[redacted]

Name_____Course_____

Be brief but show your reasoning (partial credit?). **Label your work and answers for each problem if on a different page.** Put your name on each piece of paper. You can use your notes, textbook, calculator, and computer and the internet to access course materials.

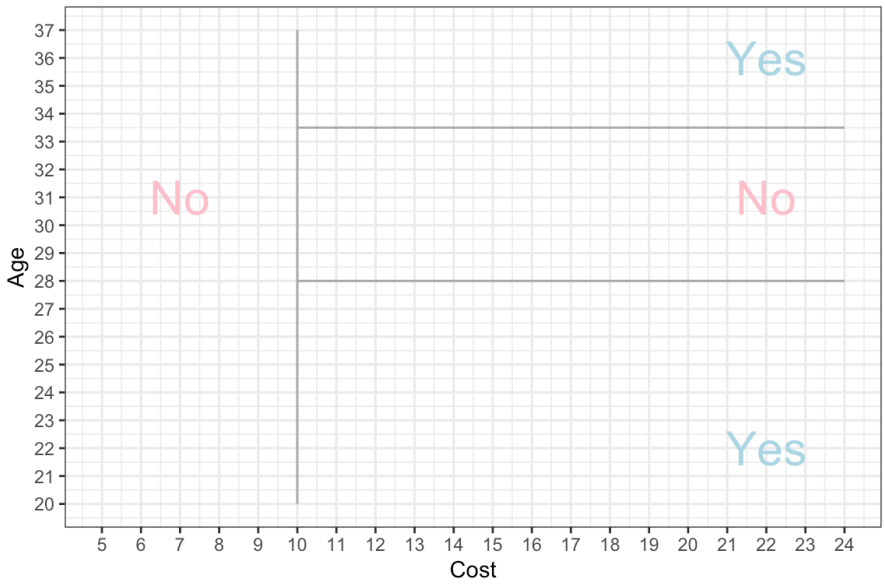
- Each problem is 20 points. Total points = 40. Time = 2 hr 30 min.

1 Insurance Predictions: Do by hand.

An insurance company wants to predict if a new customer will have a major operation within 10 years. They select 11 customers at random as training data to predict an operation based on the customer's age and their average annual cost of medical care.

Person	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**I**	**J**	**K**
Age	20	22	23	24	25	26	27	29	32	35	37
Average Cost	5	11	8	9	15	19	24	21	18	20	17
Had an Operation	N	Y	N	N	N	Y	Y	N	N	Y	Y

To classify future insured, the company partitions its predictor space (Age, Cost) as on the right.



Partition of Insured by Age and Average Annual Cost

(a) Plot the persons *using their label* on the above partition plot.

- Draw a classification tree that corresponds to this partition.

- At each internal node, state the splitting condition and threshold.
- At each terminal (leaf) node, state the predicted response.

(b) Fill in the following table with the prediction for each point based on your tree.

- Fill out the confusion matrix.
- Calculate the *training* classification rate from the confusion matrix.

Person	A	B	C	D	E	F	G	H	I	J	K
Age	20	22	23	24	25	26	27	29	32	35	37
Average Cost	5	11	8	9	15	19	24	21	18	20	17
Had an Operation	N	Y	N	N	N	Y	Y	N	N	Y	Y
Prediction?											

	Pred Yes	Pred No
Actual-Yes		
Actual-No		

Classification Rate?

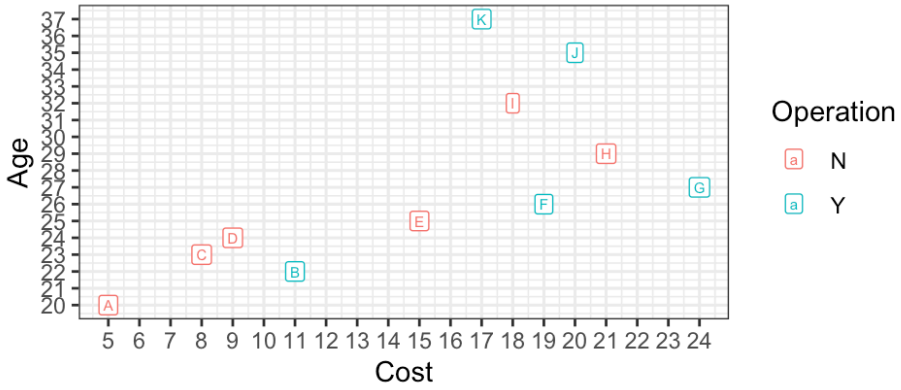
(c) A random forest is constructed for the same data. It has 3 trees and each is pruned to 1 split with 2 terminal nodes:

- The 1st tree sample is persons A, C, C, D, E, E, F, F, F, G, I. It splits on Cost.
- The 2nd tree sample is persons A, C, D, D, E, F, F, G, G, H, K. It splits on Age.
- The 3rd tree sample is persons A, A, D, D, E, E, F, G, J, K, K. It splits on Cost.

