Read this carefully:

- This exam contains 20 questions. You have 105 minutes (1h45m) to solve it.
- The questions are sequential, and their order is randomized so **manage your time** properly. Plan to have an average of 5 minutes to answer each question.
- The final grade is the total number of points.
- The points per question are:

  No answer = 0 p

  Correct answer = 1 p

  Wrong answer = -0.33 p

Sample questions

1. The most challenging part in the implementation of data projects in companies is:
   a. Obtaining qualified IT staff to manage the databases
   b. Deal with the estimation of different models
   c. Find interesting questions
   d. Articulate problems in a way that can be addressed with big data methods
2. For which reason would we estimate a ridge regression model?
   a. To improve prediction error
   b. To perform variable selection
   c. To improve prediction and perform variable selection
   d. To obtain the optimal regularization parameter
3. How is the type II error different from the type I error?
   a. They are not different
   b. Type I error is very difficult to calculate while type II error is easy to calculate
   c. Type II error is very difficult to calculate while type I error is easy to calculate
   d. Type I error is normally set by the researcher while type II error is not under control of the researcher.
4. What is the tradeoff of Ridge regression?
   a. Generalization error vs Prediction error
   b. Bias vs Variance
   c. Prediction error vs Bias
   d. Prediction error vs Variance.
5. Which methods can we use to perform variable selection in the linear model?
   a. Lasso and Ridge
   b. Lasso
   c. Ridge
   d. Decision trees
6. Explain how K-fold cross validation works.
   a. Data is sampled randomly into a validation sample
   b. Data is randomly split with replacement into different bins
   c. Data is randomly split without replacement into different bins
   d. Data is sequentially allocated into different bins

7. Regarding the testing error:
    a. Testing error is calculated in the sample used to estimate the model
    b. Testing error is calculated in the sample not used to estimate the model
    c. Testing error is calculated in the sample used to regularize the model
    d. Testing error is a proxy for training error.

8. What is the subtree replication problem?
    a. When multiple trees are used to classify a case
    b. When multiple trees are used to replicate the results.
    c. When multiple splits are replicated at different points in the tree
    d. When multiple splits are used to replicate results at different points in the tree

9. What is the difference between Bagging and Random Forests?
    a. In bagging we average across multiple trees while in random forests we randomly select one of a set of multiple trees
    b. In bagging we average across multiple trees while in random forests we randomly select a subset of the predictors for each tree
    c. In bagging we randomly select a subset of the predictors for each tree while in random forests we perform sequential variable selection
    d. In bagging we randomly select a subset of the predictors for each tree while in random forests we average across multiple trees

We are interested in understanding which variables can help to predict heart disease. We estimated a simple linear model with penalization (Lasso) and obtained the following results.

**ANOVA**

| | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|
| Regression | 28,220 | 8 | 3,528 | 3,774 | ,000 |
| Residual | 274,780 | 294 | ,935 | | |
| Total | 303,000 | 302 | | | |

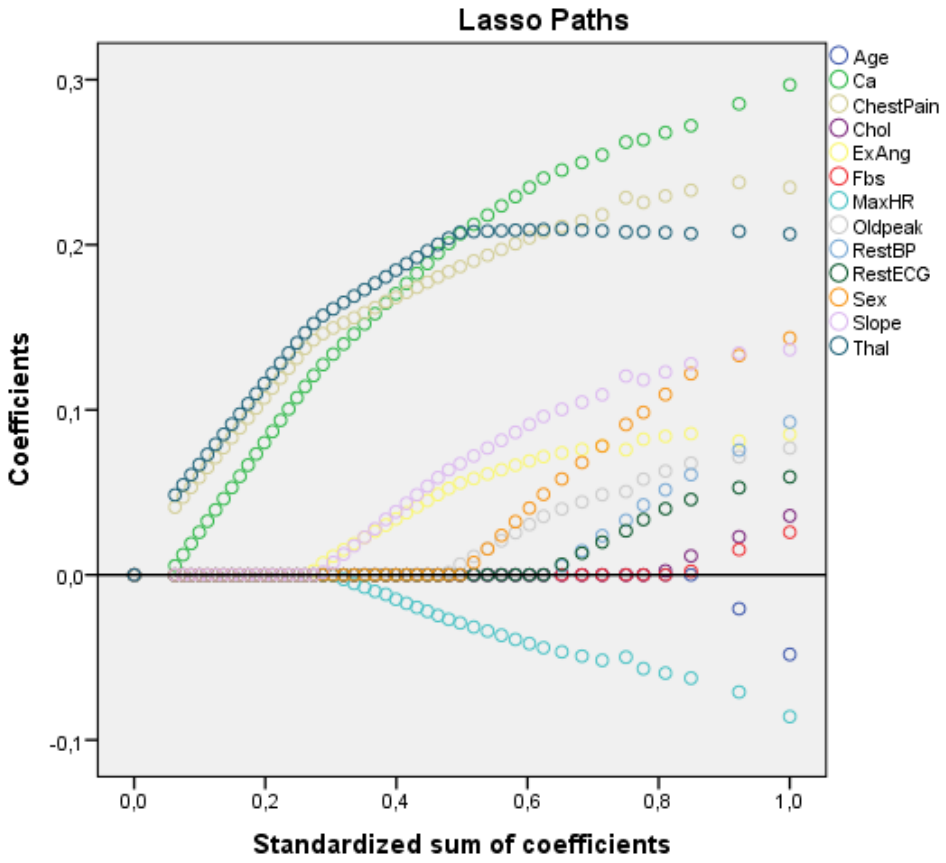
Dependent Variable: AHD
Predictors: Age Sex ChestPain RestBP Chol Fbs RestECG MaxHR ExAng Oldpeak Slope Ca Thal

