Contraceptive Methods in Indonesia

PSTAT 131

Samantha Lee

Professor Sang-Yun Oh

TABLE OF CONTENTS

ABSTRACT

The purpose of this project is to analyze the behavior of parents in Indonesia to further understand the best methods of contraception and improve the “Family Planning Program” implemented in 1976. To achieve this goal, data retrieve from the UCI Machine Learning repository was analyzed using data mining algorithms including decision trees, LDA, QDA, and random forests in R/Rstudio. The primary packages used were MASS, ISLR, randomForest, and ggplot for visualizations. The findings of the research showed that using a random forest method on the data was best to classify the model; however, all methods hovered around 50% accuracy. What has been drawn can be implemented in the “Familiy Planning Program” and help control the population in Indonesia for years to come. This data can be used to advance the quality of life and control the population growth in Indonesia.

INTRODUCTION

Indonesia is the world’s first largest country, with an estimated population of about 260 million people. Between 1976 and 2002, President Suharto implemented the “Family Planning Program”, which resulted in an increase in contraceptive use and a decrease in the fertility rate. The program focused on community education and the distribution of free contraceptives. As the country continues its efforts to manage its population, it is important that they understand the behavior of the nation’s parents.

The data strives to predict the most common method of contraception used by females in Indonesia (whether it is short term, long term, or none at all) and is a classification problem with three levels. The variables that are being predicted on are wife’s age, wife’s education, husband’s education, number of children ever born, wife’s religion, wife’s job status, husband’s occupation, standard-of-living index and exposure to media. The data comes from the UCI Machine Learning Repository and the software used for analysis is R/Rstudio. Packages used include randomForest, ggplot2, MASS and ISLR. The data is collected from surveys taken by married Indonesia women who were not pregnant (or unaware of any pregnancy) at the time of the survey. Despite this data being three decades old, it is still applicable to information collected today.

The study has shown that with statistical learning, we can better understand how to educate the people based on their demographic. The report includes graphical analysis between variables, to see how the response works in accordance with the predictors. The methods analyzed in the report are decision trees, random forest, linear discriminant analysis, quadratic discriminant analysis, and neural networks. The error rate of the methods all ended up being close, averaging about 55%. The use of data mining has proven that the simplest method may often be just as effective as a more complex one.

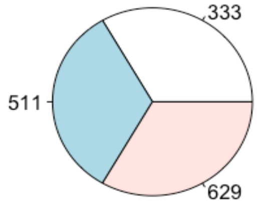
PREPROCESSING

```r
library(ggplot2)
library(dplyr)
library(tidyr)
library(ISLR)
library(reshape2)
library(plyr)
library(party)
library(rpart)
library(caret)
library(MASS)
library(dplyr)
library(class)
library(cluster)
library(randomForest)
library(rattle)
```

```
cmc <- read.table("~/Desktop/finalProject/cmcdata.txt", sep = ",")
#adding column headers
colnames(cmc) <- c("WifeAge", "WifeEducation", "HusbandEducation", "NumChildren",
                "WifeReligion", "IsWifeWorking", "HusbandOccupation", "StandardofLiving","MediaExposure",
"ContraceptiveMethodUsed")
View(cmc)
```

This is a pie chart that outputs the distribution of the variable I am predicting on. There are 629 women who do not use contraception, 333 women who used long term contraceptive methods and 511 women who used short term methods.

```
methodFreq <- table(cmc$ContraceptiveMethodUsed)
methodFreq

##
##   1   2   3
## 629 333 511

pie(table(methodFreq))
```



Now, I start the preprocessing of the data. I converted the variables into something more legible and readible.

