Constructing a Fertility Classification Model

Hailey Tan haileyt

Due Friday, April 18, at 11:59PM

Contents

```r
set.seed(151)
library("knitr")
library("dplyr")
library("kableExtra")
library("pander")
library("readr")
library("magrittr")
library("car")
library("MASS")
library("klaR")
library("tree")
library("rpart")
library("rpart.plot")
```

Introduction

Reproduction is fundamental to the long-term survival of the human species, and for many individuals, the ability to have children is a deeply important aspect of life. However, fertility testing can be costly and difficult to access, creating barriers for those seeking answers about their reproductive health. As a result, developing a model that can predict fertility outcomes using easily obtainable covariate data would be highly valuable. In this paper, we explore basic machine learning classification techniques to determine whether a fertility test would indicate normal or altered fertility, aiming to provide a more accessible approach to fertility assessment.

Exploratory Data Analysis

Background and Variables

The data used for this model was collected by a group of researchers from the University of Alicante in Spain. 100 men participated in this study, where they were tested for fertility and asked to fill out a questionnaire regarding their life habits.

The following predictor variables were extracted from each participant:

Season: Winter, Spring, Summer, Fall

Age: age in years

ChildishDisease: if the patient has ever had a child disease (chicken pox, measles, mumps, polio, . . .)

SeriousTrauma: if the patient has ever had an accident or serious trauma

SurgicalIntervention: if the patient has ever had a surgical intervention

Fevers1year: high fevers in the last year - less than three months ago, more than three months ago, no fevers

AlcoholUse: frequency of alcohol consumption: (1) several times a day, (2) every day, (3) several times a week, (4) once a week, (5) hardly ever or never

Smoking: never, occasional, daily

and our response variable to be predicted:

Output: diagnosis of fertility test (normal, altered)

Summary of Labels in the Training Dataset

Of the 70 observations in the training dataset, 8 fertility tests (11.4%) were classified as "altered," while 62 tests (88.6% of the tests) were classified as "normal." They are shown in the tables below:

```r
table(fertility_train$Output)
```

```
##
## altered  normal
##       8      62
```

```r
prop.table(table(fertility_train$Output))
```

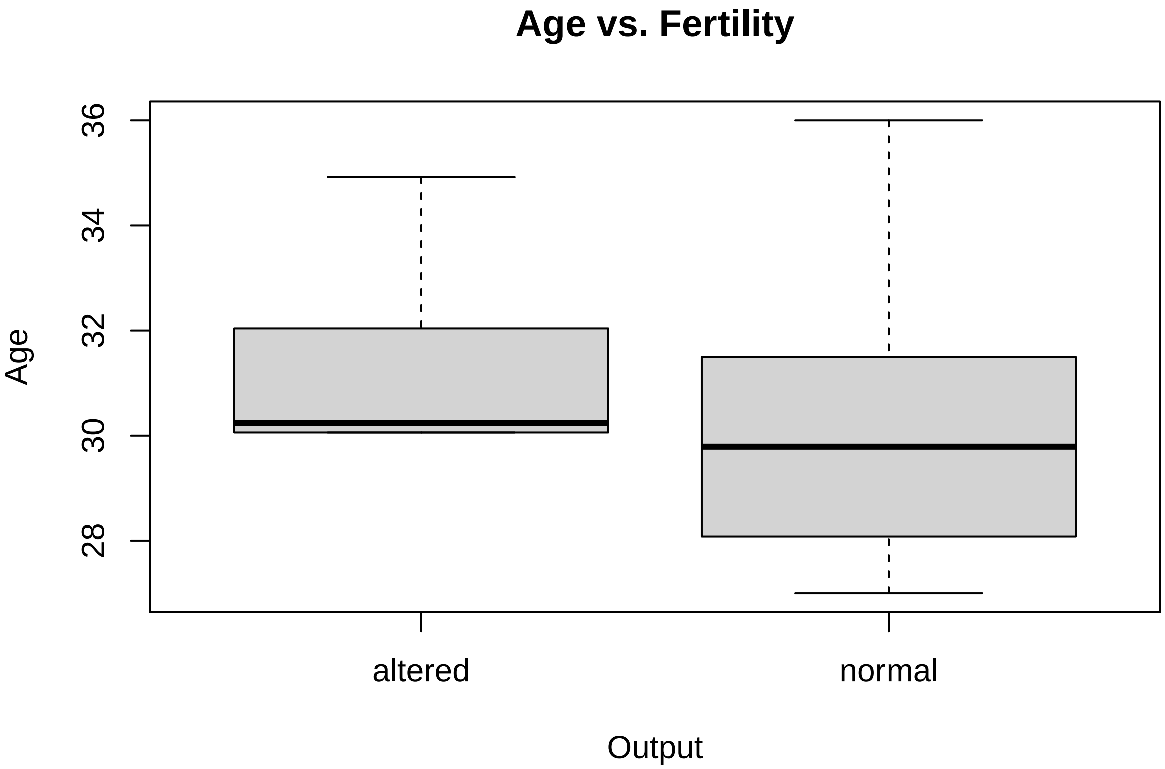
```
##
##   altered    normal
## 0.1142857 0.8857143
```

