STK2100
Oblig 2

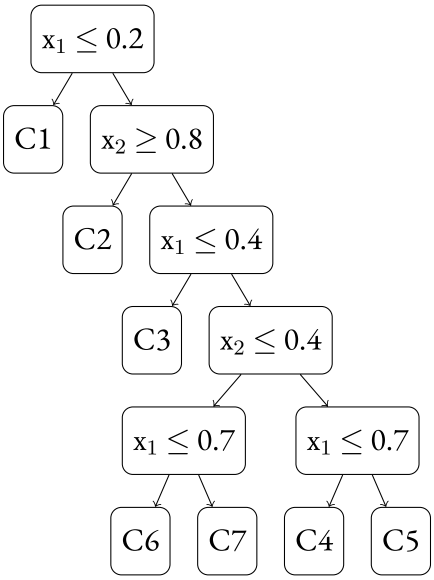
Egil Furnes
Studentnummer: 693784

Problem 1

(a)

The divison of covariate space produced by regression trees must first and foremost have partitions that are either *horisontal* or *vertical* and with *rectangular* particions. Therefore, figure A could have been produced by a regression tree, the others have not.

(b)



(c)

$$f(x_1 = 0.6, x_2 = 0.6) \Rightarrow C_4$$
$$f(x_1 = 0.1, x_2 = 0.6) \Rightarrow C_1$$
$$f(x_1 = 0.6, x_2 = 0.1) \Rightarrow C_6$$

Problem 2

(a)

(i)

```r
library(tidyverse)
library(readr)
library(caret)
library(Metrics)
set.seed(1705)

wage <- read.csv("https://www.uio.no/studier/emner/matnat/math/
    STK2100/v25/oblig/wage.csv",
           header=TRUE)

ind <- createDataPartition(wage$wage, p=.7, list = FALSE)
train <- wage[ind, ] %>% as_tibble()
test <- wage[-ind, ] %>% as_tibble()
```

(ii)

```r
lm1 <- lm(wage~., data = wage)
lm1 %>% summary()
```

```
Call:
lm(formula = wage ~ ., data = wage)

Residuals:
     Min       1Q   Median       3Q      Max
-100.303  -18.682   -3.311   13.496  211.086

Coefficients:
                              Estimate Std. Error
(Intercept)                 -2.399e+03  6.162e+02
year                         1.235e+00  3.073e-01
age                          3.092e-01  5.866e-02
maritlUnmarried             -1.531e+01  1.437e+00
raceOther                    1.596e+00  3.081e+00
raceWhite                    5.145e+00  2.142e+00
education2. HS Grad          7.518e+00  2.363e+00
education3. Some College     1.806e+01  2.512e+00
education4. College Grad     3.094e+01  2.532e+00
education5. Advanced Degree  5.358e+01  2.790e+00
jobclass2. Information       3.639e+00  1.324e+00
```

```
health2. >=Very Good         6.587e+00  1.421e+00
health_ins2. No             -1.756e+01  1.404e+00
                            t value Pr(>|t|)
(Intercept)                  -3.894 0.000101 ***
year                          4.018 6.02e-05 ***
age                           5.271 1.45e-07 ***
maritlUnmarried             -10.652  < 2e-16 ***
raceOther                     0.518 0.604490
raceWhite                     2.402 0.016357 *
education2. HS Grad           3.182 0.001478 **
education3. Some College      7.189 8.21e-13 ***
education4. College Grad     12.218  < 2e-16 ***
education5. Advanced Degree  19.202  < 2e-16 ***
jobclass2. Information        2.749 0.006009 **
health2. >=Very Good          4.636 3.71e-06 ***
health_ins2. No             -12.510  < 2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.02 on 2987 degrees of freedom
Multiple R-squared:  0.3381, Adjusted R-squared:  0.3354
F-statistic: 127.1 on 12 and 2987 DF,  p-value: < 2.2e-16
```

Looking at the `summary()` of `lm1` it seems that all covariats except from `raceOther` is significant at a `0.05` level. Looking at the adjusted R^2 the model explains `0.3354` of the variation in the predicted variable `wage`.