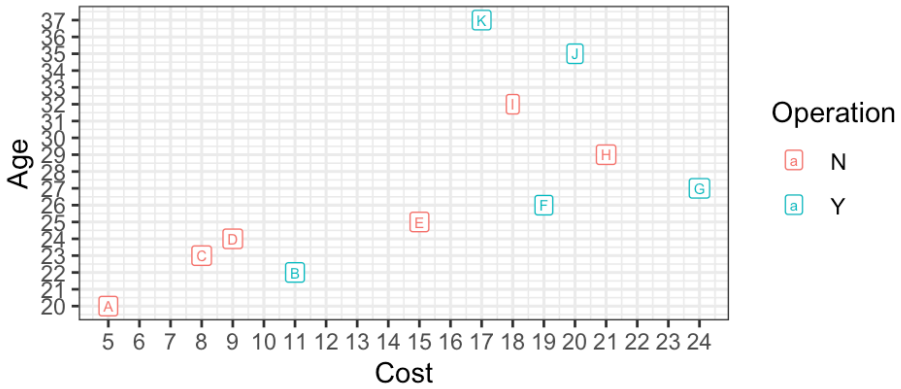
The following table summarizes the data for the forest.

Person	A	B	C	D	E	F	G	H	I	J	K
Age	20	22	23	24	25	26	27	29	32	35	37
Average Cost	5	11	8	9	15	19	24	21	18	20	17
Had an Operation	N	Y	N	N	N	Y	Y	N	N	Y	Y
Tree 1	1		2	1	2	3	1		1		
Tree 2	1		1	2	1	2	2	1			1
Tree 3	2			2	2	1	1			1	2

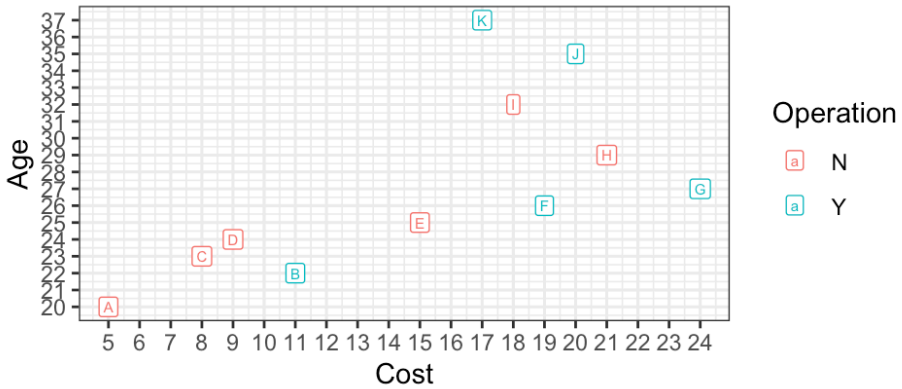
- In the following plots, put a number next to each point for how many times it is in the tree and line through the Out-Of-Bag points
- For each tree, draw a single partition for the given variable at the threshold which creates the terminal nodes as pure as possible.
- Below the three plots, for each tree, indicate the number of pure nodes (if any) for each tree and your calculation of the threshold value based on the closest points on either side of the threshold. You do **not** need to calculate the Gini impurity index.
- For each tree, plot a New customer with age 26 with an average cost of 18.
- Use this random forest to predict whether the New customer will have an operation and explain your reasoning for the prediction.



Tree 1



Tree 2



Tree 3

Number of pure nodes, Threshold, and Prediction for each tree?

The 1st tree:

The 2nd tree:

The 3rd tree:

Overall Prediction for the new Customer and Rationale:

(d) (STAT-627 only) Calculate the *prediction* error rate of the random forest in question (c), using out-of-bag (OOB) data.

- Using your thresholds from part (c), fill in the following table with the predictions for only the OOB persons in each tree.
- Calculate the overall OOB prediction for each OOB person.
- Indicate if each overall OOB prediction is an error or not (1, 0).
- Calculate the prediction error rate.
- What would you recommend to improve the prediction error rate?