EDA relationships between fertility and predictor variables

We now move to performing elementary data analysis (EDA) on the variables to visualize their individual relationships with the response variable (fertility).

For quantitative predictors, we will uses boxplots, depicted below:

```
boxplot(Age ~ Output,
        main = "Age vs. Fertility",
        data = fertility_train)
```



We note that since the boxes between “altered” and “normal” overlap heavily, age is likely not a strong predictor for a male’s fertility.

For the categorical variables, we create barplots to visualize the conditional proportions for normal or altered fertility based on each categorical variable.

```
prop.table(table(fertility_train$Output, fertility_train$Season))
```

```
##
##                 Fall     Spring     Summer     Winter
##   altered 0.07142857 0.02857143 0.00000000 0.01428571
##   normal  0.28571429 0.31428571 0.02857143 0.25714286
```

```
prop.table(table(fertility_train$Output, fertility_train$ChildishDisease))
```

```
##
##                   no        yes
##   altered 0.08571429 0.02857143
##   normal  0.77142857 0.11428571
```

```
prop.table(table(fertility_train$Output, fertility_train$SeriousTrauma))
```

```
##
##                   no        yes
##   altered 0.01428571 0.10000000
##   normal  0.40000000 0.48571429
```

```
prop.table(table(fertility_train$Output, fertility_train$SurgicalIntervention))
```

```
##
##                   no        yes
##   altered 0.07142857 0.04285714
##   normal  0.45714286 0.42857143
```

```
prop.table(table(fertility_train$Output, fertility_train$Fevers1year))
```

```
##
##           less3months more3months         no
##   altered  0.01428571  0.07142857 0.02857143
##   normal   0.07142857  0.55714286 0.25714286
```

```
prop.table(table(fertility_train$Output, fertility_train$AlcoholUse))
```

