*Create a title page here*

**Using decision tree to classify the contraceptive method used by women**

**1. Introduction**

Contraception plays a vital role in family planning and women's reproductive health. The choice of contraceptive method is a critical decision for women and plays a significant role in family planning, reproductive health, and overall well-being, influenced by various factors such as age, marital status, education, and cultural beliefs. Understanding the factors that contribute to a woman's choice of contraceptive method is essential for healthcare professionals, policymakers, and researchers to provide better family planning services and support. The primary objective of this study is to classify the contraceptive method used by women based on her demographic and socio-economic status. In essence, we aim to answer the pivotal question: What are the key determinants that influence a woman's choice of contraceptive method? By employing sophisticated analytical techniques, we intend to unearth insights that can assist healthcare providers and researchers in making informed decisions and optimizing pregnancy care strategies.

Contraception is a global concern, but its significance varies across different countries and cultures. In some parts of the world, access to contraception is limited, leading to unintended pregnancies and potentially adverse health outcomes. In others, the choice of contraceptive method is a matter of personal preference, influenced by various socio-demographic factors, cultural norms, and healthcare accessibility.

According to the World Health Organization (WHO), approximately 214 million women of reproductive age in developing countries who wish to avoid pregnancy are not using a modern contraceptive method. Additionally, each year, there are an estimated 74 million unintended pregnancies, many of which result from inadequate access to contraception or insufficient information about available methods.

Given these challenges and the potential impact on women's health and family planning, the need for accurate predictive models to understand and address contraceptive choice is evident. This analysis aims to shed light on the factors affecting contraceptive choices, ultimately contributing to improved healthcare services and policy decisions.

The choice of decision trees as the analytical method for this study is grounded in its suitability for classification tasks and its interpretability. Decision trees are a widely used machine learning technique that excels in making predictions based on structured data, such as socio-demographic and behavioral attributes. They are particularly well-suited for this analysis because they can handle both categorical and numerical data, making them versatile for modeling the diverse set of features that influence contraceptive choices.

By leveraging decision tree modeling, we aim to uncover the most influential factors that contribute to the choice of contraceptive method among women. This can help healthcare providers tailor their counseling and support services, empower women in addressing barriers to contraception access and utilization. In summary, the decision tree approach is a well-founded and practical choice for this analysis, given its ability to provide actionable insights into a complex and important public health issue.

* **Model Fitting**: explain the key steps and activities you perform to fit the model. Experiment (as appropriate) with parameters tuning. This is key, what separates highly accurate model from a less accurate one is the amount of performance tuning performed.

**2.1 Understanding the Data**

The dataset for this was selected from the module’s recommended datasets list. It is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey. The samples are married women who were either not pregnant or do not know if they were at the time of interview. Here is the Definitions of the columns of the data.

**Table 1.0: Description of Data Variables**

|  |  |  |
| --- | --- | --- |
| **Variable Name** | **Code** | **Description** |
| Wife's age | Numerical | Age of the wife |
| Wife's education | Categorical  1=low, 2, 3, 4=high | Education level of wife |
| Husband's education | Categorical  1=low, 2, 3, 4=high | Education level of husband |
| Number of children ever born | Numerical | Number of children ever born by the wife |
| Wife's religion | Binary  0=non-Islam, 1=Islam | Religion of wife |
| Wife's now working | Binary  0=Yes, 1=No | Occupation status of wife |
| Husband's occupation | Categorical  1, 2, 3, 4 | Occupation status of husband |
| Standard-of-living index | Categorical  1=low, 2, 3, 4=high | Living status |
| Media exposure | Binary  0=Good, 1=Not good | Exposure to media/ICT |
| Contraceptive method used | Class attribute  1=No-use, 2=Long-term, 3=Short-term | Choice of contraceptive method by the wife |

**2.2. Exploratory Analysis**

Using the ‘str’ function, it was observed that the variables in the dataset consist of integers. The summary function was used to obtain the minimum and maximum, as well as measures of central tendency (mean, median) and spread (1st and 3rd quartiles) for each of these variables.

