Stat 228 Final Project: Male Fertility Analysis

Huda Saeed and
Marisa Papagelis

Spring 2022

M

TABLE OF CONTENTS

M

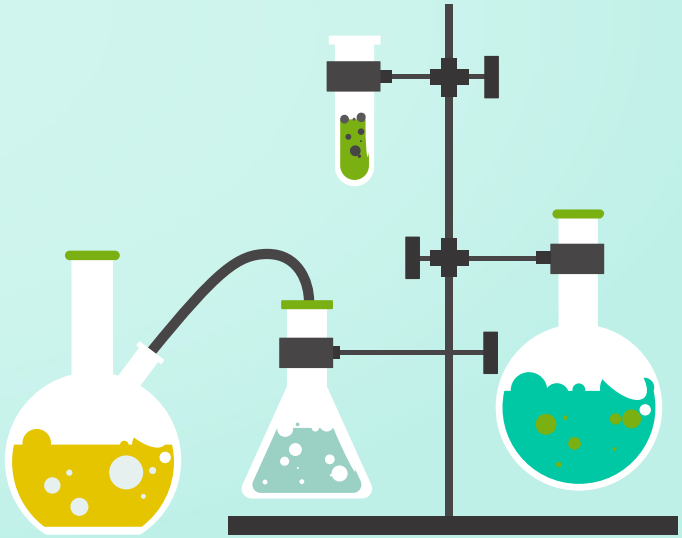
Introduction

Motivation, Data, Research Questions

M

Why are we interested in this topic?

- Joint interest in biostatistics & public health
- To create more conversation and less stigma around the topic
- Potentially identify variables that are associated with higher risks of infertility for further research (or lifestyle changes!)
- Ethics surrounding medical data collection



M

Our data set

General

- Dataset is from UCI Machine Learning (2013)
- 100 observations collected from volunteers
- 10 variables related to socio-demographic data, environmental factors, health status, and life habits

Variables

- **Response of Interest:** fertility (binary categorical)
- **Predictors:** age, season, childish diseases, accident/serious trauma, surgical intervention, high fevers, alcohol consumption frequency, hours sitting per day

M

Research Questions

1. *Which factors* contribute to best predicting male fertility?
2. *Are there differences* in male fertility based on socio-demographic, environmental, health and lifestyle factors?

H

Data Cleaning

02

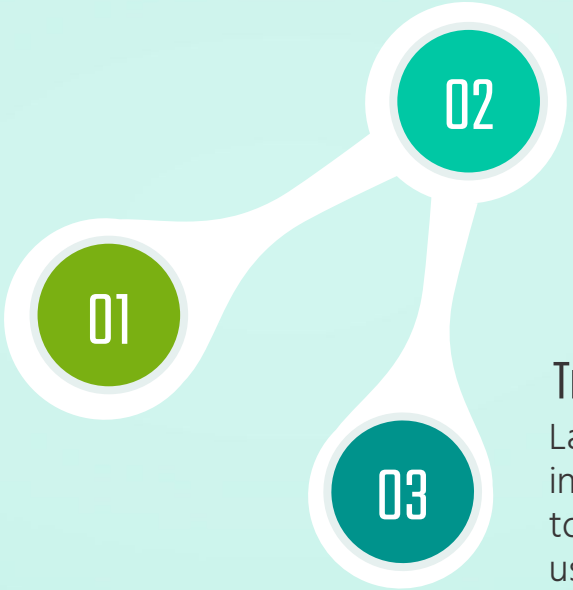
Missingness & multicollinearity,
variable creation, training/test

H

Initial Cleaning

Missingness & Multicollinearity

Our dataset did not contain any missing data, and most of our predictors and our response were categorical, so we didn't need to worry about multicollinearity



Variable Creation

We created a dummy variable (1/0) representing Altered/Normal for the boosting model which used a Bernoulli distribution

Training/Test Sets

Lastly, we divided our data into training and test sets to run our models on. We used 70% of our data for the training set and 30% for the test set *this form of CV suffers from randomness

H

Modeling & Classification

03

Regression
Bayesian
Tree-Based Methods

H

Our Plan

01

Logistic Regression

Automatic model selection to determine which predictors are the most significant and determine differences in male fertility

