’ being the target variable and the remaining nine variables being independent variables. Printing the model shows the results which include the number of leaf nodes, the dependent and independent variables, the number of observations in the training set, and the variables used for the splits in the tree. The tree structure can be visualized by plotting a graph along with a simplified version. As a final step, confusion matrices will be made to check the classification accuracy for the training data and to evaluate the model based on the test data.

Results

The figure below shows the results of the decision tree model that was run. This figure can also be seen in Figure 1 in Appendix A. The results show that the tree has 11 terminal nodes and ‘WifeEducation’ is the first splitting attribute. If the value is less than or equal to 3, the next splitting attribute is ‘WifeAge’. If ‘WifeEducation > 3’, the next splitting attribute is ‘NumChildren’.

The second splitting attribute is ‘WifeAge’ if ‘WifeEducation <= 3’. If ‘WifeAge <= 37’, the next splitting attribute is ‘NumChildren’. If ‘WifeAge > 37’, the next splitting attribute is ‘LivingStandardIndex’.

```
> print(cmc_tree)

	 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 <= 3; criterion = 1, statistic = 78.212
  2) WifeAge <= 37; criterion = 1, statistic = 30.13
    3) NumChildren <= 0; criterion = 1, statistic = 56.418
      4)*  weights = 42
    3) NumChildren > 0
      5) NumChildren <= 2; criterion = 1, statistic = 20.749
        6) WifeAge <= 29; criterion = 0.994, statistic = 14.713
          7)*  weights = 143
        6) WifeAge > 29
          8)*  weights = 36
      5) NumChildren > 2
        9)*  weights = 237
  2) WifeAge > 37
    10) LivingStandardIndex <= 3; criterion = 0.985, statistic = 12.782
      11)*  weights = 116
    10) LivingStandardIndex > 3
      12) WifeWorking <= 0; criterion = 0.999, statistic = 18.267
        13)*  weights = 13
      12) WifeWorking > 0
        14)*  weights = 48
1) WifeEducation > 3
  15) NumChildren <= 0; criterion = 1, statistic = 40.75
    16)*  weights = 26
  15) NumChildren > 0
    17) WifeAge <= 38; criterion = 0.999, statistic = 18.187
      18)*  weights = 269
    17) WifeAge > 38
      19) NumChildren <= 2; criterion = 0.989, statistic = 13.395
        20)*  weights = 19
      19) NumChildren > 2
        21)*  weights = 95
```

The second splitting attribute is ‘NumChildren’ if ‘WifeEducation > 3’. If ‘NumChildren <= 0’, this leads to terminal node 16. If ‘NumChildren > 0’, the next splitting attribute is ‘WifeAge’: if ‘WifeAge <= 38’, this leads to terminal node 18; and if ‘WifeAge > 38’, the next splitting attribute is ‘NumChildren’. If ‘NumChildren <= 2’, this leads to terminal node 20. If ‘NumChildren > 2’, this leads to terminal node 21.

The third splitting attribute is ‘NumChildren’ if ‘WifeEducation <= 3’ and ‘WifeAge <= 37’. If ‘NumChildren <= 0’, this leads to terminal node 4. If ‘NumChildren > 0’, the next splitting attribute is again NumChildren. If ‘NumChildren <= 2’, the next splitting attribute is ‘WifeAge’: if ‘WifeAge <= 29’, this leads to terminal node 7; and if ‘WifeAge > 29’, this leads to terminal node 8. Lastly, if ‘NumChildren > 2’, this leads to terminal node 9.

The third splitting attribute is ‘LivingStandardIndex’ if ‘WifeEducation <= 3’ and ‘WifeAge > 37’. If ‘LivingStandardIndex <= 3’, this leads to terminal node 11. If ‘LivingStandardIndex > 3’, the next splitting attribute is WifeWorking: if ‘WifeWorking <= 0’, this leads to terminal node 13; and if ‘WifeWorking > 0’, this leads to terminal node 14.