(iii)

```r
pred1 <- predict(lm1, newdata = test)
mean((test$wage-pred1)^2)
```

```
[1] 1202.368
```

(b)

(i)

```r
library(splines)

gam1 <- lm(
  wage ~ ns(year, df = 4) + ns(age, df = 4) + maritl + race +
     education + jobclass + health + health_ins,
  data = train
```

```
)

gam1 %>% summary()
```

```
Call:
lm(formula = wage ~ ns(year, df = 4) + ns(age, df = 4) + maritl +
    race + education + jobclass + health + health_ins,
    data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-102.730  -18.581   -3.058   14.034  210.536

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                     61.205      6.282   9.742  < 2e-16 ***
ns(year, df = 4)1                8.273      3.953   2.093   0.0365 *
ns(year, df = 4)2                5.304      3.335   1.590   0.1119
ns(year, df = 4)3                9.105      4.764   1.911   0.0561 .
ns(year, df = 4)4                5.684      2.746   2.070   0.0386 *
ns(age, df = 4)1                31.679      4.536   6.984 3.83e-12 ***
ns(age, df = 4)2                15.585      4.547   3.428   0.0006 ***
ns(age, df = 4)3                38.593     10.875   3.549   0.0004 ***
ns(age, df = 4)4                 9.597      8.579   1.119   0.2634
maritlUnmarried                -13.052      1.746  -7.474 1.13e-13 ***
raceOther                        2.944      3.653   0.806   0.4204
raceWhite                        4.818      2.481   1.942   0.0523 .
education2. HS Grad              8.515      2.778   3.065   0.0022 **
education3. Some College        19.749      2.958   6.677 3.12e-11 ***
education4. College Grad        30.398      2.953  10.294  < 2e-16 ***
education5. Advanced Degree     52.348      3.248  16.118  < 2e-16 ***
jobclass2. Information           3.893      1.549   2.514   0.0120 *
health2. >=Very Good             5.493      1.676   3.278   0.0011 **
health_ins2. No                -16.847      1.645 -10.240  < 2e-16 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.37 on 2083 degrees of freedom
Multiple R-squared:  0.3503
Adjusted R-squared:  0.3447
F-statistic: 62.4 on 18 and 2083 DF, p-value: < 2.2e-16
```

(ii)

Splines are used to model relationships between predicting and predictor variables that are smooth and continuos, or in other words, numerical and quantitative. It does not make sense to use those for qualitative variables, as these have no meaningful rank-order relationships, as quantitative ones does.

(iii)

That depends on the nature of the relationship between the covariate and the predicted variable. One could think that `wage` across `age` has a more exponential relationship before plateauing. On the other hand, `year` and `wage` might have a more linear relationship. Therefore, using splines only on `age` might make more sense.a

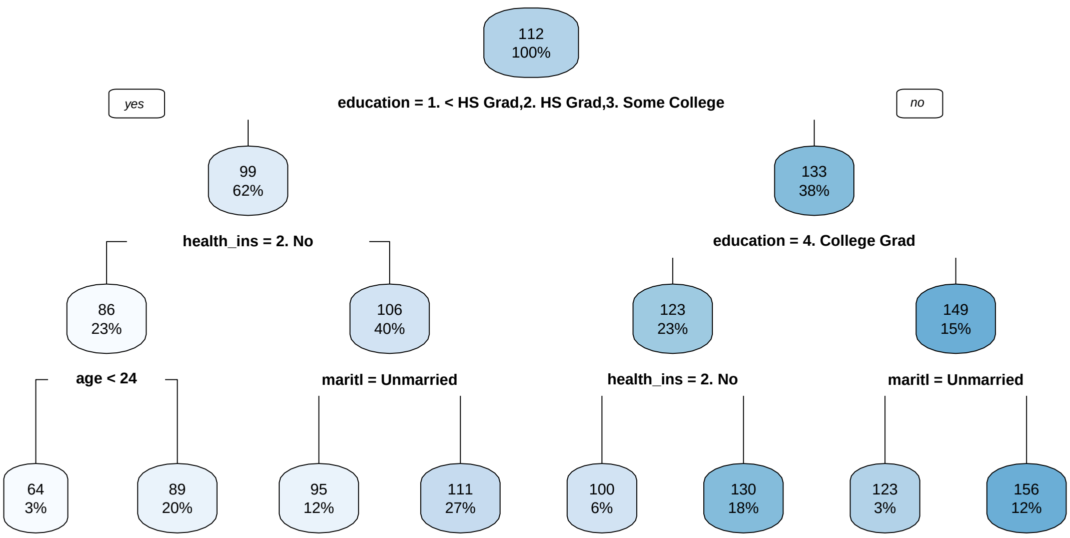
(iv)

Both `raceOther` and `raceWhite` are still not significant, and some of the splines are not neither.