02

Multivariate Discriminant Analysis

Variable selection + model fit to determine which predictors are the most significant and determine differences in male fertility

03

Tree-Based Methods

CART, Bagging, Random Forest and Boosting classification methods to determine significant predictors + determine differences in male fertility

H

Logistic Regression

- Fitted model on training set
- Conducted stepwise regression with AIC and BIC criteria → both resulted in **no significant predictors**
- Conducted GOF test → the model fit the data adequately well (G(M,M0)=40.057,df=54,p=.92)
- Further analysis is not possible without predictors

```
Call:
glm(formula = Diagnosis ~ 1, family = binomial("logit"), data = data.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0255   0.5246   0.5246   0.5246   0.5246

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.9136     0.3571   5.359 8.36e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 53.713  on 69  degrees of freedom
Residual deviance: 53.713  on 69  degrees of freedom
AIC: 55.713

Number of Fisher Scoring iterations: 4
```


```
Call:
glm(formula = Diagnosis ~ 1, family = binomial("logit"), data = data.train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0255   0.5246   0.5246   0.5246   0.5246

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.9136     0.3571   5.359 8.36e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 53.713  on 69  degrees of freedom
Residual deviance: 53.713  on 69  degrees of freedom
AIC: 55.713

Number of Fisher Scoring iterations: 4
```


*same AIC/BIC output; no significant predictors, only intercept

H

Multivariate Discriminant Analysis

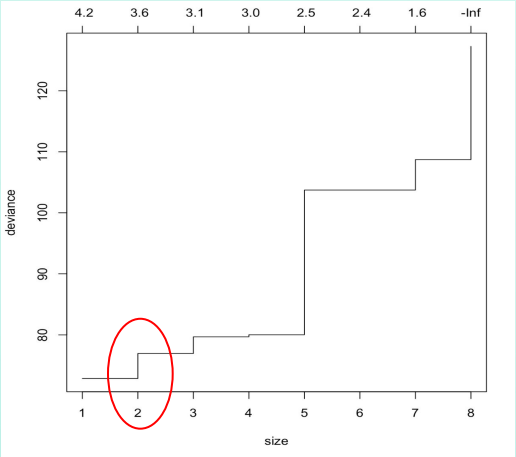
- Constant variance assumption failed via Box's M-test ($p<.01$)
  → cannot conduct LDA
- Through variable selection we found Age and Hours sitting per day (our only quantitative variables) were not significant
  → left with categorical predictors
  → violates normality assumption
  → **cannot conduct multivariate discriminant analysis**

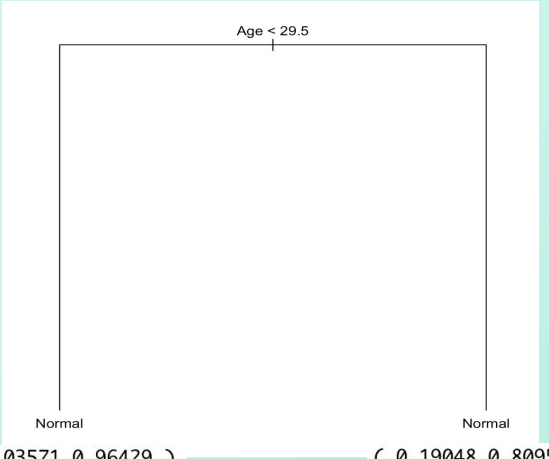
H

CART

- Fitted model on training set
- Through 10-fold CV we found that 2 was the optimal size → Age was the most significant predictor
- A threshold of 0.19 was chosen via CV
- Predictions on observations from test set → computed misclassification rate (0.5), sensitivity (1), AUC (0.278)

Two terminal nodes is the optimal size





H

Advanced Tree-Based Methods: Bagging, Random Forest, Boosting

- Fitted models on training set
- Threshold chosen through CV
- Predictions on observations from test set → computed misclassification rate, sensitivity, AUC