The decision tree structure can be examined and explained in a simple way below: first splitting attribute: ‘WifeEducation’

- ‘WifeEducation <= 3’:
    - ‘WifeAge <= 37’:
        - ‘NumChildren <= 0’: Leads to terminal node 4.
        - ‘NumChildren > 0’:
            - ‘NumChildren <= 2’:
                - ‘WifeAge <= 29’: Leads to terminal node 7.
                - ‘WifeAge > 29’: Leads to terminal node 8.
            - ‘NumChildren > 2’: Leads to terminal node 9.
    - WifeAge > 37:
        - ‘LivingStandardIndex <= 3’: Leads to terminal node 11.

        - ‘LivingStandardIndex > 3’:
            - ‘WifeWorking <= 0’: Leads to terminal node 13.
            - ‘WifeWorking > 0’: Leads to terminal node 14.
- WifeEducation > 3:
    - ‘NumChildren <= 0’: Leads to terminal node 16.
    - ‘NumChildren > 0’:
        - ‘WifeAge <= 38’: Leads to terminal node 18.
        - ‘WifeAge > 38’:
            - ‘NumChildren <= 2’: Leads to terminal node 20.
            - ‘NumChildren > 2’: Leads to terminal node 21.

Terminal nodes are the final nodes where predictions are made. Each terminal node represents a unique subset of the population with similar characteristics, and the decision tree captures the hierarchical relationship among these variables. The decision tree helps to understand how different factors like the wife's education, age, number of children, living standard index, and wife’s working status influence the choice of contraceptive method. For example, women with lower education levels, ‘WifeEducation <= 3’, and younger ages ‘WifeAge <= 37’ with no children are predicted to belong to a certain contraceptive method category of terminal node 4. Whereas women with higher education levels, ‘WifeEducation > 3’, and with children are predicted to choose a different contraceptive method based on their age and the number of children they have, which leads to terminal nodes 18, 20, and 21. Table 1 in Appendix A contains all of the terminal nodes and the paths leading to them. A snippet of Table 1 can be seen below.

Table 1

Table of the terminal nodes breakdown of the decision tree.

Terminal Node	Path
4	‘WifeEducation <= 3’, ‘WifeAge <= 37’, and ‘NumChildren <= 0’
7	‘WifeEducation <= 3’, ‘WifeAge <= 37’, ‘NumChildren > 0’, ‘NumChildren <= 2’, and ‘WifeAge <= 29’
8	‘WifeEducation <= 3’, ‘WifeAge <= 37’, ‘NumChildren > 0’, ‘NumChildren <= 2’, and ‘WifeAge > 29’

The partial tree structure was printed using the nodes function. This function took the model name and the starting node number as the input, which can be seen in the figure below. This figure can also be seen in Figure 2 in Appendix A. Here, the tree starts from node 2. This partial tree is significantly smaller since node siblings are excluded.

```
> nodes(cmc_tree, 2)
[[1]]
2) WifeAge <= 37; criterion = 1, statistic = 30.13
  3) NumChildren <= 0; criterion = 1, statistic = 56.418
    4)*  weights = 42
  3) NumChildren > 0
    5) NumChildren <= 2; criterion = 1, statistic = 20.749
      6) WifeAge <= 29; criterion = 0.994, statistic = 14.713
        7)*  weights = 143
      6) WifeAge > 29
        8)*  weights = 36
    5) NumChildren > 2
      9)*  weights = 237
2) WifeAge > 37
  10) LivingStandardIndex <= 3; criterion = 0.985, statistic = 12.782
    11)*  weights = 116
  10) LivingStandardIndex > 3
    12) WifeWorking <= 0; criterion = 0.999, statistic = 18.267
      13)*  weights = 13
    12) WifeWorking > 0
      14)*  weights = 48
```

This partial decision tree represents the decisions made based on the attributes ‘WifeAge’, ‘NumChildren’, ‘LivingStandardIndex’, and ‘WifeWorking’ to predict the ‘ContraceptiveMethod’. The first splitting attribute is ‘WifeAge’ where if ‘WifeAge <= 37’, the next splitting attribute is NumChildren; and if ‘WifeAge > 37’, the next splitting attribute is ‘LivingStandardIndex’.