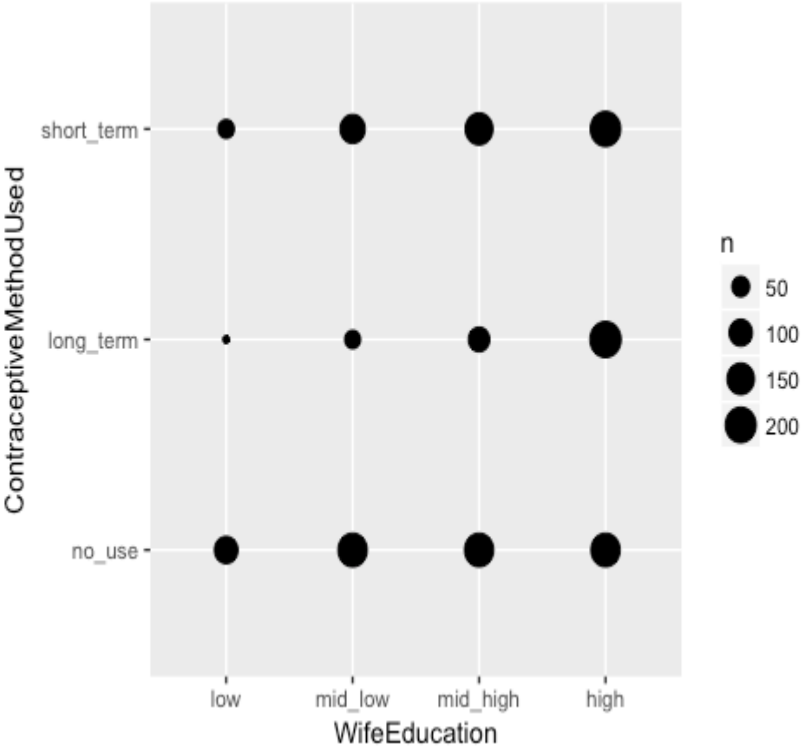
```
#convert binary variables into boolean variables
cmc$WifeReligion <- ifelse(cmc$WifeReligion == 1, TRUE, FALSE)
cmc$IsWifeWorking <- ifelse(cmc$IsWifeWorking == 1, FALSE, TRUE)
cmc$MediaExposure <- ifelse(cmc$MediaExposure == 1, FALSE, TRUE)

#label education as factor
cmc$WifeEducation <- factor(cmc$WifeEducation, labels = c("low", "mid_low", "mid_high", "high"))
cmc$HusbandEducation <- factor(cmc$HusbandEducation, labels = c("low", "mid_low", "mid_high", "high"))

#store husband's occupation and standard of living as a factor
#unsure what the integers actually mean
cmc$HusbandOccupation <- as.factor(cmc$HusbandOccupation)
cmc$StandardofLiving <- factor(cmc$StandardofLiving, labels = c("low", "mid_low", "mid_high", "high"))

#factor the response variable
cmc$ContraceptiveMethodUsed <- factor(cmc$ContraceptiveMethodUsed, labels = c("no_use", "long_term",
"short_term"))
```

```
#check all data types
sapply(cmc[,1:10], FUN = function(x) {class(x)})
```

```
##               WifeAge          WifeEducation       HusbandEducation
##             "integer"               "factor"               "factor"
##           NumChildren            WifeReligion           IsWifeWorking
##             "integer"              "logical"              "logical"
##      HusbandOccupation        StandardofLiving          MediaExposure
##              "factor"               "factor"              "logical"
## ContraceptiveMethodUsed
##              "factor”
```

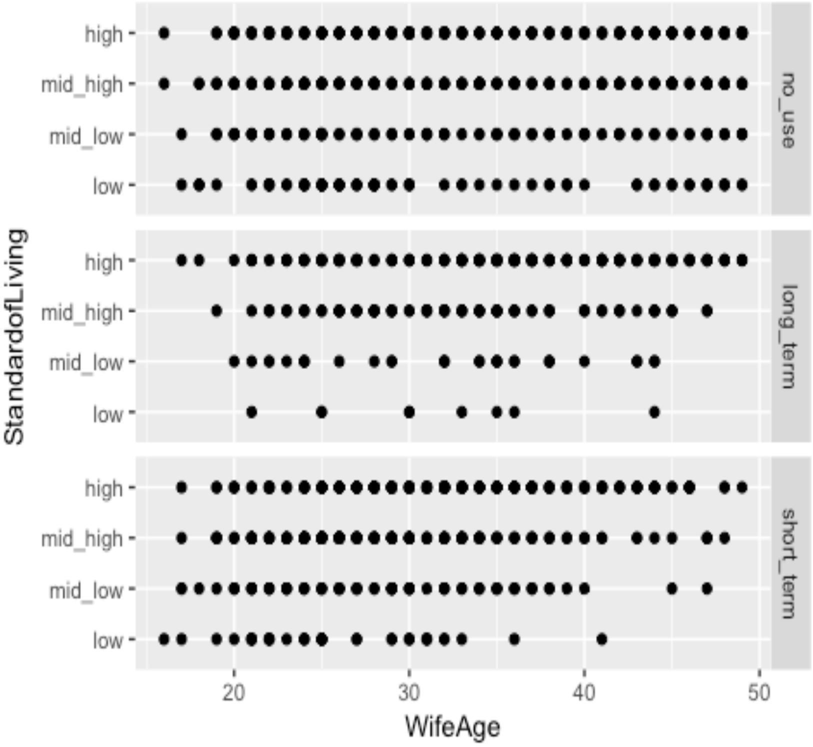
Plot of the wife's education versus the type of contraceptive method used. We can see that with a higher education level, more long and short term contraceptive methods were used. The number of women who did not use contraception at every level of education is fairly evenly distributed.

```
#plot the wife's education vs contraceptive method
ggplot(cmc, aes(x = WifeEducation, y = ContraceptiveMethodUsed)) +
  geom_count() +
  scale_x_discrete(limits =  c("low", "mid_low", "mid_high", "high"))
```