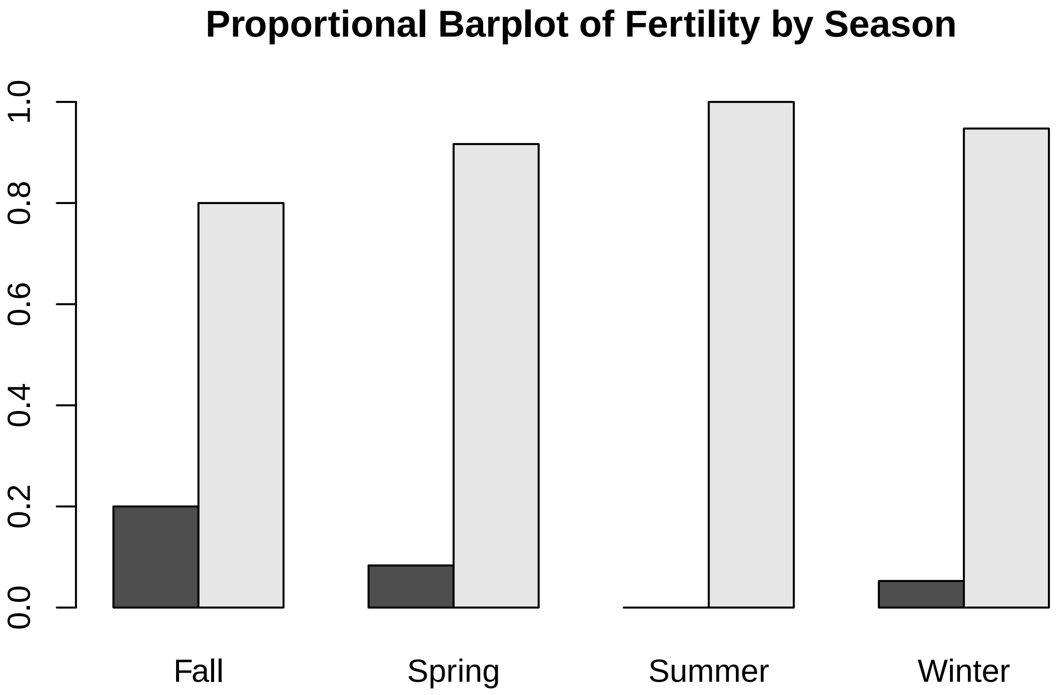
```
##
##                    2          3          4          5
##   altered 0.00000000 0.05714286 0.02857143 0.02857143
##   normal  0.01428571 0.15714286 0.31428571 0.40000000
```

```
prop.table(table(fertility_train$Output, fertility_train$Smoking))
```

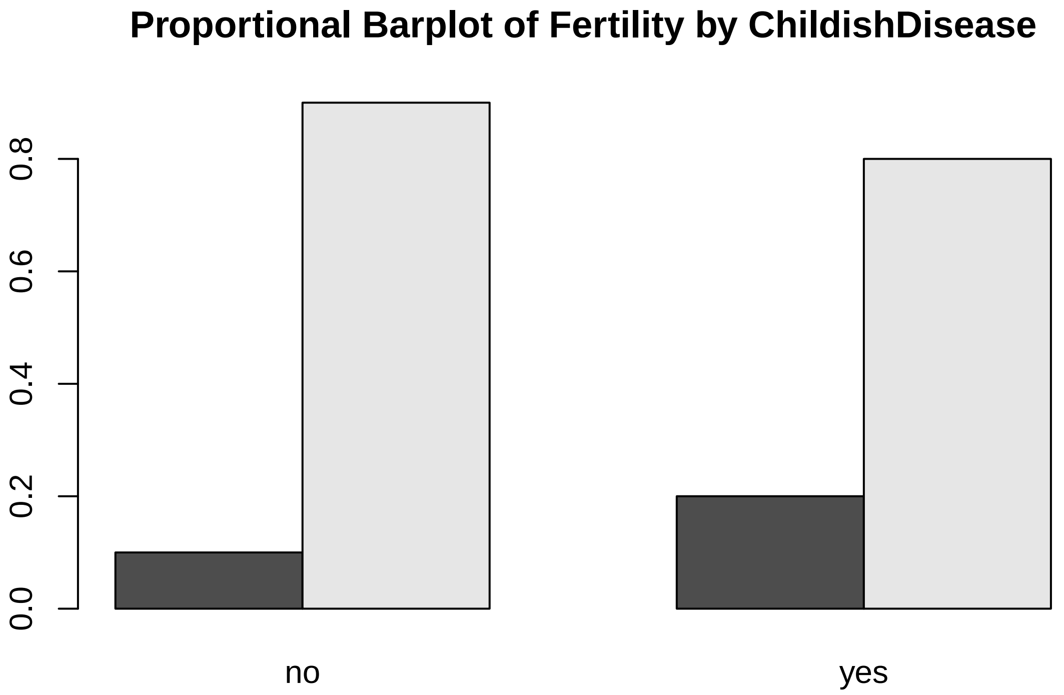
```
##
##                daily      never occasional
##   altered 0.04285714 0.04285714 0.02857143
##   normal  0.21428571 0.52857143 0.14285714
```

We are also able to examine barplots to view these proportions visually.