The second splitting attribute is ‘NumChildren’ if ‘WifeAge <= 37’. If ‘NumChildren <= 0’, it leads to terminal node 4. If ‘NumChildren > 0’, ‘NumChildren <= 2’, the next splitting attribute is ‘WifeAge’. If ‘WifeAge <= 29’ it leads to terminal node 7. If ‘WifeAge > 29’ it leads to terminal node 8. If ‘NumChildren > 2’, it leads to terminal node 9.

The second splitting attribute is ‘LivingStandardIndex’ if ‘WifeAge > 37’. If ‘LivingStandardIndex <= 3’, it leads to terminal node 11. If ‘LivingStandardIndex > 3’, the next splitting attribute is ‘WifeWorking’. If ‘WifeWorking <= 0’, it leads to terminal node 13. If ‘WifeWorking > 0’, it leads to terminal node 14.

This partial decision tree helps to understand how different factors like the wife's age, number of children, living standard index, and wife’s working status influence the choice of contraceptive method. For example, when ‘WifeAge <= 37’, if the number of children is 0, they are predicted to follow a certain contraceptive method of terminal node 4. If they have 1 or 2 children, the wife's age further determines the method: if ‘WifeAge <= 29’, they fall into terminal node 7; if ‘WifeAge > 29’, they fall into terminal node 8. If they have more than 2 children, they fall into terminal node 9. When ‘WifeAge > 37’, if the living standard index is 3 or less, they are predicted to follow a certain contraceptive method of terminal node 11. If the living standard index is higher than 3, the wife's working status further determines the method: if ‘WifeWorking <= 0’, they fall into terminal node 13; if ‘WifeWorking > 0’, they fall into terminal node 14. Table 2 in Appendix A contains all of the terminal nodes and the paths leading to them for this partial tree.

The graph visualization for the decision tree can be seen in Figure 3 in Appendix A. Similarly, a simplified version of the decision tree graph can be seen in Figure 4 in Appendix A.

Using the predict command, a confusion matrix was built to check the classification accuracy of the training data. The confusion matrix for the training data shows how the predicted labels (1, 2, 3) align with the actual labels (1, 2, 3) for the training dataset. The number of correctly classified instances was 610. The total number of instances in the training set was 1044. The classification accuracy of the training model is (610/1044)*100 = 58.43%. A proportion table was also built which computes the probability for each matrix entry. For example, the first entry of 0.28735632 means that 28.74% of the total predictions were correctly classified as class 1. Figure 5 in Appendix A shows the confusion matrix and proportion table for the training data.

The same can be done for the test data. The number of correctly classified instances was 229. The total number of instances in the training set was 429. The classification accuracy of the training model is (229/429)*100 = 53.38%. Figure 6 in Appendix A shows the confusion matrix test data.

The training data's accuracy of 58.43% can be interpreted as the model correctly predicting the contraceptive method for 58.43% of the training data instances. This means that out of all the instances the model saw during training, it correctly identified the contraceptive method for approximately 58% of them. While the model has learned some patterns and relationships in the training data, it is not highly accurate. About 41.57% of the instances were misclassified, indicating that the model may not fully capture the complexities of the factors influencing contraceptive method choice. This level of accuracy suggests that the model is somewhat effective but not robust in predicting the outcomes based on the given features, e.g.: 'WifeAge', 'WifeEducation', and 'NumChildren'.

