Factors that affect Men's Fertility

Nathan Yao nyao

Due Monday, July 29, at 11:59PM

Contents

Introduction

A common reproduction issue that is studied around the world is fertility. More specifically, the focus has been on the fertility of adult men. Fertility levels have decreased in the past two decades and getting fertility tested can be a financial struggle.

Instead, doctors, researchers, and statisticians have focused on possible factors that may predicts fertility. There are a variety of different possible factors that include environmental and health measures. This paper will explore how to classify fertility through various machine learning techniques, and use data from "Predicting seminal quality with artificial intelligence methods". [data from: David Gil, Jose Luis Girela, Joaquin De Juan, M. Jose Gomez-Torres, and Magnus Johnsson. Predicting seminal quality with artificial intelligence methods. Expert Systems with Applications, 39(16):12564 12573, 2012]

Exploratory Data Analysis

Background and Variables:

The data that is going to be analyzed is provided by a research paper written by David Gil and a number of other machine learning and artificial intelligence researchers. The data was collected in 2012 where 99 men participated in a study that tested their fertility. A number of explanatory variables were recorded:

- **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

The response variable that is being classified is:

- **Output**: diagnosis of fertility test (normal, altered).

Summary of the Response Labels in the Training Dataset

The data sets that will be analyzed include a sample of 99 men that were tested in the study. The fertility training set will include 70 men and the fertility testing set will contain 29 men. The training set has 8 altered fertility outputs and 62 normal fertility outputs with proportions of 11.43% and 88.57% respectively. The following tables describe Output variable of the training data set:

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

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

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

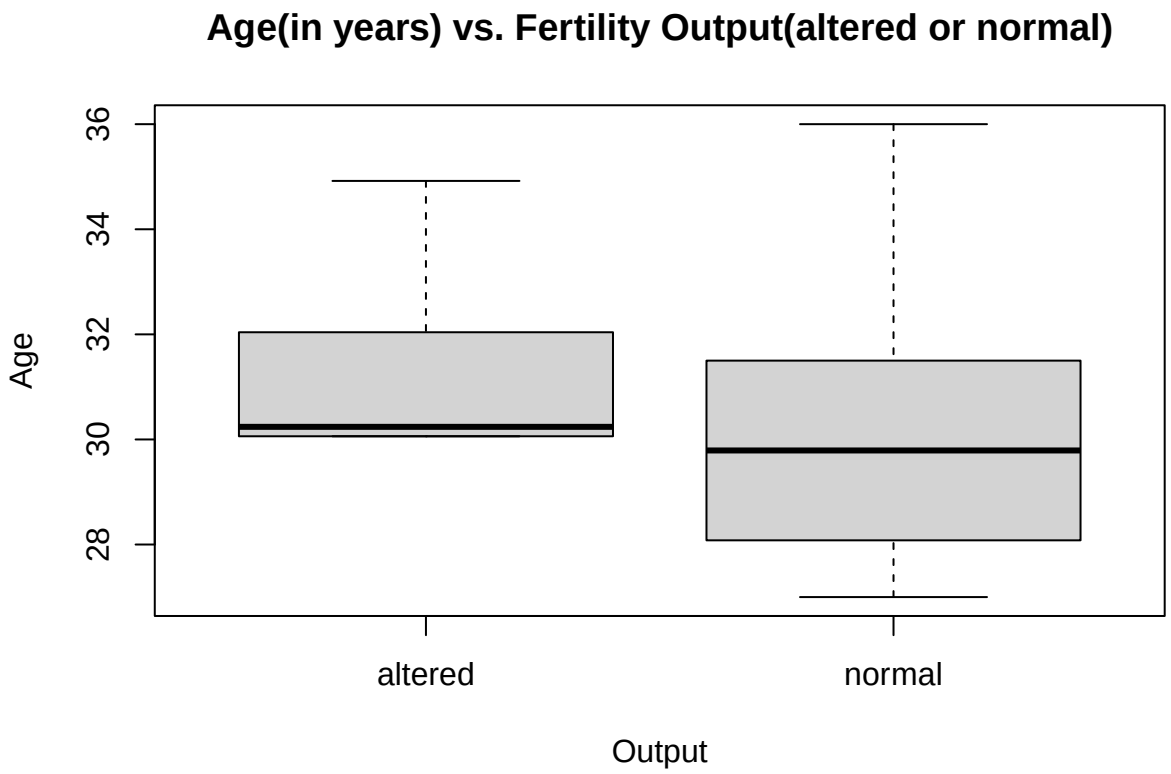
```
##
##   altered    normal
## 0.1142857 0.8857143
```