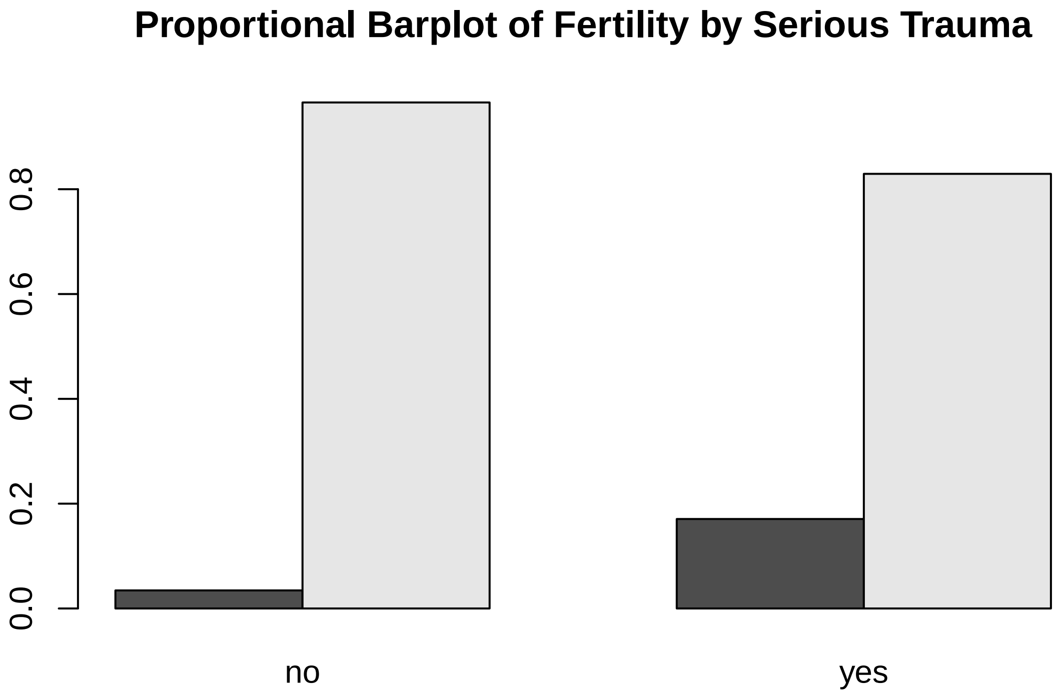
```
barplot(prop.table(table(fertility_train$Output, fertility_train$Season),
                   margin = 2),
        beside = TRUE,
        main = "Proportional Barplot of Fertility by Season")
```