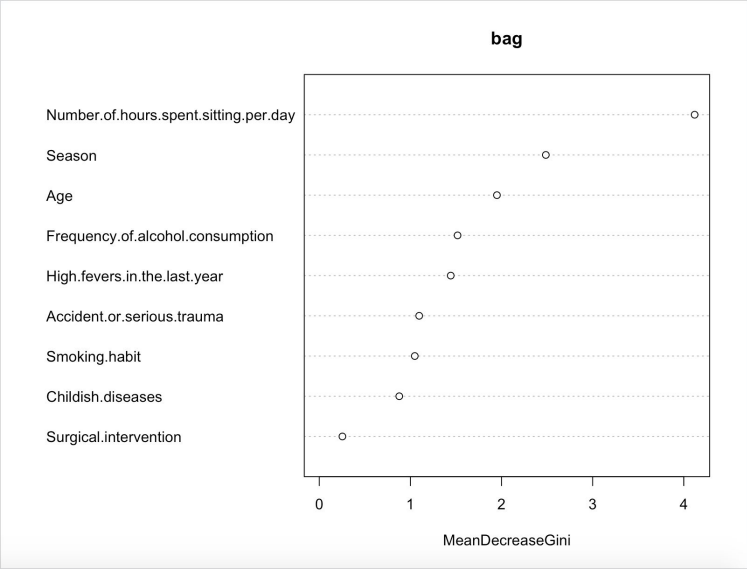
Method	Threshold	Misclassification	Sensitivity	AUC
Bagging	.44	0.067	0.667	0.648
Random Forest	.41	0.033	0.667	0.728
Boosting	.3	0.067	0.667	0.765

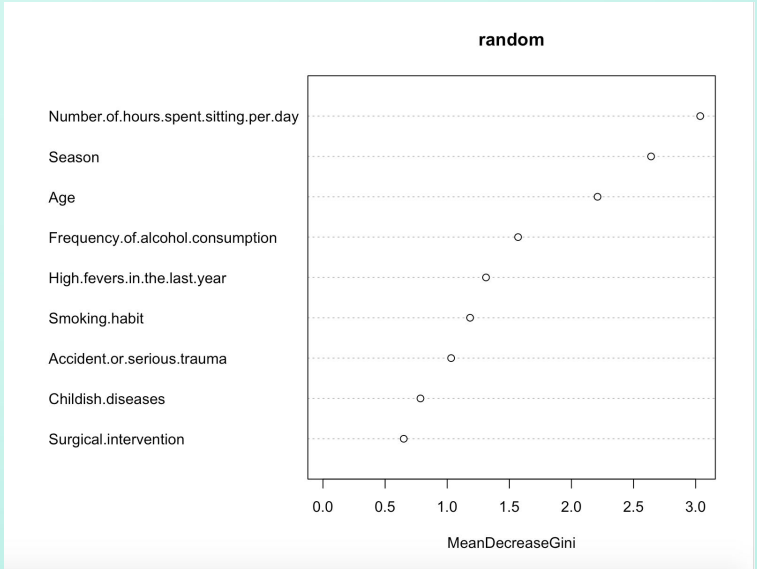
All three models had the same sensitivity
Boosting had the best overall performance (AUC)
Random forest had the best misclassification rate

H

Random Forest Variable Importance Plot

- varImpPlot() - measures variable importance in Random Forest models
- Excludes a variable in predicting each class of the response and measures the drop in % of accuracy
  → returns mean decrease in accuracy
  → variables with a higher drop are more important





H

Variable Relationships in Boosting

- boost.obj$var.monotone indicates whether variables have a positive (+1) negative (-1) or monotone (0) relationship with response
  - All predictors returned 0

H

Model Comparison

Misclassification, Sensitivity, AUC, Observations

M

Model Comparison

Method	Misclassification	Sensitivity	AUC
CART .3	0.1724138	0.667	0.8397
Bagging .3	0.03448276	0.667	0.7372
Random Forest .3	0.03448276	0.667	0.7949
Boosting .5	0.06896552	1	0.8846

- Our data set did not allow us to obtain metrics for logistic regression or LDA
- CART has a relatively high misclassification rate and a low overall performance likely due to underfitting of the model
- CART sensitivity is high due to over-identifying altered observations (because of high misclassification)
- The advanced tree-based methods are comparable
- We find **boosting** to be the most effective model due to its high AUC value at the slight expense of misclassification to random forest

M

Conclusion & Further Considerations

Data Ethics, Future Considerations & Improvements

M

Conclusion

1. *Which factors* contribute to best predicting male fertility?
    - From stepwise regression with the **logistic model** and var.monotone with **boosting** we conclude that **none of them** contribute best to predicting male fertility
    - From **bagging** and **random forest**, **number of hours sitting per day** appears to be most important
    - **CART** identified **age** as the most important (which was also ranked high in bagging and random forest)
2. *Are there differences* in male fertility based on socio-demographic, environmental, health and lifestyle factors?