On the other hand, the test data had an accuracy of 53.38%, which can be interpreted as the model correctly predicting the contraceptive method for 53.38% of the test data instances, which is lower than the training accuracy. The drop in accuracy from the training data to the test data suggests that the model may not generalize well to new data. This decrease can indicate potential overfitting, where the model has learned specific patterns in the training data that do not apply to unseen data. In the context of the contraception study, this means that while the model can somewhat predict contraceptive method choice, its performance is less reliable when applied to new individuals.

The accuracies indicate that the decision tree model has moderate capability in predicting the contraceptive method choice. It captures some underlying trends but fails to provide highly accurate predictions. The accuracy rates also reflect how well the model has understood the influence of various factors on contraceptive method choice. For instance, education levels, number of children, and living standards may significantly impact the decision-making process for contraception. For practical applications, such as advising on contraceptive methods in a healthcare setting, the model's moderate accuracy suggests it can provide some insights but should not be solely relied upon. Healthcare professionals might use the model as a supplementary tool rather than a definitive predictor.

The decision tree model shows a reasonable performance but also highlights some challenges in correctly classifying the different contraceptive methods. The confusion matrices indicate that while the model can identify patterns in the data, it struggles with distinguishing between similar classes. Further refinement, such as pruning, parameter tuning, or using more complex models, might improve the classification accuracy.

Conclusion

This decision tree analysis aimed to predict the current contraceptive method used by women in Indonesia based on various demographic and socio-economic factors. The study found that factors such as the wife's education, age, number of children, living standard, and employment status significantly influence contraceptive method choice. Specifically, women with higher education levels and certain age groups tend to choose different methods compared to those with lower education levels and different age groups.

However, several limitations were noted in the analysis. Firstly, the dataset may not fully capture all relevant factors influencing contraceptive choices, such as cultural or personal preferences. Additionally, the moderate accuracy of the predictions indicates that while some patterns were identified, the model may not comprehensively reflect the complexities of contraceptive decision-making. Lastly, the analysis might have been influenced by the specific demographic and socio-economic context of the dataset, limiting its applicability to other regions or populations.

References

Central Bureau of Statistics, National Family Planning Coordinating Board, & Institute for Resource Development/Westinghouse. (1998). National Indonesia Contraceptive Prevalence Survey 1987 Preliminary Report. In *U.S. Agency for International Development*. https://pdf.usaid.gov/pdf_docs/PNABD430.pdf

Indonesia - National Contraceptive Prevalence Survey 1987 | Documentation. (2013, February 26). The World Bank. https://microdata.worldbank.org/index.php/catalog/1398/related-materials. The webpage contains the link to download the pdf of the original questionnaire used in the survey.

Indonesia - National Contraceptive Prevalence Survey 1987 | Study Description. (2013, February 26). The World Bank. https://microdata.worldbank.org/index.php/catalog/1398

Lim, T.-S. (1997, July 6). *Contraceptive Method Choice*. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/30/contraceptive+method+choice

Appendix A

All figures, tables, and visualizations mentioned in the report can be seen below. The R code is attached as a separate file. The original questionnaire is attached as a separate pdf. .

Figure 1

Figure of the decision tree model results.

```
> print(cmc_tree)

	 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 <= 3; criterion = 1, statistic = 78.212
  2) WifeAge <= 37; criterion = 1, statistic = 30.13
    3) NumChildren <= 0; criterion = 1, statistic = 56.418
      4)*  weights = 42 
    3) NumChildren > 0
      5) NumChildren <= 2; criterion = 1, statistic = 20.749
        6) WifeAge <= 29; criterion = 0.994, statistic = 14.713
          7)*  weights = 143 
        6) WifeAge > 29
          8)*  weights = 36 
      5) NumChildren > 2
        9)*  weights = 237 
  2) WifeAge > 37
    10) LivingStandardIndex <= 3; criterion = 0.985, statistic = 12.782
      11)*  weights = 116 
    10) LivingStandardIndex > 3
      12) WifeWorking <= 0; criterion = 0.999, statistic = 18.267
        13)*  weights = 13 
      12) WifeWorking > 0
        14)*  weights = 48 
1) WifeEducation > 3
  15) NumChildren <= 0; criterion = 1, statistic = 40.75
    16)*  weights = 26 
  15) NumChildren > 0
    17) WifeAge <= 38; criterion = 0.999, statistic = 18.187
      18)*  weights = 269 
    17) WifeAge > 38
      19) NumChildren <= 2; criterion = 0.989, statistic = 13.395
        20)*  weights = 19 
      19) NumChildren > 2
        21)*  weights = 95 
```