Attempt was made to create a new binary variable for the wife age and number of children i.e., ‘WifeAge’ and ‘NumChildren’ based on the condition of the wife's age being greater than 25 and number of children being greater than 4. The frequency table for this new variable to count the occurrences of "Yes" and "No" values is shown below.

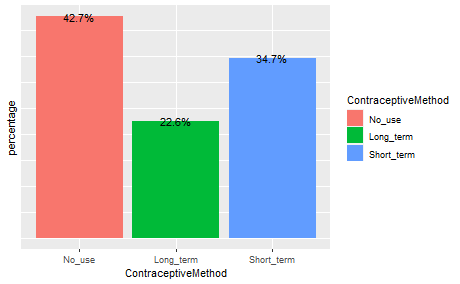
**Table 2.0: Frequency table for Age and Number of Children**

|  |  |  |  |
| --- | --- | --- | --- |
| **WifeAge** | | **NumChildren** | |
| No | Yes | No | Yes |
| 36 | 1117 | 1105 | 368 |

Now, to ensure that all the variables including the target variable are in factor. Encoding was done as described in Table 1.0, column two.

**Variable Analysis**

42.7% of women does not use contraceptive method, 22.6% uses a long term contraceptive method, while 34.7% uses a short term contraceptive method



**Figure 2.1: Bar plot of Contraceptive method**

Older women do not seem to be using contraceptive. Also, women with the more children tend to use a long-term contraceptive.

|  |  |
| --- | --- |
|  |  |
| **Figure 2.2: Box plot of Contraceptive method by Age and Number of children** | |

**2.3. Preprocessing - Check for missing data**

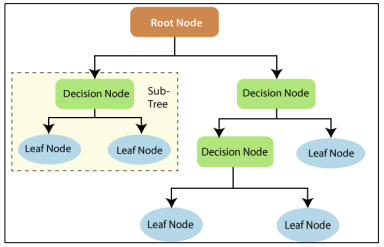
Next, we check how many NA records we have, per column. Fortunately, there are no records of missing data in the dataset as shown below

**Table 3.0: Counts for NAs in the dataset**

|  |  |
| --- | --- |
| **Columns** | **NAs** |
| ContraceptiveMethod | 0 |
| WifeAge | 0 |
| NumChildren | 0 |
| HighWifeAge1 | 0 |
| HighNumChildren1 | 0 |
| HighWifeAge | 0 |
| HighNumChildren | 0 |
| WifeEducation | 0 |
| HusbandEducation | 0 |
| WifeReligion | 0 |
| WifeWorking | 0 |
| HusbandOccupation | 0 |
| LivingStandardIndex | 0 |
| MediaExposure | 0 |

**2.4. Decision tree algorithm**

One of the widely used techniques in data mining is systems that create classifiers. In data mining, classification algorithms are capable of handling a vast volume of information. It can be used to make assumptions regarding categorical class names, to classify knowledge on the basis of training sets and class labels, and to classify newly obtainable data. Classification algorithms in machine learning contain several algorithms, one of which is the decision tree algorithm.



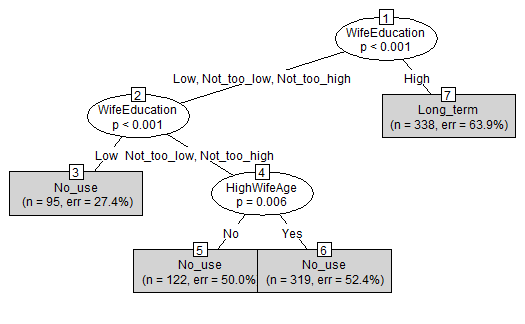
**Figure 2.4: Decision tree algorithm**

**2.5. Model Fitting**

The dataset was divided into training and test set using R code, after which we perform the decision tree model and evaluate the confusion matrix with model methods such as sensitivity, specificity, positive prediction value, negative prediction value, and prevalence of the data.

**3. Results**

**3.1 Output**

****

**Figure 3,0: Decision tree algorithm for Contraceptive method**

From above tree flow, we can have our inferences like the wife age and education level are the most significant variables as they tell us which contraceptive method is mostly used by women.