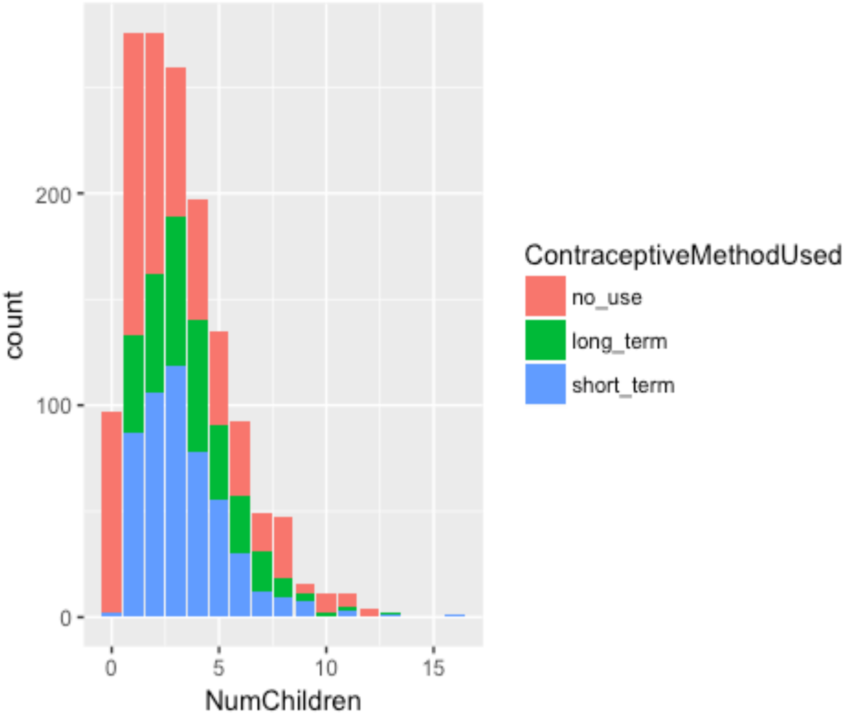


We can see that for no contraceptive used, the standard of living is roughly the same for every age. For the long-term use, standard of living tends to be higher. For short-term use, there is still a skew towards the higher standard of living, but there is more variance.

```
#plot standard of living vs age of the wife against contraceptive method used

qplot(WifeAge, StandardofLiving, data=cmc, facets=ContraceptiveMethodUsed ~.)
```



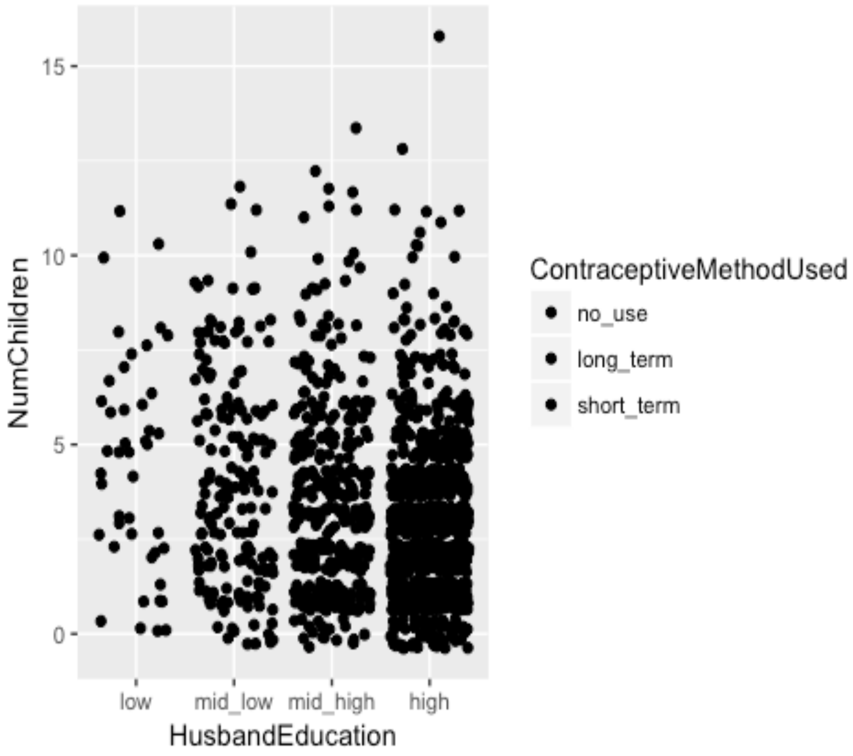
The number of children is typically from 1 - 4. Rarely long term use of contraceptives, primarily short term or none used at all.