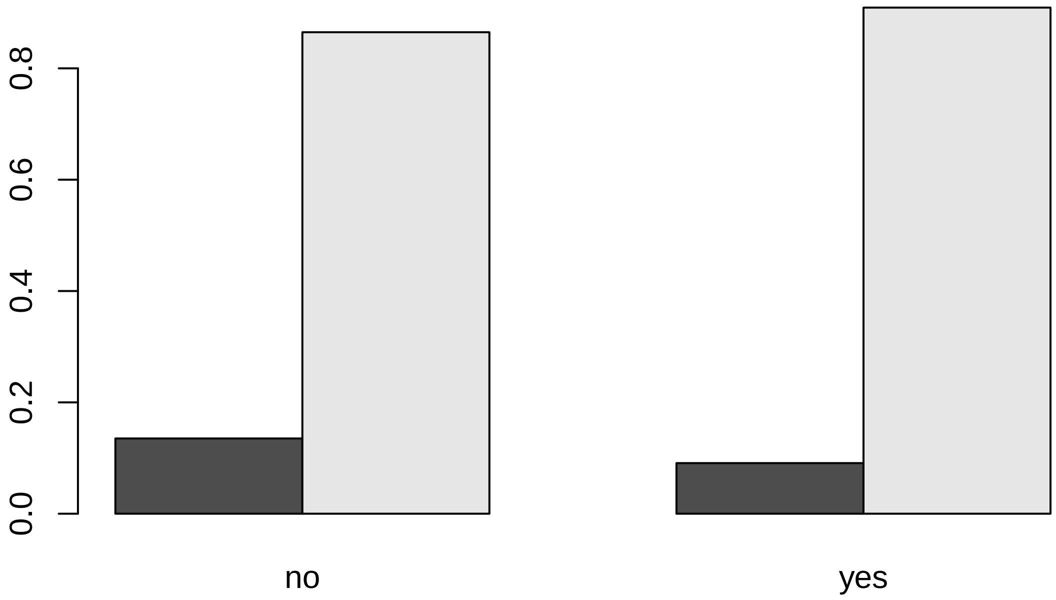
```
barplot(prop.table(table(fertility_train$Output, fertility_train$ChildishDisease),
                   margin = 2),
        beside = TRUE,
        main = "Proportional Barplot of Fertility by ChildishDisease")
```



```
barplot(prop.table(table(fertility_train$Output, fertility_train$SeriousTrauma),
                   margin = 2),
        beside = TRUE,
        main = "Proportional Barplot of Fertility by Serious Trauma")
```

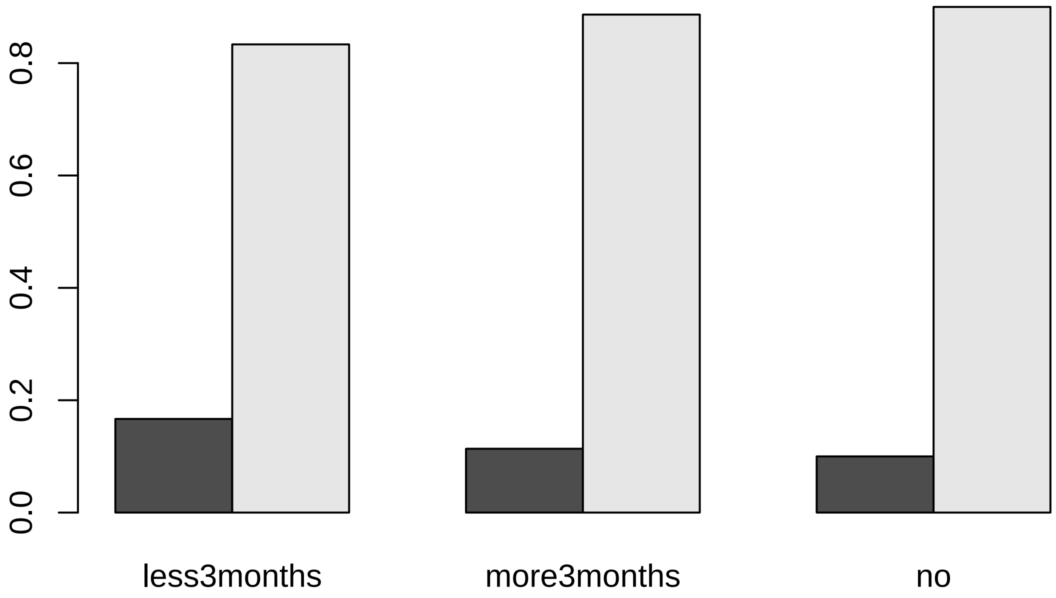


```r
barplot(prop.table(table(fertility_train$Output, fertility_train$SurgicalIntervention),
                   margin = 2),
        beside = TRUE,
        main = "Proportional Barplot of Fertility by Surgical Intervention")
```