Some EDA on relationships between fertility output and the quantitative variable

The explanatory variables listed above include both quantitative and categorical variables. The quantitative explanatory variable, Age, will be analyzed first with a box plot with the response variable of fertility output. The histogram of the age distribution and box plot predicting fertility output is below:

```
hist(fertility_train$Age, main = "Histogram of Age of Patients", xlab = "Age in years")
```


```
boxplot(Age ~ Output, main="Age(in years) vs. Fertility Output(altered or normal)", data = fertility_tra
```



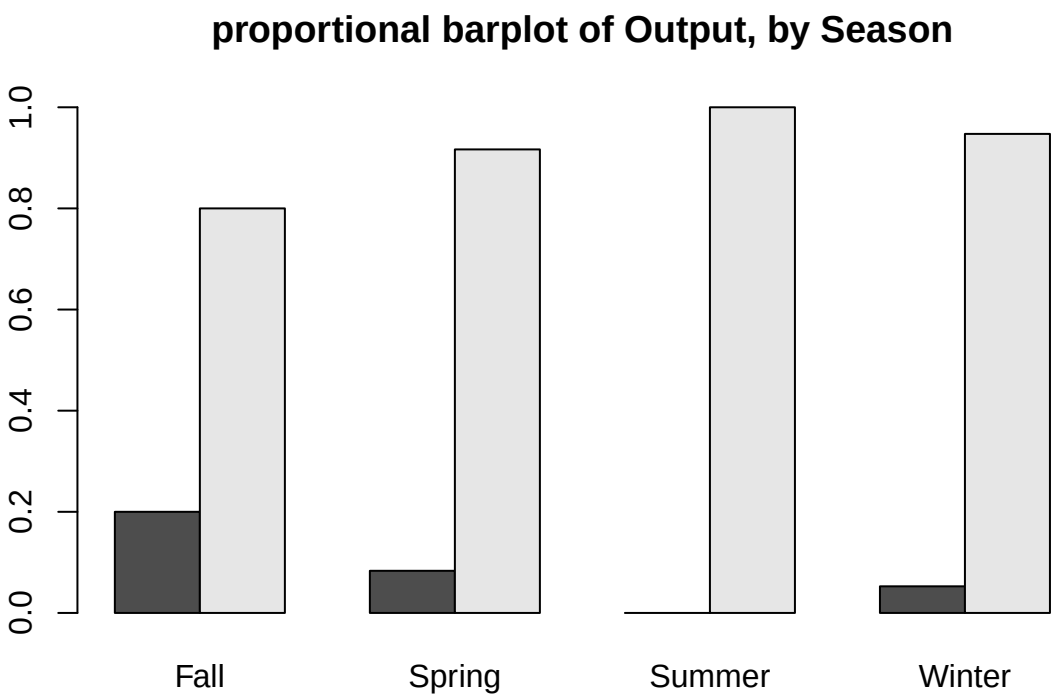
From the histogram above, we can see that the distribution of age in the training data set is bimodal with peaks at the 28-29 years old mark and the 30-31 years old mark. The distribution is also right skewed without any obvious outliers.

Looking at the box plot, it can be noted that ages of the altered fertility outputs tend to be older than the ages of the normal fertility outputs. The median of the two box plots are very similar but the normal fertility output subjects seems to be concentrated on the lower end of the ages displayed. The range of the altered group also has a smaller range compared to the normal group. The observations above makes it seem like age could be a classifier for fertility.