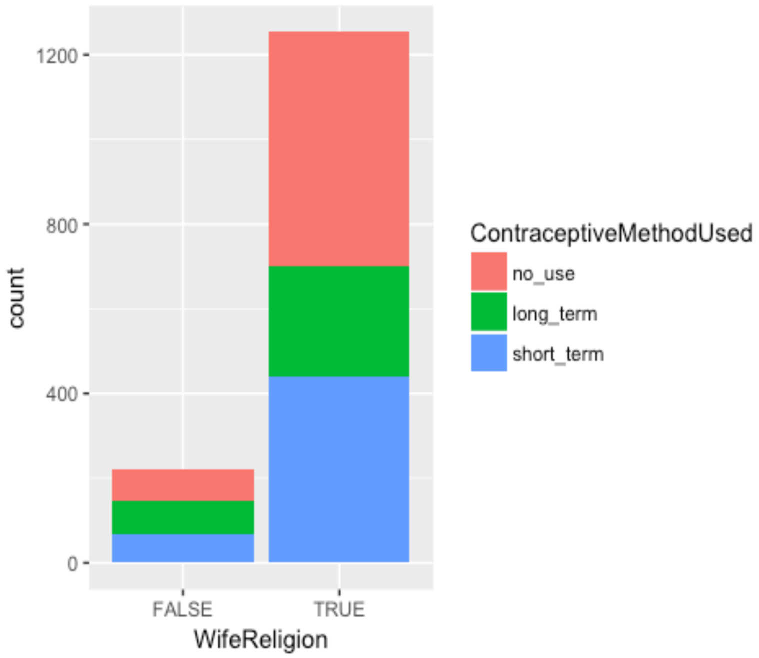


```
#plot contraceptive use based on number
of children
ggplot(cmc, aes(x=NumChildren)) +

geom_bar(aes(fill=ContraceptiveMethod
Used))
```

Husbands with higher education tend to have a larger cluster of contraceptive method used. The usage of contraceptive tends to scatter toward husbands with higher education and children from 0 - 5.

```
#Measure of husbands! Y = number of children, X = education
ggplot(cmc, aes(x=HusbandEducation, y =NumChildren)) +

geom_jitter(aes(fill=ContraceptiveMethodUsed))
```





The distribution of contraceptive methods is fairly even between women who are and are not Islamic. Women who were not Islamic had a lower count of using contraceptions.