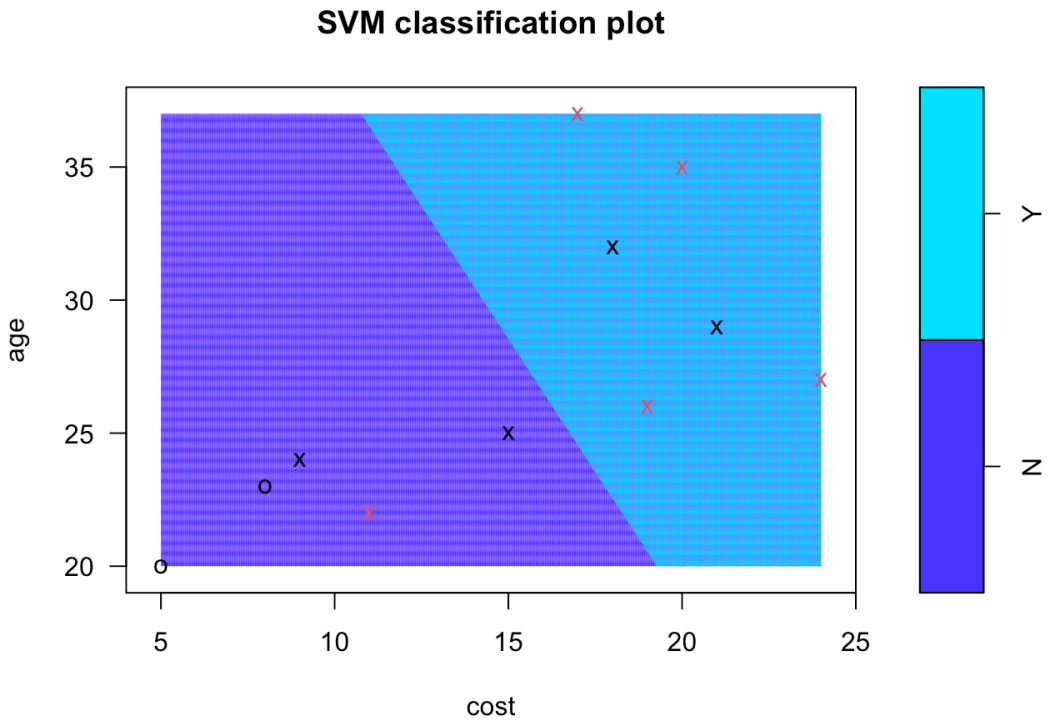
Person	A	B	C	D	E	F	G	H	I	J	K
Tree 1											
Tree 2											
Tree 3											
————	—	—	—	—	—	—	—	–	—	—	—
Predicted											
Had an Operation	N	Y	N	N	N	Y	Y	N	N	Y	Y
Error?											

OOB Prediction Error Rate?

(e) Is there a *maximal margin* classifier for this dataset? Explain your answer. Consider drawing some lines on the scatterplot you created in (a).

(f) Consider the following plot from an SVM classifier.

- How many support vectors are there?
- What type of kernel is being used based on the plot and why?
- What is the training classification rate?



Name: _____Course:_____

2 College Scorecard Data

The US Department of Education collects data from every "college" level institution in America and makes a lot of data available under the College Scorecard.

- This question uses a curated extract of the college scorecard data. The variable names and definitions are at the end.
- This dataset has 23 variables of data on 1,695 four-year colleges.
- This dataset is on Canvas or at "https://raw.githubusercontent.com/AU-datascience/data/main/427-627/college_scorecard_extract_sep_2023.csv".

We want to predict the Endowment of a new colleges given the other variables as potential predictors.

In the following steps, build models to predict the College Endowment (`ENDOWBEGIN`) and use $K = 10$-fold cross-validation to tune and evaluate predictive performance with **`set.seed(123)` as appropriate.**

- For each problem, describe your approach, your R code, the most important results, and your interpretation of the results.
- You may write your responses on this document by hand after running code in R and/or submit a file on Canvas with your approach, code, results, and interpretation of results.

2.1 Multiple Linear Regression Regularization

- Load the data and assign the name `college` to it. Get rid of any records with `NA`s and divided `ENDOWBEGIN` by 1 million to reduce the scale. Glimpse `college`.

````
        ```{r}
        #| message: false
        library(tidyverse)
        college <- read_csv("https://raw.githubusercontent.com/AU-
datascience/data/main/427-627/college_scorecard_extract_sep_2023.csv")
        college <- na.omit(college)
        college$ENDOWBEGIN <- college$ENDOWBEGIN/1000000
        #college <-
read_csv("./data/college_scorecard_extract_sep_2023.csv")
        #glimpse(college)
        ```
````

- Fit a multiple linear regression of Endowment ($ENDOWBEGIN$) on all the other variables as a full model.
- How many predictors appear important with a p value less than 0.1?