It shows 4 terminal nodes and a vector representing the proportion of instances in the node representing one of the 3 class values. For example, terminal node 5 states that 122 instances apply to this exact classification. 50% of married women out of the 122 do not use contraceptive, for terminal node 7, 63.9% of women out of 338 use a long term contraceptive

Education, being of the highest significance and our first assumption for the likelihood of contraception use, may have played a part in the level of accuracy of the model. Additionally, it was interesting to start off with the fact that the survey gave no option for a husband to not be holding an occupation. For instance, if the wife was working a job, then we do not know if the husband also was working or was not. This is a factor that we were not able to consider.

**3.2. Model Properties**

We observed 43% accuracy of the model on test data i.e 43% of the test values are correctly classified, and misclassification rate is around 57%. The rate at which there was no information, i.e., "No Information Rate" is 60%.

This p-value is associated with a hypothesis test comparing the model's accuracy to the accuracy achieved by the "No Information Rate." In this case, the p-value is 1, which suggests that there is no statistically significant difference between the model's accuracy and the "No Information Rate."

**3.3. Evaluation of the model**

The confusion matrix, which is often used to evaluate the performance of a classification model is provided below. In a confusion matrix, the rows represent the actual classes or categories, while the columns represent the predicted classes or categories. Each cell in the matrix represents the number of observations that fall into a particular combination of actual and predicted classes. There are 174 instances where the actual class was "No\_use," and the model correctly predicted "No\_use." Also, here are 92 instances where the actual class was "Long\_term," and the model correctly predicted "Long\_term."

**Table 3.0: Confusion Matrix**

|  |  |  |  |
| --- | --- | --- | --- |
|  | *Predict* | | |
|  | No use | Long term | Short term |
| No use | 174 | 69 | 0 |
| Long term | 50 | 92 | 0 |
| Short term | 146 | 82 | 0 |

The model is observed to be 43% accurate, with 47% and 72% Sensitivity and Specificity respectively

|  |  |  |  |
| --- | --- | --- | --- |
|  | **No use** | **Long term** | **Short\_term** |
| Sensitivity | 0.4703 | 0.3786 | NA |
| Specificity | 0.716 | 0.8649 | 0.6281 |
| Pos Pred Value | 0.716 | 0.6479 | NA |
| Neg Pred Value | 0.4703 | 0.6794 | NA |
| Prevalence | 0.6036 | 0.3964 | 0 |
| Detection Rate | 0.2838 | 0.1501 | 0 |
| Detection Prevalence | 0.3964 | 0.2316 | 0.3719 |
| Balanced Accuracy | 0.5932 | 0.6217 | NA |

**4. Conclusion**

**4.1 Summary**

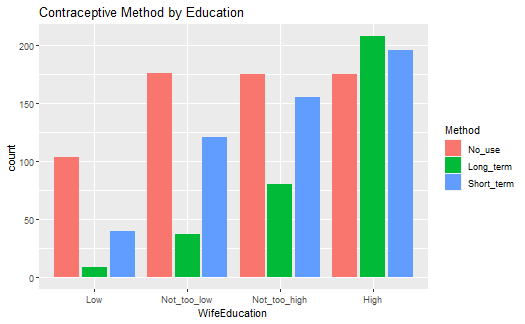
As we see, using the accuracy of decision tree is 43%, and classification error is 57%. So, we can conjecture that using decision tree wouldn’t build a good model as we expect, though reliable. It was deduced that the wife age and education are the most important determinant of contraceptive metho used by women.

**4.2 Limitations of the study**

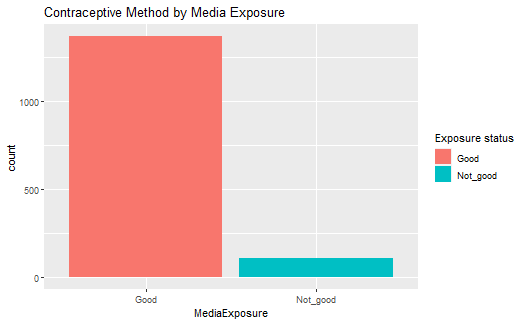
* Limited number of variables in the data**:** This may result in a loss of valuable information that could have contributed to a more comprehensive understanding of the study.
* The values of the attribute named Media Exposure are mostly “Good”, so we cannot estimate whether this attribute works or not.
* It is hard to explain the detail of each attribute because there is no reference about this dissertation. Because of that, we maybe make some misunderstanding to attributes.
* The outcomes change by the period and environment when the dataset is collected. For example, the environment of this dataset is Indonesia and the people there are Muslim. We think the outcomes are seldom the same in Taiwan.