Figure 2

Figure of the partial tree structure using the nodes function.

```
> nodes(cmc_tree, 2)
[[1]]
2) WifeAge <= 37; criterion = 1, statistic = 30.13
  3) NumChildren <= 0; criterion = 1, statistic = 56.418
    4)*  weights = 42
  3) NumChildren > 0
    5) NumChildren <= 2; criterion = 1, statistic = 20.749
      6) WifeAge <= 29; criterion = 0.994, statistic = 14.713
        7)*  weights = 143
      6) WifeAge > 29
        8)*  weights = 36
    5) NumChildren > 2
      9)*  weights = 237
2) WifeAge > 37
  10) LivingStandardIndex <= 3; criterion = 0.985, statistic = 12.782
    11)*  weights = 116
  10) LivingStandardIndex > 3
    12) WifeWorking <= 0; criterion = 0.999, statistic = 18.267
      13)*  weights = 13
    12) WifeWorking > 0
      14)*  weights = 48
```

Figure 3

Figure of the decision tree graph.

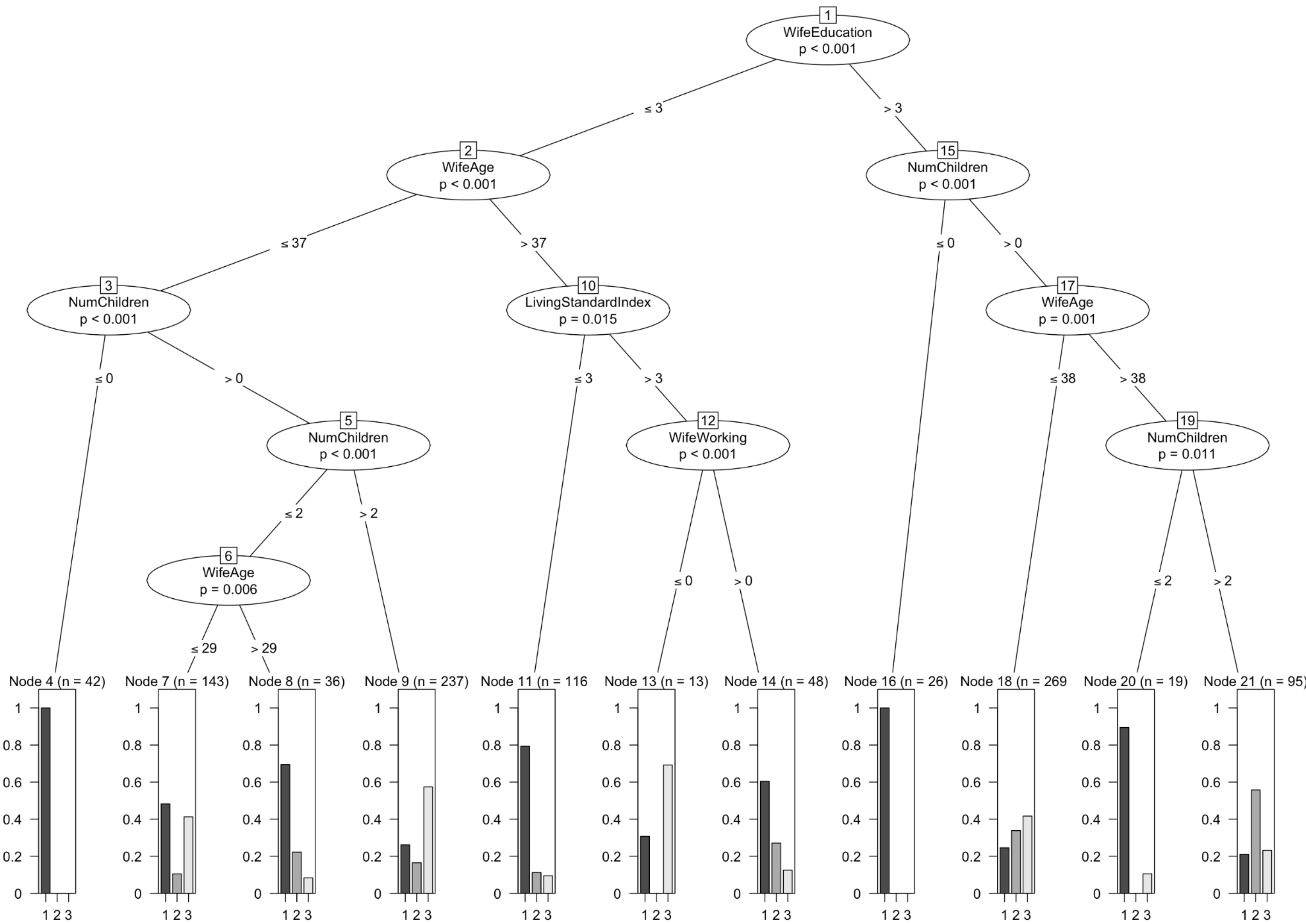

Figure 4

Figure of the simplified decision tree graph.