- As shown by **CART**, **age** appears to be the most important factor
    - A greater proportion of participants aged 29.5 years or younger were fertile than those older than 29.5 years
    - This matches previous literature we found on male infertility

M

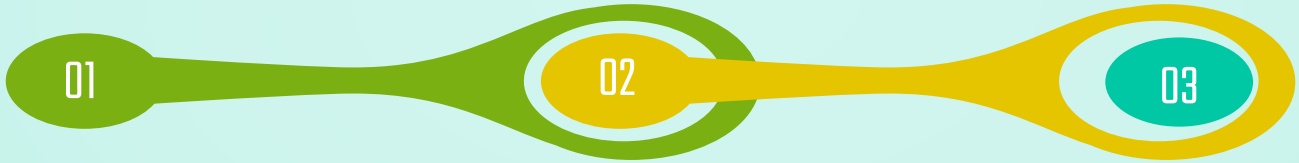
Future Considerations & Improvements

Larger Dataset

We can find a larger dataset with additional (quantitative) significant factors (i.e. tobacco use, overweight, etc.)

External Validity

Our dataset contains limited data, so our models are likely not externally valid and cannot be applied to outside datasets



More Recent Dataset

We can find a more recent dataset which would more accurately represent the population with current technology and medical findings

M

Data Ethics

Transparency

The data was originally collected by UCI in 2013; however, we obtained it from Kaggle and at least four edits/contributions have been made since then. It is not transparent what the edits were

Volunteer Bias

The data was collected from a sample of 100 volunteers, so there may be inherent differences between those who volunteered to participate and those who did not

Ownership

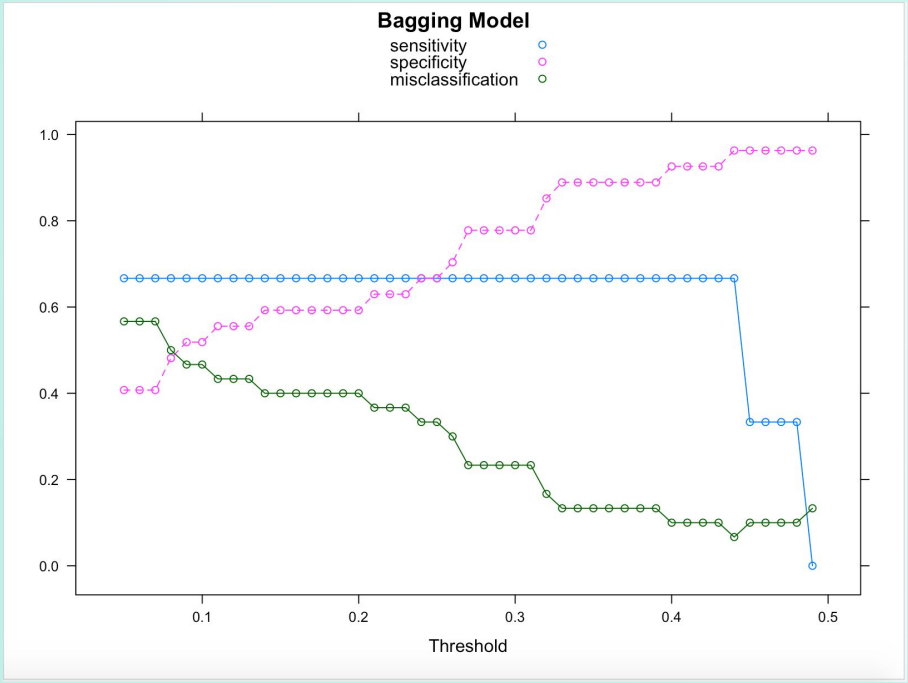
The data overview clearly states the initial owners of the data set with contact information as well as affiliations!!

M

M

Thresholds for Tree-Based Methods

- Sensitivity matters most to us because we are interested in determinants of infertility
- We plotted the sensitivity, specificity, and misclassification over a range of thresholds and chose the optimal one



Optimal threshold: .44

H