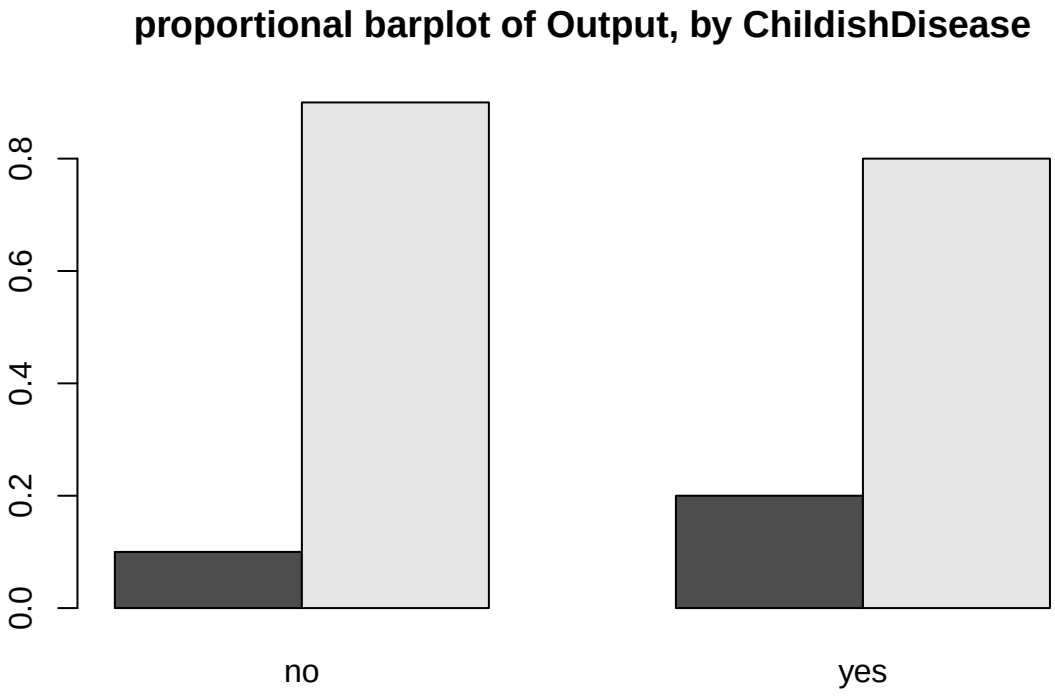
EDA on relationships between fertility output and the categorical variables

The rest of the variables that were measured were categorical. These variables include season, if the patient has ever had a child disease, if the patient has ever had an accident or serious trauma, if the patient has ever had a surgical intervention, high fevers in the last year, frequency of alcohol consumption, and the frequency of how much the patient smokes. The following are proportional bar plots between each of these categorical variables with the response variable, fertility output:

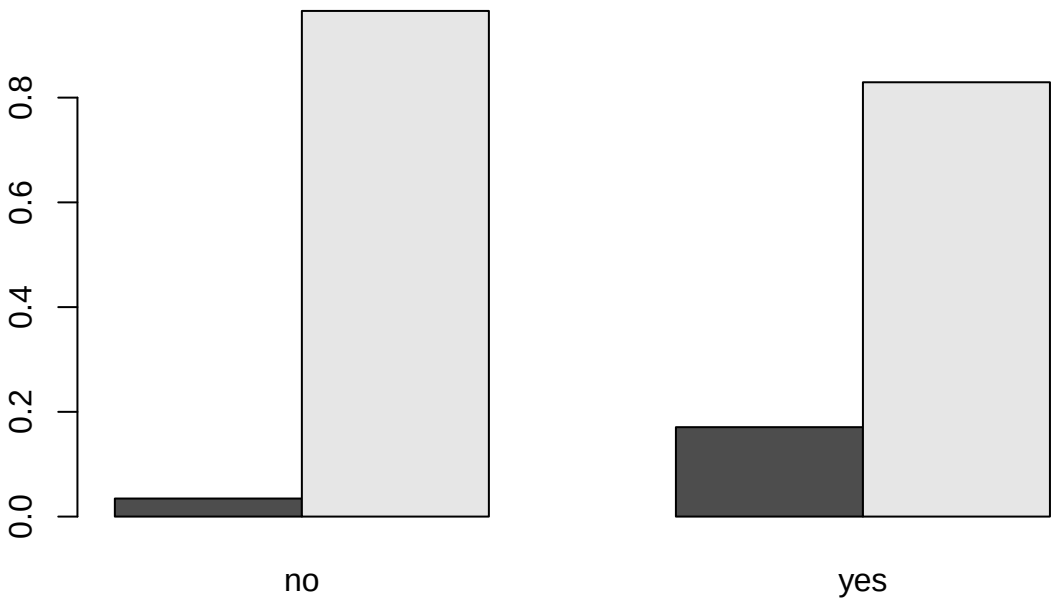
```
barplot(prop.table(table(fertility_train$Output, fertility_train$Season),margin = 2), beside = TRUE,mai
```