**4.3 Recommendation/Improvement Areas:**

* There should be increased efforts to ensure adequate keeping of information or updating of database of contraceptive method used by women for future analysis.
* Efforts should be made to increase awareness and proper documentation of other factors that may affect pregnancy.



**Figure 2.3: Contraceptive method by Education**



**Figure 2.3: Contraceptive method by Media Exposure**

```{r}

install.packages("rpart", dependencies=TRUE)

install.packages("rpart.plot", dependencies=TRUE)

install.packages("rattle", dependencies=TRUE)

install.packages("party", dependencies=TRUE)

install.packages("partykit", dependencies=TRUE)

```

```{r}

library(corrplot)

library(caTools)

library(tidyverse)

library(data.table)

library(dplyr)

library(party)

library(partykit)

library(tree)

library(MASS)

library(ggplot2)

library(cowplot)

library(rpart)

library(rpart.plot)

library(rattle)

```

```{r}

# Importing the data file

cmc <- read.csv("cmc.csv")

# Converting the data to a dataframe

data <- as.data.frame(cmc)

# Top rows of the data

head(data)

```

# EDA

```{r}

# Summary of the data

summary(data)

```

```{r}

# Understanding the datatype of dataset

str(data)

```

```{r}

# Creating the age flag for and define as WifeAge variable

data = data %>% mutate(HighWifeAge1 = if\_else(WifeAge > 25,"Yes","No"))

table(data$HighWifeAge)

```

```{r}

# Creating the age flag for and define as WifeAge variable

data = data %>% mutate(HighNumChildren1 = if\_else(NumChildren > 4,"Yes","No"))

table(data$HighNumChildren)

```

```{r}

head(data)

```

Also, encoding all categorical variables as factors

```{r}

# Transforming the target and data variables that I will use to submit to the categorical algorithm into factors

ContraceptiveMethod = factor(data$ContraceptiveMethod, levels = c(1,2,3), labels = c("No\_use", "Long\_term", "Short\_term"))

data = data[,-10]

HighWifeAge = factor(data$HighWifeAge1)

WifeEducation = factor(data$WifeEducation, levels = c(1,2,3,4), labels = c("Low", "Not\_too\_low", "Not\_too\_high", "High"))

data = data[,-2]

HusbandEducation = factor(data$HusbandEducation, levels = c(1,2,3,4), labels = c("Low", "Not\_too\_low", "Not\_too\_high", "High"))

data = data[,-2]

HighNumChildren = factor(data$HighNumChildren1)

WifeReligion = factor(data$WifeReligion, levels = c(0,1), labels = c("non\_Islam", "Islam"))

data = data[,-3]

WifeWorking = factor(data$WifeWorking, levels = c(0,1), labels = c("Yes", "No"))

data = data[,-3]

HusbandOccupation = factor(data$HusbandOccupation)

data = data[,-3]

LivingStandardIndex = factor(data$LivingStandardIndex, levels = c(1,2,3,4), labels = c("Low", "Not\_too\_low", "Not\_too\_high", "High"))

data = data[,-3]

MediaExposure = factor(data$MediaExposure, levels = c(0,1), labels = c("Good", "Not\_good"))

data = data[,-3]

```

```{r}

head(data)

```

```{r}

# Adding the newly recoded varaibles to our data

cmcdata <- cbind(ContraceptiveMethod, data, HighWifeAge, HighNumChildren, WifeEducation, HusbandEducation, WifeReligion, WifeWorking, HusbandOccupation, LivingStandardIndex, MediaExposure)

```

```{r}

head(cmcdata)