**Model Summary**

| Multiple R | R Square | Adjusted R Square | Regularization "R Square" (1-Error) | Apparent Prediction Error | Expected Prediction Error | | |
|---|---|---|---|---|---|---|---|
| | | | | | Estimate[a] | Std. Error | N |
| ,653 | ,427 | ,411 | ,093 | ,907 | ,473 | ,029 | 303 |

Penalty ,920
Dependent Variable: AHD
Predictors: Age Sex ChestPain RestBP Chol Fbs RestECG MaxHR ExAng Oldpeak Slope Ca Thal

a. Mean Squared Error (10 fold Cross Validation).



**Coefficients**

| | Standardized Coefficients | | | | |
|---|---|---|---|---|---|
| | Beta | Bootstrap (1000) Estimate of Std. Error | df | F | Sig. |
| Age | ,000 | ,000 | 0 | . | . |
| Sex | ,000 | ,000 | 0 | . | . |
| ChestPain | ,041 | ,045 | 3 | ,822 | ,483 |
| RestBP | ,000 | ,000 | 0 | . | . |
| Chol | ,000 | ,000 | 0 | . | . |
| Fbs | ,000 | ,000 | 0 | . | . |
| RestECG | ,000 | ,000 | 0 | . | . |
| MaxHR | ,000 | ,005 | 0 | ,000 | . |
| ExAng | ,000 | ,006 | 0 | ,000 | . |
| Oldpeak | ,000 | ,000 | 0 | ,000 | . |
| Slope | ,000 | ,003 | 0 | ,000 | . |
| Ca | ,005 | ,030 | 2 | ,033 | ,967 |
| Thal | ,048 | ,048 | 3 | 1,009 | ,389 |

Dependent Variable: AHD

10. How many variables are relevant predictors?
    a. Ca, Thal, and Chestpain
    b. Ca, Thal, MaxHR, ExAng, Oldpeak, Slope, and Chestpain
    c. We can't say.
    d. None. They are all statistically insignificant.

11. What is the testing error?
    a. 0.403
    b. 0.907
    c. 0.093
    d. We can't say

Alternatively, we have decided to use a method which can be visually more appealing. To do this we fitted a decision tree and obtained the following results:

**Classification**

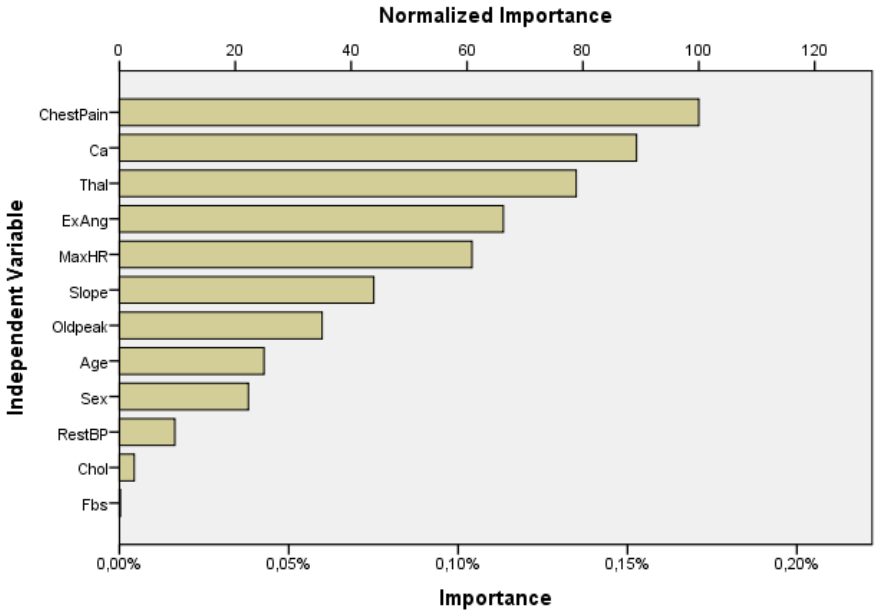
| Observed | Predicted | | |
|---|---|---|---|
| | No | Yes | Percent Correct |
| No | 154 | 10 | 93,9% |
| Yes | 58 | 81 | 58,3% |
| Overall Percentage | 70,0% | 30,0% | 77,6% |