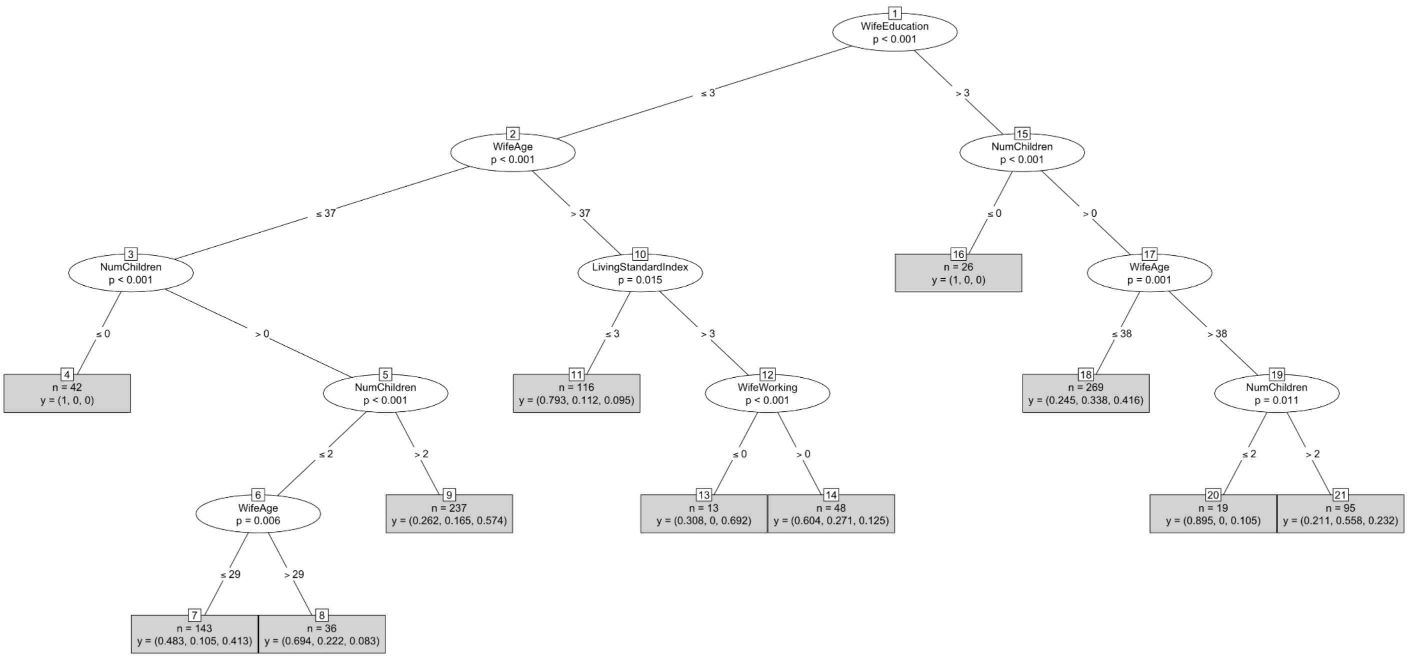

Figure 5

Figure of the confusion matrix for the training data.

```
> # Confusion matrix for the training set
> table(predict(cmc_tree), train_data$ContraceptiveMethod)

      1   2   3
  1 300  49  81
  2  20  53  22
  3 132 130 257
> prop.table(table(predict(cmc_tree), train_data$ContraceptiveMethod))

             1          2          3
  1 0.28735632 0.04693487 0.07758621
  2 0.01915709 0.05076628 0.02107280
  3 0.12643678 0.12452107 0.24616858
```

Figure 6

Figure of the confusion matrix for the test data.

```
> # Evaluating the model based on the test data
> testPred <- predict(cmc_tree, newdata = test_data)
> table (testPred, test_data$ContraceptiveMethod)

testPred   1   2   3
       1 109  21  41
       2   7  17   7
       3  61  63 103
```

Table 1

Table of the terminal nodes breakdown of the decision tree.

Terminal Node	Path
4	‘WifeEducation <= 3’, ‘WifeAge <= 37’, and ‘NumChildren <= 0’
7	‘WifeEducation <= 3’, ‘WifeAge <= 37’, ‘NumChildren > 0’, ‘NumChildren <= 2’, and ‘WifeAge <= 29’
8	‘WifeEducation <= 3’, ‘WifeAge <= 37’, ‘NumChildren > 0’, ‘NumChildren <= 2’, and ‘WifeAge > 29’