(c)

(i)

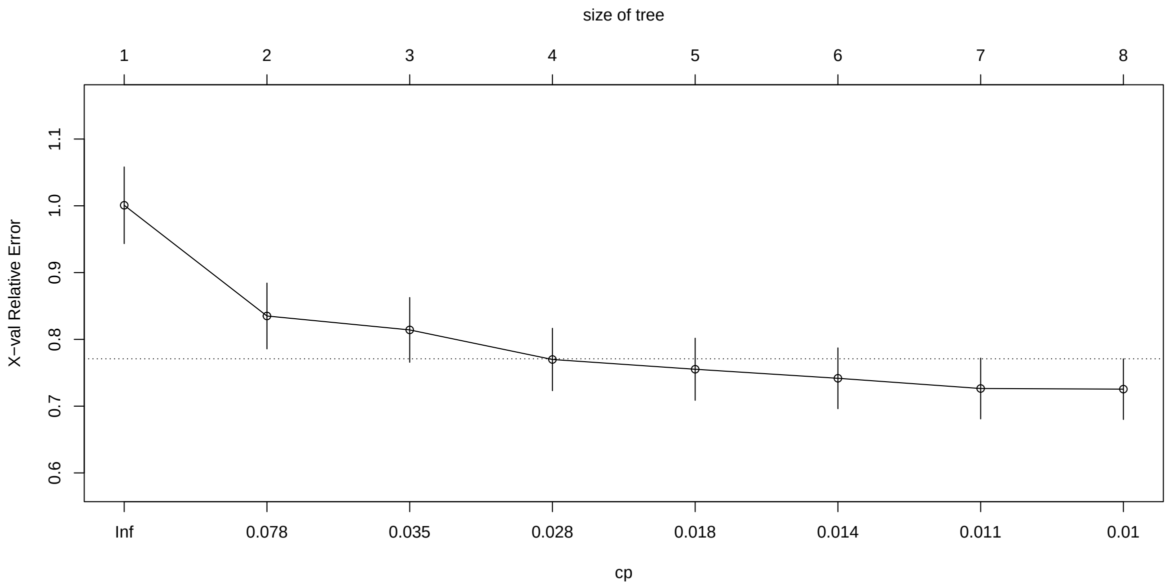
```
library(rpart)
library(rpart.plot)
library(caret)

tree1 <- rpart(wage~., data = train, method = "anova")
rpart.plot(tree1)
```