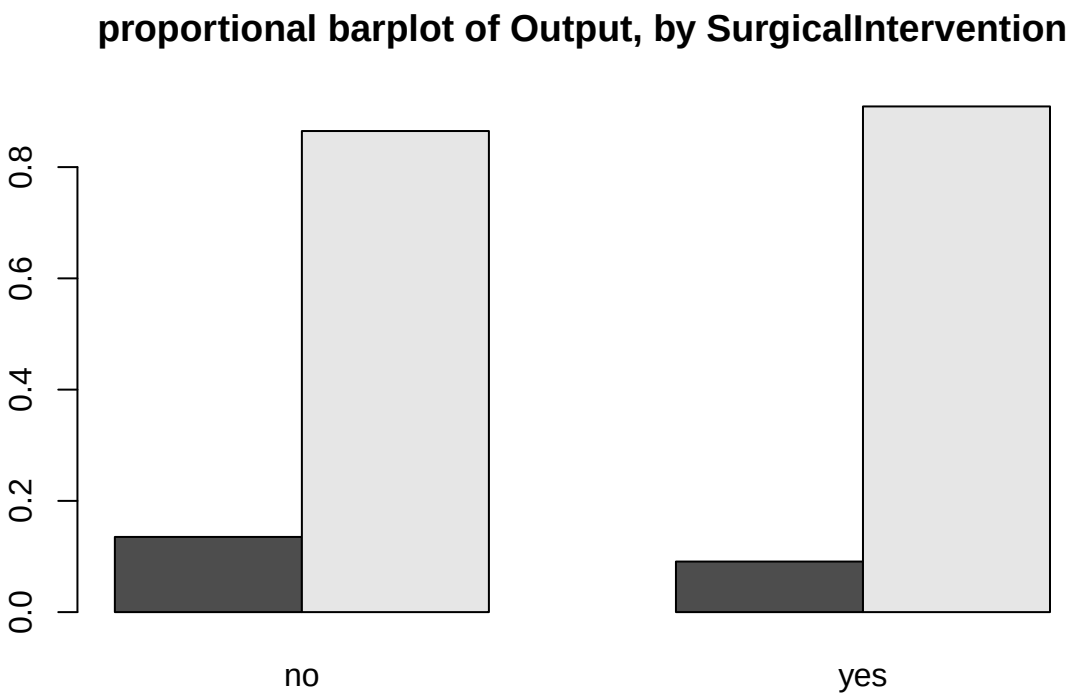

```
barplot(prop.table(table(fertility_train$Output, fertility_train$ChildishDisease),margin = 2), beside =
```



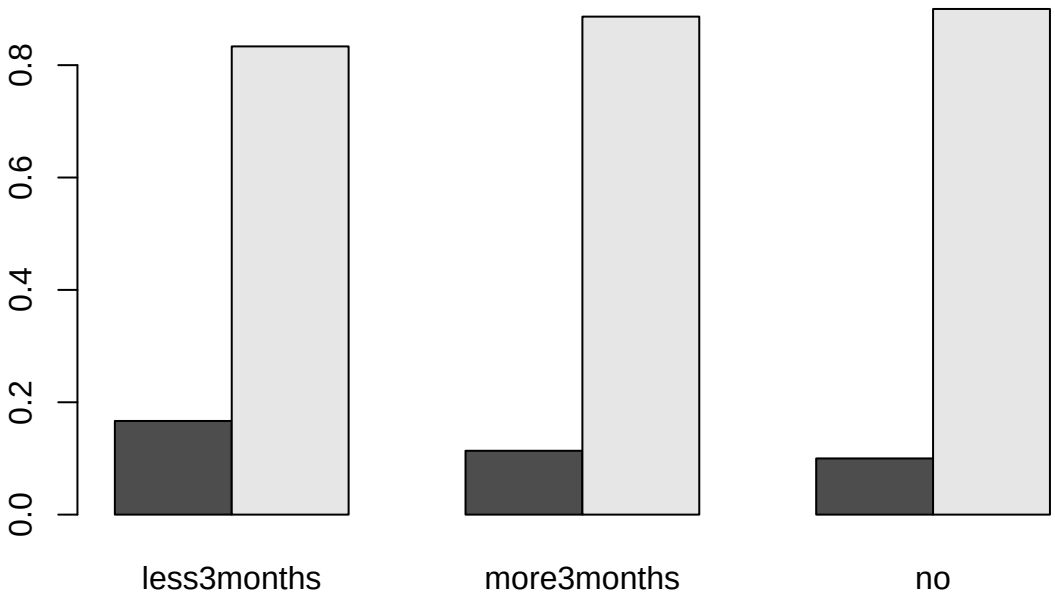

```
barplot(prop.table(table(fertility_train$Output, fertility_train$SeriousTrauma),margin = 2), beside = TI
```