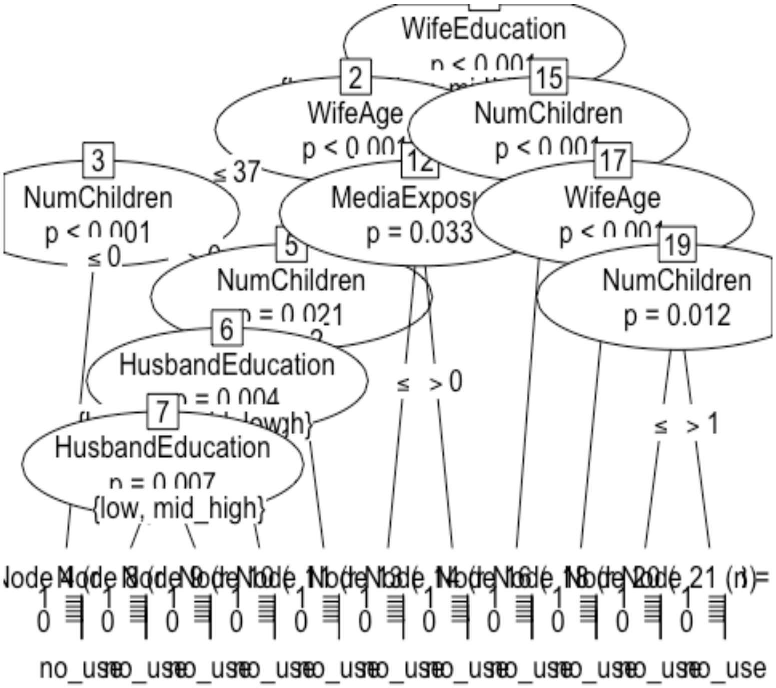
```
#relationship of religion to the contraceptive method used
ggplot(cmc, aes(x = WifeReligion)) +
  geom_bar(aes(fill = ContraceptiveMethodUsed))
```

Now that we have finished graphically exploring the data and the preprocessing, it is time to start analyzing the data and trying to find which algorithm is best predictive of the method of contraception used by women. I am applying decision trees, random forests, linear discriminant analysis, quadratic discriminant analysis, and neural networks.

The first thing that needs to be done is splitting the data. I have split the data into 75% training set and 25% testing set.

METHODS

```
#split data in training and test set
set.seed(1357) #set a random seed
index <- sample(1:dim(cmc)[1], floor(0.75 * dim(cmc)[1]))
train <- cmc[index,] #75% training data
test <- cmc[-index,] #25% test data

#build decision tree on training data
cmcFormula <- ContraceptiveMethodUsed ~ WifeAge + WifeEducation + HusbandEducation + NumChildren +
            WifeReligion + IsWifeWorking + HusbandOccupation + StandardofLiving + MediaExposure
cmc_tree <- ctree(cmcFormula, data = train)
table(predict(cmc_tree), train$ContraceptiveMethodUsed)

##
##              no_use long_term short_term
##   no_use       302        52        87
##   long_term     48       129        82
##   short_term   120        68       216

plot(cmc_tree)
```



The number of children plays an important role in the decision tree. It breaks the tree into two sets: those who have had children before and those who have not. Women that have not yet had a child are almost certain to be using no contraception at all.

After that initial split, the age of the wife plays a key role in the split between choices. Next is the wife's education level, and another split on number of children.

```
#build decision tree on test data
testPred <- predict(cmc_tree, newdata = test)
table(testPred, test$ContraceptiveMethodUsed)


## testPred     no_use long_term short_term
##   no_use         99        20         29
##   long_term      19        30         30
##   short_term     41        34         67
```

```
#find the accuracy of a decision tree
cmc.rpart <- rpart(ContraceptiveMethodUsed ~. , data = cmc, method = "class")
cmc.predict <- predict(cmc.rpart, test, type = "class")
results <- cmc.predict == test$ContraceptiveMethodUsed
(accuracy <- sum(results) / length(results))

## [1] 0.5392954
```

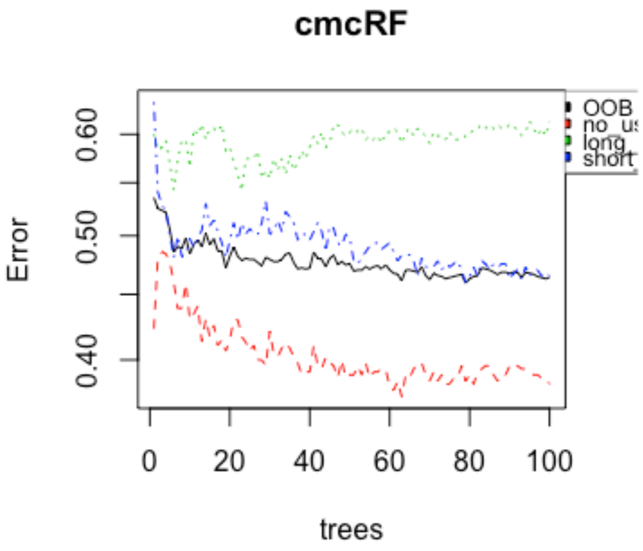
The accuracy for a decision tree is about 54%.

```
train(ContraceptiveMethodUsed ~ ., data = cmc, method = "rpart")

## CART
##
## 1473 samples
##    9 predictor
##    3 classes: 'no_use', 'long_term', 'short_term'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 1473, 1473, 1473, 1473, 1473, 1473, ...
## Resampling results across tuning parameters:
##
##   cp          Accuracy   Kappa
##   0.02310427  0.5300309  0.26754322
##   0.03791469  0.5025314  0.22021389
##   0.06931280  0.4494629  0.08295314
##
## Accuracy was used to select the optimal model using  the largest value.
## The final value used for the model was cp = 0.02310427.
```