```

```{r}

# Understanding the new datatype of dataset

str(cmcdata)

```

```{r}

options(repr.plot.width = 4, repr.plot.height = 3)

## ggplot theme

theme <- theme(

axis.text.y = element\_blank(), axis.ticks.y = element\_blank(),

)

cmcdata %>%

group\_by(ContraceptiveMethod) %>%

summarize(

n = n()

) %>%

mutate(

percentage = round(n / sum(n), 3),

n = NULL

) %>%

ggplot(aes(x = ContraceptiveMethod, y = percentage)) + geom\_col(aes(fill = ContraceptiveMethod)) + theme +

geom\_text(

aes(x = ContraceptiveMethod, y = percentage, label = paste(percentage\*100, "%", sep = ""))

)

```

```{r}

## ggplot theme

theme <- theme(

axis.text.y = element\_blank(), axis.ticks.y = element\_blank(),

legend.position="none"

)

# Decrease graph size from standard

options(repr.plot.width = 7, repr.plot.height = 3)

plot\_grid(

data %>%

filter(ContraceptiveMethod == "Yes") %>%

group\_by(WifeAge) %>%

summarize(

n = n()

) %>%

mutate(

Percentage = round(n / sum(n), 3)

) %>%

# Create plot

ggplot(

aes(x = WifeAge, y = Percentage, color = WifeAge)

) +

stat\_smooth(method = "lm", col = "red") +

geom\_point(alpha = 2/3) +

# Clean graph visual a bit

theme +

labs(

x = "WifeAge", y = "ContraceptiveMethod (%)"

),

ggplot(

data = cmcdata,

aes(y = WifeAge, x = ContraceptiveMethod, color = ContraceptiveMethod)

) +

theme +

geom\_boxplot()

, align = "h")

```

```{r}

# Decrease graph size from standard

options(repr.plot.width = 7, repr.plot.height = 3)

plot\_grid(

data %>%

filter(ContraceptiveMethod == "Yes") %>%

group\_by(NumChildren) %>%

summarize(

n = n()

) %>%

mutate(

Percentage = round(n / sum(n), 3)

) %>%

# Create plot

ggplot(

aes(x = NumChildren, y = Percentage, color = NumChildren)

) +

stat\_smooth(method = "lm", col = "red") +

geom\_point(alpha = 2/3) +

# Clean graph visual a bit

theme +

labs(

x = "NumChildren", y = "ContraceptiveMethod (%)"

),

ggplot(

data = cmcdata,

aes(y = NumChildren, x = ContraceptiveMethod, color = ContraceptiveMethod)

) +

theme +

geom\_boxplot()

, align = "h")

```

```{r}

ggplot(cmcdata, aes(WifeEducation, fill = factor(ContraceptiveMethod))) +

geom\_bar(position = "dodge2") +

labs(title = "Contraceptive Method by Education", fill = "Method")

```

```{r}

ggplot(cmcdata, aes(MediaExposure, fill = factor(MediaExposure))) +

geom\_bar(position = "dodge2") +

labs(title = "Contraceptive Method by Media Exposure", fill = "Exposure status")

```

Check for missing data

```{r}

cmcdata %>%

summarise\_all(

funs(sum(is.na(.)))

)

```

```{r}

# Deleting variables for the tree model

cmcdatatree = subset(cmcdata, select = c(-WifeAge, -NumChildren, -HighWifeAge1, -HighNumChildren1))

```

```{r}

head(cmcdatatree)

```

```{r}

# Splitting data into training and testing with 60 and 40 percent respectively

index <- sample(2, nrow(cmcdatatree), replace=TRUE, prob = c(0.60, 0.40))

traindata <- cmcdatatree[index==1, ]

testdata <- cmcdatatree[index==2, ]

```

# Building the ctree classification model

```{r}

treefit <- ctree(ContraceptiveMethod ~., data = traindata)

plot(tree\_fit, type="simple")

```

```{r}

# Making prediction with the above model

predict <- predict(treefit, newdata = testdata[, -1], type="response")

matriz\_conf <- table(testdata[,1], predict)

library(caret)

confusionMatrix(matriz\_conf)

```