Growing Method: CRT
Dependent Variable: AHD

**Risk**

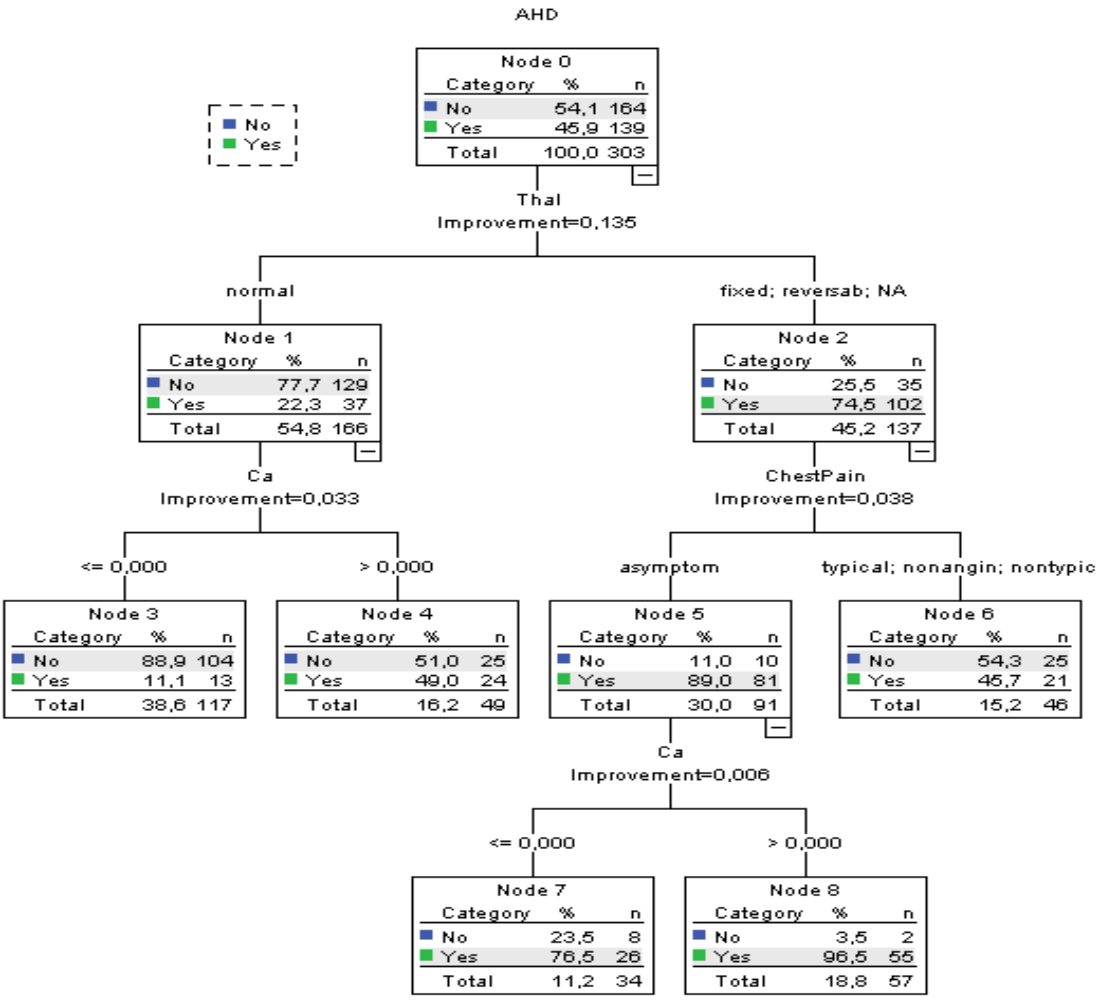
| Estimate | Std. Error |
|---|---|
| ,224 | ,024 |

Growing Method: CRT
Dependent Variable: AHD




Growing Method: CRT

Dependent Variable: AHD




12. How many nodes does the tree contain?
    a. 7
    b. 8
    c. 9
    d. We can't say
13. How many are terminal nodes?
    a. 3
    b. 4
    c. 5
    d. 6
14. What is the depth of the tree?
    a. 1
    b. 2
    c. 3
    d. 4

15. Which variables are more important in predicting heart disease?
    a. Ca, Thal, and Chestpain
    b. Ca, Thal, MaxHR, ExAng, Oldpeak, Slope, and Chestpain
    c. We can't say.
    d. None. They are all statistically insignificant.

16. What is the overall model fit?
    a. 0.224
    b. 0.776
    c. 0.583
    d. We can't say.

17. What is the overall classification error of a 3-way split that produces the above nodes?

    Node 1: N11=10,N12=3

    Node 2: N21=5,N22=1

    Node 3: N31=1,N32=5

    a. 30%
    b. 20%
    c. 25%
    d. We can't say

18. The principle of parsimony states that:
a) Models should be selected on the base of overall fit
b) Models should be selected on their simplicity
c) Both a) and b)
d) Neither a) or b)

19. A terminal node of a decision tree is a node with:
a) no incoming edge and two or more outgoing edges
b) one incoming edge and two or more outgoing edges
c) one incoming edge and no outgoing edges
d) one incoming edge and one outgoing edge.

20. Table lookup models are most fit for the cases when
a) There are many dimensions
b) There are few dimensions
c) We need to select a subset of variables
d) Neither a) or b)