Since the accuracy for a decision tree was pretty weak - about the same as tossing a coin - I utilized cross validation to create a better model. However, this did not work. The accuracy ended up lower than the decision tree.

The next method I employed was a random forest.

```
cmcRF <- randomForest(ContraceptiveMethodUsed ~ ., data = train, ntree = 100)
table(predict(cmcRF), train$ContraceptiveMethodUsed)

##
##             no_use long_term short_term
##   no_use       290        71        114
##   long_term     47       103         82
##   short_term   133        75        189

layout(matrix(c(1,2),nrow=1),
width=c(4,1))
par(mar=c(5,4,4,0)) #No margin on the right side
plot(cmcRF, log="y")
par(mar=c(5,0,4,2)) #No margin on the left side
plot(c(0,1),type="n", axes=F, xlab="", ylab="")
legend("top",
colnames(cmcRF$err.rate),col=1:4,cex=0.8,fill=1:4)
```



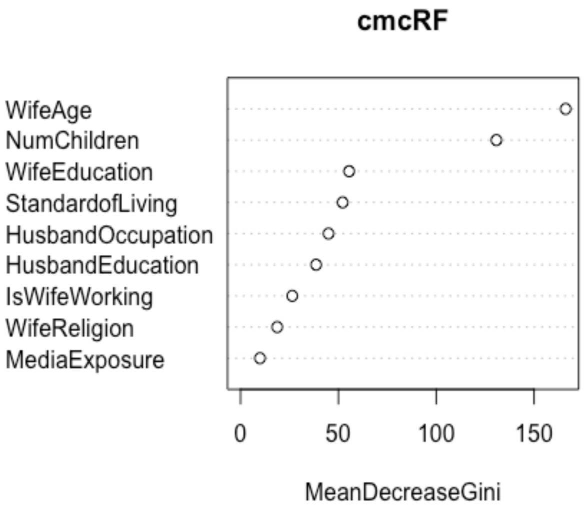
This is a plot of the out of bag error rates of the different classes of the random forest. The black curve represents the out of bag error rate curve and the others are the misclassification error rates, as stated in the legend. The error rate is the prediction error based on out-of-bag (OOB) data. The error rate of OOB is higher than the prediction error of the original data.

Next, I explored the different trees in the forest.

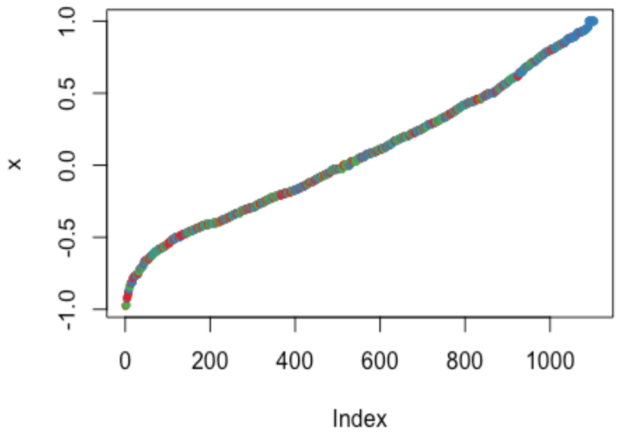
```
cmcGetTree <- getTree(cmcRF, 1, labelVar=TRUE)
```

By looking at the variable importance plot, we can see the most important predictors. Media exposure and the religion of the woman are important factors in using our random forest algorithm, and the age of the wife is not as important.

```
varImpPlot(cmcRF)
```



```
cmcPred <- predict(cmcRF, newdata = test)
table(cmcPred, test$ContraceptiveMethodUsed)

##
## cmcPred      no_use long_term short_term
##   no_use       101        19         39
##   long_term     20        31         23
##   short_term    38        34         64

plot(margin(cmcRF, test$ContraceptiveMethodUsed))
```



The margin of a data point is the proportion toward the correct class minus the maximum proportion for the other classes. A positive margin means correct classification, and we can see that about 50% of the margin has been correctly classified.