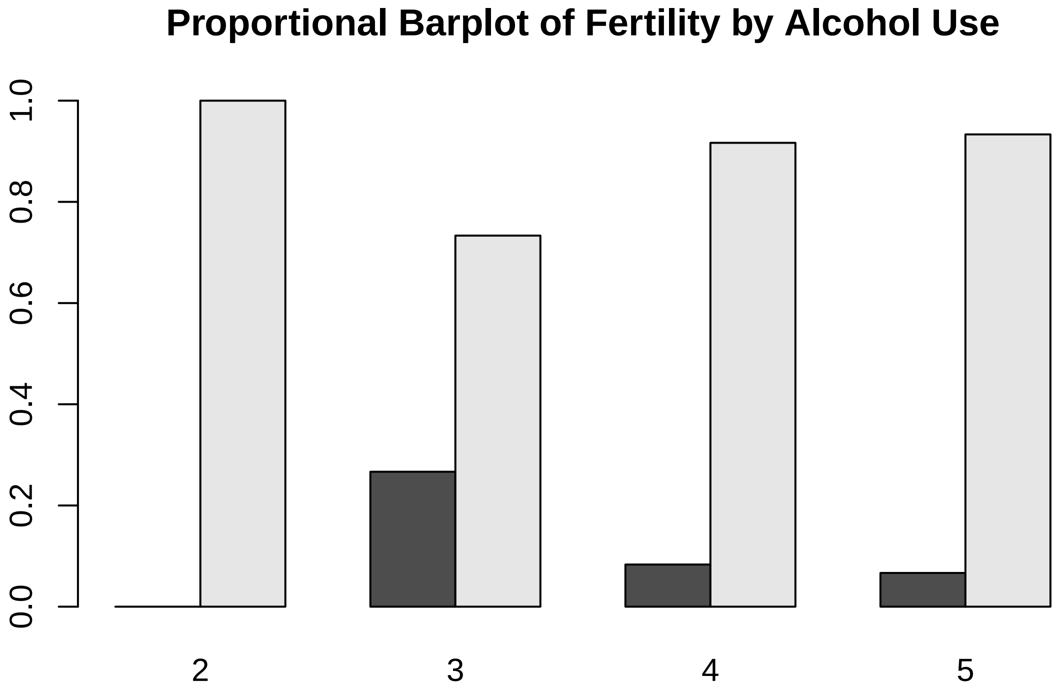
Proportional Barplot of Fertility by Surgical Intervention

```r
barplot(prop.table(table(fertility_train$Output, fertility_train$Fevers1year),
                   margin = 2),
        beside = TRUE,
        main = "Proportional Barplot of Fertility by Fevers1Year")
```