9	‘WifeEducation <= 3’, ‘WifeAge <= 37’, ‘NumChildren > 0’, and ‘NumChildren > 2’
11	‘WifeEducation <= 3’, WifeAge <= 37’, ‘NumChildren > 0’, and NumChildren > 2’
13	‘WifeEducation <= 3’, ‘WifeAge > 37’, ‘LivingStandardIndex > 3’, and ‘WifeWorking <= 0’
14	‘WifeEducation <= 3’, ‘WifeAge > 37’, ‘LivingStandardIndex > 3’, and ‘WifeWorking > 0’
16	‘WifeEducation > 3’ and ‘NumChildren <= 0’
18	‘WifeEducation > 3’, ‘NumChildren > 0’, and ‘WifeAge <= 38’

20	'WifeEducation > 3', 'NumChildren > 0', 'WifeAge > 38', and 'NumChildren <= 2'
21	'WifeEducation > 3', 'NumChildren > 0', 'WifeAge > 38', and 'NumChildren > 2'

Table 2

Table of the terminal nodes breakdown of the partial decision tree.

Terminal Node	Path
4	'WifeAge <= 37' and 'NumChildren <= 0'
7	'WifeAge <= 37', 'NumChildren > 0', 'NumChildren <= 2', and 'WifeAge <= 29'
8	'WifeAge <= 37', 'NumChildren > 0', 'NumChildren <= 2', and 'WifeAge > 29'
9	'WifeAge <= 37', 'NumChildren > 0', and 'NumChildren > 2'
11	WifeAge <= 37' and 'LivingStandardIndex <= 3'
13	'WifeAge > 37', 'LivingStandardIndex > 3', and 'WifeWorking <= 0'
14	'WifeAge > 37', 'LivingStandardIndex > 3', and 'WifeWorking > 0'