I obtained the confusion matrix of the data to find the accuracy. The highest value is for long-term contraceptive use, which still stays around 58%.

```
conf <- cmcRF$confusion
conf[, 'class.error']

##     no_use  long_term short_term
##  0.3829787  0.5863454  0.5090909
```

My next method is LDA - linear discriminant analysis. I used cross validation and leave-one-out cross validation to help fit the model. The LDA function tries to detect if the within-class covariance matrix is singular. The first few linear discriminants emphasize the differences between groups within the weights given by the rotation of the linear discriminants within their space.

We can see the means of each probability of each response group in relation to the variables in the predictor, as well as the coefficients of the discriminants in LD1 and LD2.

```
CV <- trainControl(method = "cv", number = 10, classProbs = TRUE) #10-fold cross validation
LOOCV <- trainControl(method = "LOOCV", classProbs = TRUE)

lda_CV <- train(ContraceptiveMethodUsed ~ . , data = cmc, method = "lda", metric = "Accuracy",
            trControl = CV)

lda_loocv <- train(ContraceptiveMethodUsed ~ ., data = cmc, method = "lda", metric = "Accuracy",
              trControl = LOOCV)

lda_vs <- lda(ContraceptiveMethodUsed~ . , data = train)

lda_CV$results

##   parameter  Accuracy     Kappa AccuracySD    KappaSD
## 1      none 0.5098685 0.2349374 0.03881762 0.05880821

lda_loocv$results

##   parameter Accuracy     Kappa
## 1      none 0.509165 0.2323845
```

The accuracy for LDA is still about 50%. Cohen's kappa, which measures inter-rate agreement for the categorical variables, is also fairly low. We have still not found a sufficient model.

```
lda_pred <- predict(lda_vs, test)
accuracy_lda <- mean(lda_pred$class == test$ContraceptiveMethodUsed)
accuracy_lda

## [1] 0.4796748
```

I applied QDA to see if this would be better, again for CV and LOOCV. The accuracy ended up even lower - less than 45%.

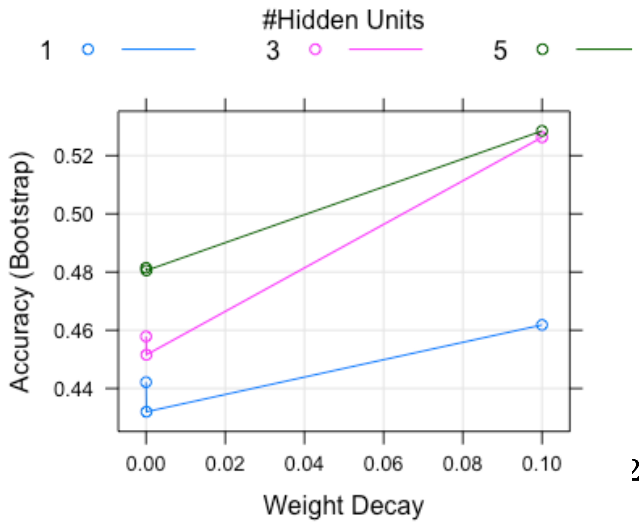
```
set.seed(111)
qda_CV <- train(ContraceptiveMethodUsed~., data = cmc, method = "qda", metric = "Accuracy",
            trControl = CV)

qda_loocv <- train(ContraceptiveMethodUsed~., data = cmc, method = "qda", metric = "Accuracy",
              trControl = LOOCV)
```

```
qda_vs <- qda(ContraceptiveMethodUsed~., data = train)

qda_CV$results

##   parameter Accuracy     Kappa AccuracySD    KappaSD
## 1      none 0.461624 0.2168181 0.03213909 0.04204934

qda_loocv$results

##   parameter  Accuracy    Kappa
## 1      none 0.4663951 0.223769

 qda_pred <- predict(qda_vs, test)
accuracy_qda <- mean(qda_pred$class == test$ContraceptiveMethodUsed)
accuracy_qda

## [1] 0.4444444
```

This method only takes into account the numerical attributes, such as number of children or age of the wife, so therefore it isn’t necessarily the strongest method to use when there are so many other predictor variables available.