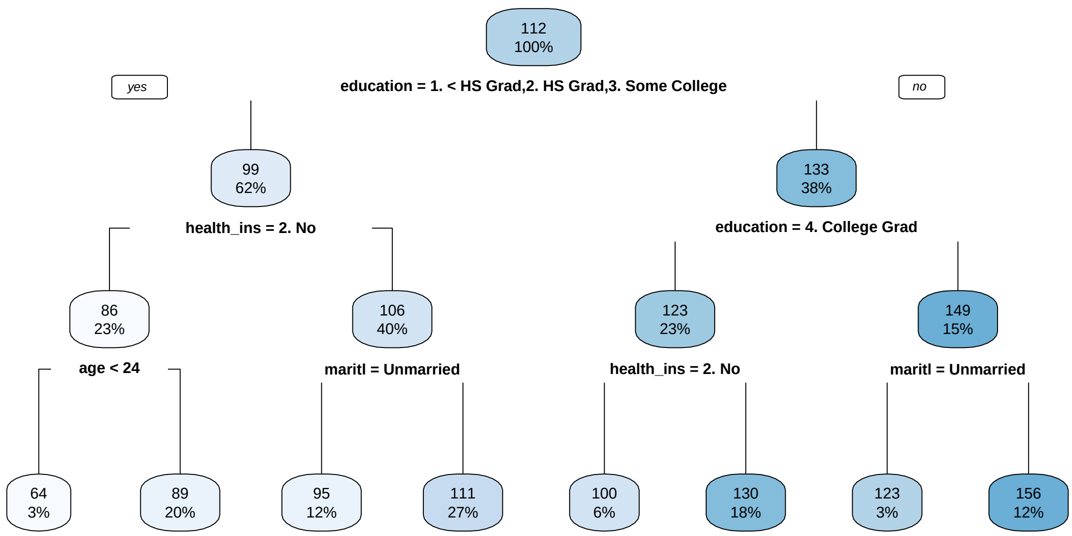


The following covariates contributes to the model, `education, health, age, maritl, health_ins`. Unlike the linear and additive models, the regression tree captures interactions and threshold effects automatically but may result in a less smooth and less stable predictions.

(ii)

```r
plotcp(tree1)
opt <- tree1$cptable[which.min(tree1$cptable[,"xerror"]), "CP"]
tree2 <- prune(tree1, cp=opt)
rpart.plot(tree2)
```





The pruned tree uses the same covariates in the model, and in fact has not changed since before pruning.

(iii)

```r
pred3 <- predict(tree2, newdata = test)
mse(test$wage, pred3)
```

```
[1] 1320.415
```

Purely based on mean squared error I would pick the one with the lowest `mse`, which in this case is the linear model.

Problem 3

(a)

(i)

```r
library(tidyverse)
library(caret)
set.seed(1705)

vert <- read.csv("https://www.uio.no/studier/emner/matnat/math/STK2100/v25/oblig/vertebral-column.csv",
                 header = TRUE) %>% as_tibble()

ind <- createDataPartition(vert$class, p = 2/3, list = FALSE)
train <- vert[ind, ]
test <- vert[-ind, ]
```

(ii)

```r
logit1 <- glm(class ~ ., data = train, family = "binomial")
summary(logit1)
```

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

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  15.833522   4.468339   3.543 0.000395 ***
pelvInc     -34.535054  45.811455  -0.754 0.450938
pelvTilt     34.627536  45.821240   0.756 0.449824
lumbLord     -0.008563   0.029358  -0.292 0.770526
SacrSl       34.437412  45.810809   0.752 0.452213
pelvRad      -0.118640   0.032837  -3.613 0.000303 ***
degrS         0.149402   0.026748   5.586 2.33e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 264.88  on 206  degrees of freedom
Residual deviance: 117.68  on 200  degrees of freedom
AIC: 131.68

Number of Fisher Scoring iterations: 8
```

Among the covariates, the `Intercept` and `pelvRad` and `degrS` is highly significant at a ≈0.

(iii)

```
prob4 <- predict(logit1, newdata = test, type = "response")
pred4 <- ifelse(prob4 > 0.5, 1, 0)
error4 <- mean(pred4 != test$class)
error4
```

```
[1] 0.1650485
```

(b)

(i)

```
library(MASS)
lda1 <- lda(class ~ ., data = train)
```

(ii)

```
pred5 <- predict(lda1, newdata = test)$class
error5 <- mean(pred5 != test$class)
error5
```

```
[1] 0.2038835
```

It seems that the mean classification error is lower for the logistic regression model compared to the linear discriminant analysis.

(c)

(i)

```
qda1 <- qda(class ~ ., data = train)
```

(ii)

```r
pred6 <- predict(qda1, newdata = test)$class
error6 <- mean(pred6 != test$class)
error6
```

```
[1] 0.1941748
```

The quadratic discriminant analysis is now slightly better than the linear one, but still worse of than the normal logistic regression model.

(d)

(i)

```r
library(nnet)

train_std <- train %>% mutate(across(-class, scale), class = as.
   factor(class))
test_std <- test %>% mutate(across(-class, scale), class = as.factor
   (class))
nn1 <- nnet(class ~ ., data = train_std, size = 5, decay = 0.01,
   maxit = 1000, trace = FALSE)
```

(ii)

```r
pred7 <- predict(nn1, newdata = test_std, type = "class")
error7 <- mean(pred7 != test_std$class)
error7
```

```
[1] 0.1747573
```

The one-layer neural networks mean classification error ranks second now, behind logistic regression but ahead of both the discriminant alaysis models.