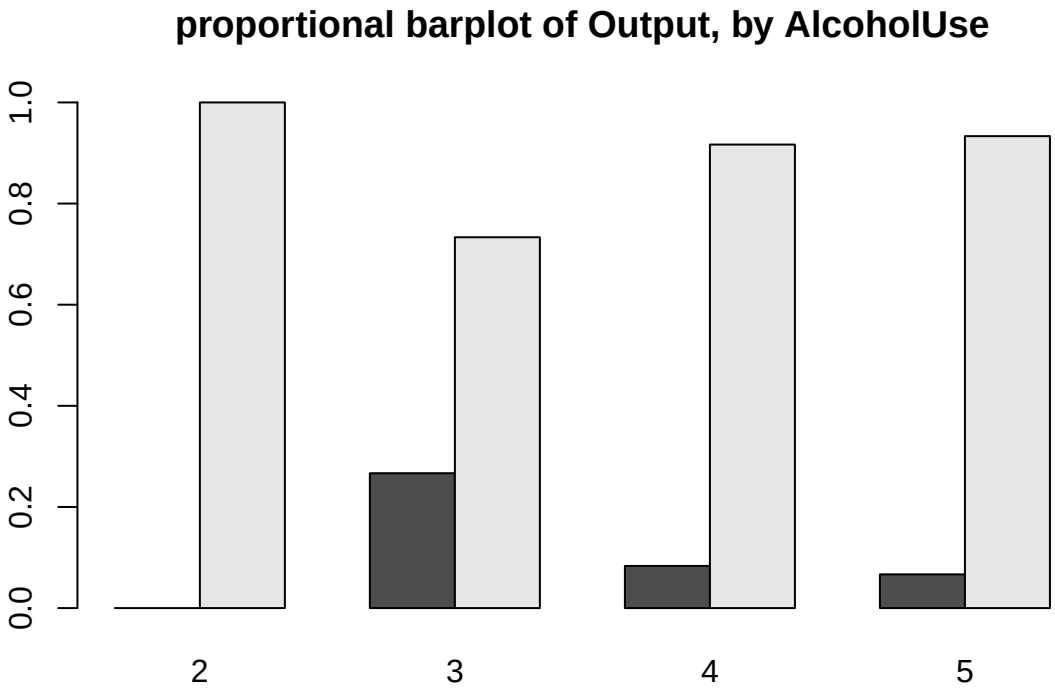

```
barplot(prop.table(table(fertility_train$Output, fertility_train$SurgicalIntervention),margin = 2), besi
```