The last method I am going to apply is a neural network. After training the neural net with about 100 iterations, the accuracy is still not better than any other method - around low 50%.

```
cmcNeuralNet <- train(ContraceptiveMethodUsed ~., data = cmc, method = "nnet")

## Loading required package: nnet

cmcNeuralNet

## Neural Network
##
## 1473 samples
##    9 predictor
##    3 classes: 'no_use', 'long_term', 'short_term'
##
## No pre-processing
## Resampling: Bootstrapped (25 reps)
## Summary of sample sizes: 1473, 1473, 1473, 1473, 1473, 1473, ...
## Resampling results across tuning parameters:
##
##   size  decay  Accuracy   Kappa
##   1     0e+00  0.4334669  0.05916020
##   1     1e-04  0.4392447  0.05749969
##   1     1e-01  0.4653978  0.14180606
##   3     0e+00  0.4501278  0.10514400
##   3     1e-04  0.4583849  0.13063631
##   3     1e-01  0.5164214  0.25515052
##   5     0e+00  0.4620439  0.16910705
##   5     1e-04  0.4779645  0.19377132
##   5     1e-01  0.5343126  0.28375773
##
## Accuracy was used to select the optimal model using  the
largest value.
```



```
## The final values used for the model were size = 5 and decay = 0.1.

plot(cmcNeuralNet)
```

This is a graph of the hidden units in a layer of the neural net, which feeds into more layers of hidden units to feed into the output layer. The three hidden units displayed here are a respectively squashed linear function of its inputs. The bootstrapped accuracy shows that the hidden unit #5 has the highest accuracy, around 52%.

METHOD	Highest Accuracy Rate
Decision Tree	0.5393
Random forest	0.58634
LDA	0.47968
QDA	0.44444
Neural net	0.534313

After working through these methods, all of the algorithms generated accuracy between 50-60%, which is slightly better than random choice, which in this case would be 33% since this is a classification on three levels. The method that had the highest accuracy was the random forest, about 58% for long-term usage. LDA and QDA only utilized the continuous predictor variables in the model – wife's age and number of children. The cause of these error rates could possibly be due to the fact that there were seven categorical variables that served as the predictor variables (before conversion) and that this is a classification problem that classifies on three factors, which leads to a class imbalance. There was not an even distribution of each type of contraceptive method used, so there could have been errors in false positive or false negatives.

It is mostly surprising that a three way classification problem could be analyzed most accurately with a model as simple as random forests. Having a neural net which trained over 100 iterations did not make a significant difference in predicting the accuracy of the model – in fact even having the five hidden units in the neural net, it still only achieved about 52% accuracy.

CONCLUSION

The ongoing study of population control in Indonesia has been mitigated and helped through the implementation of contraceptive methods. This study has furthered an understanding of the behavior and reasoning behind Indonesian women making decisions about contraceptive methods. Beyond predicting the most commonly used contraceptive method used, this analysis furthers how better to educate women and families about childbearing and raising families.

To take the most effective route for educating families about contraception, a random forest algorithm allows us to understand what the most important factors are in behavioral patterns of the women and how their demographic influences their decisions. Methods like LDA and QDA, which emphasize the different weights between groups, is not as effective because it predicts solely on continuous numerical variables. A decision tree taught us that factors such as age of the wife of number of children were important to take into account when analyzing the method of contraception used.

Because there were nine variables that were included in the predictor, considering each factor is essential when educating the demographic for women who use or do not use contraceptive methods. For example, since the median age for a woman in Indonesia to have her first child is at 22.8 years old, this demographic should be targeted for education about long-term contraceptive methods while in university. We should also take into account all factors of the demographic of the families, including the education level of the husband and whether or not the woman was religious. These are all important variables in the decision for a family to control its population or choose to not use contraception.

Overall, the metric of success was accuracy in this study. In this multi-class classification problem, being able to generate accuracy greater than 33% meant that the models were trained fairly well. This ongoing practice will help the “Family Planning Program” in Indonesia control its population and provide education for families on safer practices.

REFERENCES

Loh, WeiYin, TjenSien Lim, and YuShan Shih. "A Comparison of Prediction Accuracy, Complexity, and ..." N.p., n.d. Web. 21 Mar. 2017.

Place, Graham. "Sign In." RPubs - Modeling Contraceptive Use in Indonesia. N.p., n.d. Web. 21 Mar. 2017.

"UCI Machine Learning Repository: Contraceptive Method Choice Data Set." UCI Machine Learning Repository: Contraceptive Method Choice Data Set. N.p., n.d. Web. 21 Mar. 2017.