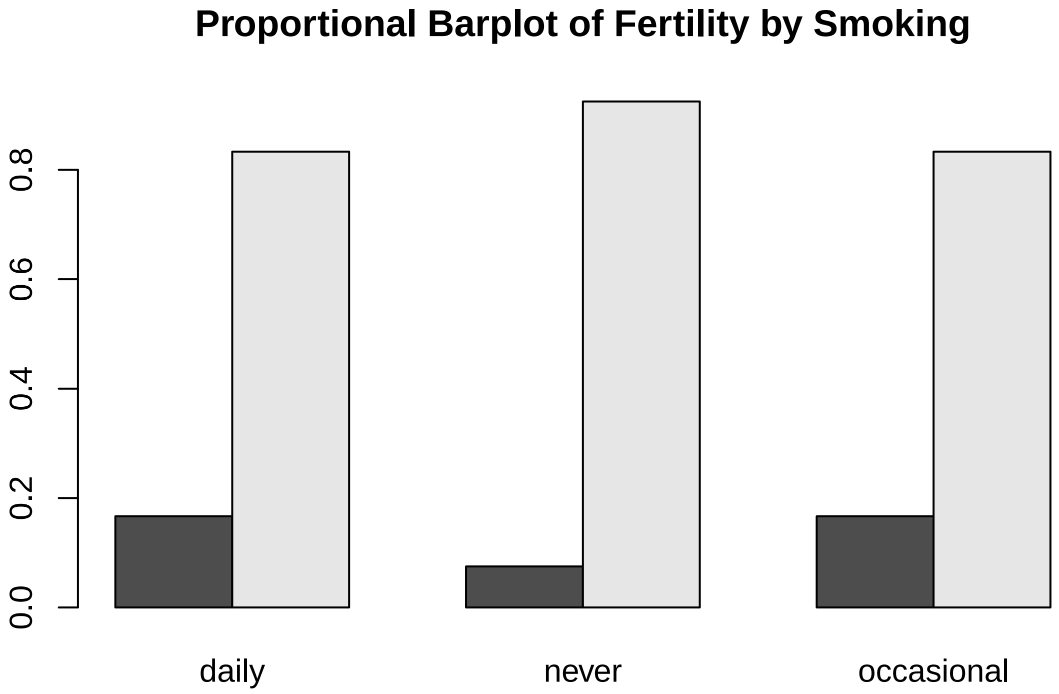


Proportional Barplot of Fertility by Fevers1Year

```r
barplot(prop.table(table(fertility_train$Output, fertility_train$AlcoholUse),
                   margin = 2),
        beside = TRUE,
        main = "Proportional Barplot of Fertility by Alcohol Use")
```



```r
barplot(prop.table(table(fertility_train$Output, fertility_train$Smoking),
                   margin = 2),
        beside = TRUE,
        main = "Proportional Barplot of Fertility by Smoking")
```



Based on the summaries above, several key observations can be made regarding factors potentially associated with altered fertility. Fall appears to be linked with higher levels of altered fertility, suggesting that seasonal changes may play a role. Additionally, individuals who experienced childhood diseases seem to have a higher likelihood of fertility issues later in life, and those who have endured serious trauma also show an increased risk of altered fertility, possibly reflecting the lasting effects of physical or emotional stress on reproductive health. However, there does not appear to be a strong association between surgical interventions or fevers and fertility, nor between alcohol use and fertility outcomes. In contrast, smoking stands out as a significant factor, with smokers exhibiting a higher likelihood of altered fertility, highlighting the long-term reproductive consequences of this lifestyle choice.

Modeling

We can now begin developing and evaluating different classifiers for predicting fertility. The four classifier types we will explore include linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), classification trees, and binary logistic regression. To prevent overfitting, we will randomly divide our data into training and testing sets. Each of the four models will be tested using the same training and testing datasets.

Linear Discriminant Analysis (LDA)

```
fertility.lda <- lda(factor(Output) ~ Age, data = fertility_train)
```

```
fertility.lda.pred <- predict(fertility.lda,
                              as.data.frame(fertility_test))
```

```
table(fertility.lda.pred$class, fertility_test$Output)
```

```
##
##           altered normal
##   altered       0      0
##   normal        4     25
```

For both LDA and QDA, we use only quantitative variables in our model. When tested on the data, LDA achieved an overall error rate of 4/29, or approximately 0.138, which is relatively low. It performed especially well at predicting normal fertility, with an error rate of 0/25, or 0%. However, LDA struggled with classifying altered fertility, resulting in an error rate of 4/4, or 100%.

Quadratic Discriminant Analysis (QDA)

```
fertility.qda <- qda(factor(Output) ~ Age, data = fertility_train)
```

```
fertility.lda.pred <- predict(fertility.lda, as.data.frame(fertility_test))
```

```
table(fertility.lda.pred$class, fertility_test$Output)
```

```
##
##           altered normal
##   altered       0      0
##   normal        4     25
```

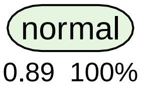
Despite being more flexible, QDA performed similarly to LDA, with identical error rates. The overall error rate for QDA was 0.114, just like LDA.