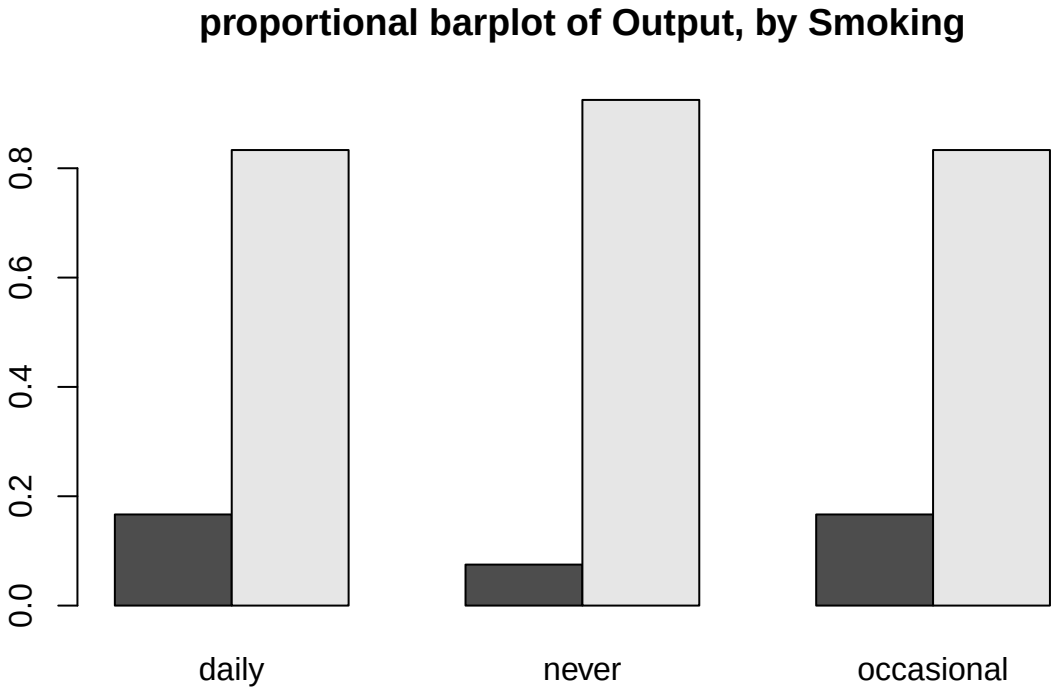

```
barplot(prop.table(table(fertility_train$Output, fertility_train$Fevers1year),margin = 2), beside = TRU
```




```
barplot(prop.table(table(fertility_train$Output, fertility_train$AlcoholUse),margin = 2), beside = TRUE
```




```
barplot(prop.table(table(fertility_train$Output, fertility_train$Smoking),margin = 2), beside = TRUE,ma
```



From the histograms above, there is evidence of a relationship between some of the categorical variables and the response variable, Output. In the proportional bar plot of output, by Season, the number of altered fertility output patients decreases from fall to summer and the number of normal fertility output patients increase simultaneously. Furthermore, the proportional bar plot of output, by ChildishDisease, seems to show that the number of altered fertility output patients increase when the patient has had a child disease and the number of normal fertility output patients seem to decrease at the simultaneously. Both the proportional bar plots of output, by SurgicalIntervention and Fevers1year, do not have a strong enough relationship to use as a classifiers of output. On the other hand, the number of altered fertility output patients decrease as alcohol

usage decreases. Additionally, the number of altered fertility output patients increase as the frequency of smoking increases. The strongest relationship seems to be predicting the output from the Serioustrauma variable which increases the number of altered outputs significantly if the patient has had serious trauma before.

From the observations made above, the best possible predictors include Season, ChildishDisease, Serious-Trauma, AlcoholUse, and Smoking.

Modeling

After picking the predictors for the fertility output, we look to create classifiers to predict if a patient's fertility output is altered or normal. The four classifiers that we will construct and analyze are binary logistic regression, linear discriminant analysis(LDA), Quadratic Discriminant Analysis(QDA), and classification trees. The classifiers will be constructed using a training set and the resulting models will be tested on the testing set. Aforementioned at the beginning of the paper, the training set has 70 subjects and the testing set has 29.

Binary Logistic Regression

The first classifier that we are going to construct is using binary logistic regression to model the fertility output. This model is able to use both quantitative and categorical variables so we will be classifying Output by Age, Season, ChildishDisease, SeriousTrauma, AlcoholUse, and Smoking. The first step is to fit the binary logistic model:

```
fertility.logit <- glm(factor(Output) ~ Age + factor(Season) + factor(ChildishDisease) + factor(Serious'
```

After fiting the binary logistic model, we then need to store the predicted probabilities from the logistic model and test it on the testing data set as follows:

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

Running the following code lets us know how the Output is ordered so we can establish a probability threshold:

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

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

A reasonable threshold is a probability of 0.5 and the classifying the test data set is below:

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

To see the final results of how the binary logistic classifier did on the testing data set, a confusion matrix is construct to do so:

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

```
##
## fertility.logit.pred altered normal
##              altered       0      1
##              normal        4     24
```

```
(4+1)/29
```

```
## [1] 0.1724138
```

```
1/25
```

```
## [1] 0.04
```

```
4/4
```

```
## [1] 1
```

The model performed on the test data with an overall error rate of 17.24% ((4+1)/29). For altered outputs, it had an error rate of 100% (4/4) which is higher than we would have liked. On the other hand, for the normal outputs, it had an error rate of 4% (1/25) which is substantially lower.