- Which of those increase the endowment?

Do any of the variables have a high generalized variance inflation factor GVIF? If any, which ones and do they make sense as having high GVIF given the other variables?

Refit a reduced model without `MN_EARN_WNE_P10` and `PCT_WHITE`. Are there any changes in significant variables?

- Check the GVIF again and comment on any changes.
- Create a new data frame with the variables below (you can use the following code).
    - Remove the rows with `REGION = "Outlying Regions"`.
    - Convert all character variables to factors.

````
        ```{r}
        college |>
          dplyr::select(
             ENDOWBEGIN, ADM_RATE, AVGFACSAL, CONTROL,FIRST_GEN,
GRAD_DEBT_MDN, PCIP27,
        PCT_ASIAN, REGION
          ) |>
          filter(REGION != "Outlying Regions") |>
          mutate(across(where(is.character), as.factor)) ->
        college2
        ```
````

- Use the {boot} package with `college2` to report the prediction MSE for a full model (`ENDOWBEGIN` on the other data) based on K-10 fold cross-validation adjusted deviance.

2.2 Regularization via Shrinkage

Use LASSO with cross validation to model `ENDOWBEGIN` on the other variables in `college2` and find the best lambda.

- Create model matrices for `X` and `Y`.
- Use `set.seed(123)` for the cross validation.
- Plot the result of the cross validation.
- Show the result and identify whether `lambda-min` or `lambda.1se` has fewer non-zero variables?
- Show the coefficients for `lambda.1se` and discuss which were driven to zero if any.
- Do any of the +/- signs of the coefficients for the variables surprise you?

What is the cross-validated predicted MSE from the model?

2.3 Principal components.

Calculate the principal components using the `X` model matrix you created earlier, with scaling, and show the scree plot.

- Interpret the scree plot

Given the scree plot, choose to create **either** a PCR **or** a PLSR model.

Create the model with scaling and K=10 fold cross-validation. (Use seed 123)

How many principal components are

- Needed to explain 90% of the total variation *among X-variables*?
- Needed to explain 40% of the total variation *of the response*, `ENDOWBEGIN`?
- What is the optimal number of PCs based on adjusted Cross-Validation RMSEP?
- What is the adjusted Cross Validation MSEP for the optimal number of PCs?
- Show the validation plot.

2.4 Summary

Create a summary table showing the method, the MSE, and the number of predictors.

- Recommend a model for predicting `ENDOWBEGIN` for new observations and explain your choice.

Method	Predicted MSE	Number of Predictors
Linear Model (Reduced)		
LASSO lambda.1se		
PCR		
PLSR		

2.5 Classification with SVM (Optional Extra Credit 4 points)

We now want to predict whether a new college is Private or Public based on the data in `college2`.

Tune a Support Vector Machine model to find the best cost and kernel.

- Use the range of costs in `seq(4.0, 6.0, 0.25)` and the linear and radial kernels.

What is the best cost value and the best kernel and the cross-validated error rate?

Create a model with the best parameters

- How many support vectors are there?

Plot the results looking at `ADM_RATE` and `AVGFACSAL`.

- Comment on the plot

2.5.1 College Scorecard Data

Variable	Definition
ADM_RATE	Admission Rate
AGE_ENTRY	Average age of entry
AVGFACSAL	Average Faculty Salary
CONTROL	Public or Private Non-Profit
COSTT4_A	Cost of an Academic Year
ENDOWBEGIN	Endowment at the Beginning of the year
FEMALE	Percent Female Students
FIRST_GEN	Percent First Generation Students
GRAD_DEBT_MDN	Median Debt at Graduation
LOCALE	City, Suburban, Town or Rural
MD_EARN_WNE_P10	Median Earnings 10 years after enrollment
MN_EARN_WNE_P10	Mean earnings 10 years after enrollment
PCIP14	Percent Engineering Degrees
PCIP27	Percent Math Stat Degrees
PFTFAC	Percent Full Time Faculty
PCT_ASIAN	Percent Asian in Home Zip Code
PCT_BLACK	Percent Black in Home Zip Code
PCT_WHITE	Percent White in Home Zip Code
PCT_HISPANIC	Percent Hispanic in Home Zip Code
PCTPELL	Percent with Pell Grant
REGION	Location in United States
SAT_AVG	Average SAT Score
UGDS	Total Undergraduates