Classification Trees

We can account for categorical variables in classification trees, which can make it advantageous over LDA and QDA classifiers.

```
fertility.tree <- rpart(factor(Output) ~ Age + factor(ChildishDisease) +
                        factor(SeriousTrauma) + factor(SurgicalIntervention) +
                        factor(Fevers1year) + factor(AlcoholUse) + factor(Smoking),
                        data = fertility_train,
                        method = "class")
```

```
rpart.plot(fertility.tree,
           type = 0,
           clip.right.labs = FALSE,
           branch = 0.1,
           under = TRUE)
```



The classification tree did not use any predictors to classify fertility; instead, it simply categorized everyone as having normal fertility. Now, we can evaluate the performance of our tree classifier on the test data, which we expect to be similar to that of the LDA and QDA classifiers due to its classification approach.

```
fertility.tree.pred <- predict(fertility.tree, as.data.frame(fertility_test),
                               type = "class")
```

```
table(fertility.tree.pred, fertility_test$Output)
```

```
##
## fertility.tree.pred altered normal
##             altered       0      0
##             normal        4     25
```

As anticipated, the classification tree resulted in the same errors as the LDA and QDA classifiers, since it classified everyone as having normal fertility.

Binary Logistic Regression

The final model we will consider for predicting fertility is binary logistic regression. Similar to classification trees, the logistic regression model can incorporate both quantitative and categorical variables.

```
fertility.logit <- glm(factor(Output) ~ Age + factor(ChildishDisease) + factor(SeriousTrauma)
                       + factor(SurgicalIntervention) + factor(Fevers1year) + factor(AlcoholUse)
                       + factor(Smoking),
                       data = fertility_train,
                       family = binomial(link = "logit"))
```

```
fertility.logit.prob <- predict(fertility.logit, as.data.frame(fertility_test),
                                type = "response")
```

Unlike the other models, the logistic regression model outputs probabilities for altered or normal fertility rather than making direct classifications. To turn these probabilities into classification predictions, we assign individuals to a category if their predicted probability of belonging to that category exceeds 0.5.

```
levels(factor(fertility_test$Output))
```

```
## [1] "altered" "normal"
```

Since the “levels” command returned “normal” as the second category, it is treated as the “yes” or 1 category in the logistic model.

With this in mind, we can now convert the predicted probabilities into classification outcomes based on the chosen threshold.

```
fertility.logit.pred <- ifelse(fertility.logit.prob > 0.5, "normal", "altered")
```

```
table(fertility.logit.pred, fertility_test$Output)
```

```
##
## fertility.logit.pred altered normal
##              altered       0      2
##              normal        4     23
```

The logistic regression model performed the worst among all the classifiers, with an overall error rate of (4+2)/29, or approximately 0.21. It failed to correctly classify any cases of altered fertility, resulting in an error rate of 4/4, or 100%, while its error rate for normal fertility was 2/25, or 0.08. As with the other models, logistic regression showed better performance when predicting normal fertility.

Final Recommendation

Among the four classifiers, LDA, QDA, and the classification tree performed equally well, while logistic regression performed slightly worse. All models showed stronger performance when predicting normal fertility, which is likely due to overfitting—particularly since the first three models classified every individual as having normal fertility. This approach led to a 100% error rate when predicting altered fertility, meaning the models were no more effective than simply assuming everyone has normal fertility.

Given these results, our final recommendation is to avoid using any of the classification models developed, as none provided better predictive accuracy than a basic, uniform classification.

Discussion

When looking solely at error rates, our models appeared to perform reasonably well in classifying fertility. However, a deeper examination reveals a significant concern: the models did not base their classifications on any meaningful patterns or specific variables. Instead, most of them defaulted to classifying every individual as having normal fertility. This is a clear red flag, as it suggests that the models were not learning from the data in a meaningful way. Such behavior implies that the variables we collected may not be strong predictors of fertility outcomes, or perhaps more critically, that our dataset—especially the subset with altered fertility—was too small to capture any real underlying patterns.

This limitation highlights the importance of sample size and data balance in predictive modeling. With so few cases of altered fertility, the models lacked the variability needed to generalize across features and make accurate distinctions. For future research, it will be essential to collect a larger and more balanced dataset that includes a sufficient number of individuals with both normal and altered fertility. Doing so will allow for a more thorough evaluation of different classifiers and help determine whether meaningful predictive relationships exist. Until then, any conclusions drawn from the current models should be approached with caution.