Linear Discriminant Analysis

In LDA, the only variable that can be used in the quantitative variable Age, which had somewhat of a relationship with fertility output in the bivariate analysis conducted in the EDA.

Similar to the previous classifier, we build the LDA on the training data set as follows:

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

After training the model on the training set, we now use the model to predict Output on the testing data set:

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

To visualize how the classifier performed on the testing data set, we analyze the confusion matrix below:

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

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

```
4/29
```

```
## [1] 0.137931
```

```
4/4
```

```
## [1] 1
```

```
0/29
```

```
## [1] 0
```

The LDA model performed on the test data set with an overall error rate of 13.79% (4/29). For the altered outputs, similar to the ligistic classifier, this model also had an error rate of 100% (4/4). However, for the normal outputs, the error rate went down to 0% (0/29) which is ideal.

Quadratic Discriminant Analysis

Similar to LDA, QDA can also only use quantitative variables which mean that we will use the predictor variable, Age, once again.

The process is almost identical to the LDA as we train the model on the training data set as follows:

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

After training the model on the training set, we now use the model to predict Output on the testing data set:

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

To visualize how the classifier performed on the testing data set, we analyze the confusion matrix below:

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

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

The QDA model performed on the test data set with the exact same results as the LDA with an overall error rate of 13.79% (4/29). For the altered outputs, this model also had an error rate of 100% (4/4). For the normal outputs, the error rate went down to 0% (0/29).

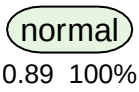
Classification Trees

The last classifier that we are doing to train and test is using classification trees. Like the binary logistic model, classification trees use both quantitative and categorical variables. We will be classifying the fertility output using the same variables as the binary logistic model: Age, Season, ChildishDisease, SeriousTrauma, AlcoholUse, and Smoking.

Similar to all the other classifiers, we first fit the classification tree onto the training data set and plot it:

```
fertility.tree <- rpart(factor(Output) ~ Age + factor(Season) + factor(ChildishDisease) + factor(Seriou

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



From the classification tree above, we can see that the classification stopped on the 0th iteration. This could be because that training data set is too small to split based on the variables above. The classification tree has classified that the training data set as a normal fertility output. There is an 89% chance of a patient having a normal fertility output and this composes 100% of the training data set.

We now use this classifier to predict output on the testing data set:

```
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
```

The classification tree has the same results as both of the discriminant analsysis classifiers with an overall error rate 13.79% (4/29). For the altered outputs, this model also had an error rate of 100% (4/4). For the normal outputs, the error rate is to 0% (0/29).

Final Recommendation

After testing all four types of classifiers available to us, the recommended classifier is LDA. We note that both QDA and the classification tree performed with the same results and error rate. However, it is in the benefit of the statistician to keep the model as simple as possible we recommend using the LDA classifier over the QDA classifier in this case. Furthermore, although the classification tree also yielded the same overall error rate, we would rather recommend the LDA classifier because the classification tree stopped on the zero iteration, predicting that all subjects in the data set were of normal output. However, that is not the case. On the other hand, the binary logistic classifier performed at higher overall error rate compared to the three classifiers mentioned above so we would recommend to not use this classifier on this data set.

Discussion

Constructing and training the four classifiers above gave us insight into the data set and predicting fertility output. The recommended classifier is the LDA model although the other three models performed very similar to the LDA.

A challenge that was presented in this paper was the lack of data in the sample. Only having 99 subjects hindered a better use of the classification tree which stopped on the 0th of iteration of the model, predicting all subjects to have a normal fertility output.

In the future, it is essential to gather a larger sample size to train so that the models can be trained to their max capacity which will make them a lot more accurate. The prediction variables did not have a strong enough relationship with fertility, so in the future, it would be beneficial to keep exploring other health and environmental factors that may help predict fertility. As fertility testing is a significant financial issue for many men around the world, it is important to continue to develop alternative way such as machine learning techniques to predict fertility.