Tarefa 4 - Classificação

# Introdução

Para este estudo, temos um banco de dados que possui variáveis categóricas e numéricas com informações sobre pacientes com câncer. Algumas variáveis categóricas não estavam definidas como fatores, assim como algumas das numéricas estavam classificadas em categorias. Portanto, se fez um ajuste inicial de forma que variáveis categóricas fossem reconhecidas como fatores e as numéricas como números, utilizando as funções **as.factor()** e **as.numeric()**.

As variáveis categóricas são: \* Paciente em remissão (**remission**): Níveis são 1 e 2, consideramos 1 = “Sim” e 2 = “Não” e recodificamos para 1 e 0, respectivamente; \* Paciente é casado (**Married**): Níveis são 0 e 1, consideramos 0 = “Não” e 1 = “Sim”; \* Paciente possui histórico familiar de câncer (**FamilyHx**): Níveis são 0 e 1, consideramos 0= “Não” e 1 = “Sim”; \* Paciente é fumante (**SmokingHx**): Níveis são “never”, “former”, “current”; \* Sexo (**Sex**): Níveis são 1 e 2, consideramos 1 = “Homem”, 2 = “Mulher”; \* Médico que atendeu o paciente(**DID**): Níveis são “A”, “B”, “C”, “D”.

As variáveis numéricas são: \* Tamanho do tumor (**tumorsize**): Contínua; \* Percentual de níveis de co2 (**co2**): Contínua; \* Escala de dor (**pain**): Inteiro de 0 a 10; \* Escala de mobilidade (**mobility**): Inteiro de 0 a 10; \* Número de tumores (**ntumors**): Inteiro de 0 a 9; \* Número de doses de morfina (**nmorphine**): Inteiro; \* Proporção de capacidade ótima do pulmão (**lungcapacity**): Contínua de 0 a 1; \* Idade (**Age**): Contínua em; \* Contagem de glóbulos brancos no sangue (**WBC**): Contínua; \* Contagem de glóbulos vermelhos no sangue (**RBC**): Contínua; \* Indicador de inflamação (**IL6**): Contínua.

Ainda, o banco possui informações faltantes que foram retirados através do recurso **na.omit()**.

Vamos avaliar alguns métodos de classificação para identificar o médico que atendeu o paciente, se o paciente está em remissão ou não e se o paciente tem ou teve o hábito de fumar. Além disso, vamos buscar determinar se o paciente está em remissão ou não a partir das variáveis sexo, histórico familiar, estado civil (casado ou não) e médico que atendeu.

Para todas as análises, a amostra foi dividida nos conjuntos de treino e teste através da função **createDatapartition** do pacote **caret** e com semente **205650**.

dataCont <- data[,c("tumorsize", "co2", "lungcapacity", "Age", "WBC", "RBC", "IL6")]

# Avaliação da identificação do médico (DID)

Para avaliar se é possível identificar o médico que atendeu o paciente, considerando os algoritmos vistos até a Terefa 3, vamos utilizar a regressão logística ordinal. Para tal, será usado o pacote **MASS** e a função **polr**. Separamos o dataset em treino (85%) e teste (15%).

set.seed(205650)  
indexes <- createDataPartition(data$DID, p = 0.90, list = FALSE)  
train = data[indexes, ]  
test = data[-indexes, ]

head(data)

## tumorsize co2 lungcapacity Age WBC RBC IL6 pain mobility ntumors nmorphine  
## 1 64 37 70 58 25 26 30 2 3 5 1  
## 2 41 24 65 54 50 23 42 6 4 5 11  
## 3 33 4 69 31 96 22 44 1 3 5 1  
## 4 86 62 67 37 85 46 48 7 4 4 13  
## 5 40 13 51 29 92 16 54 4 8 4 7  
## 6 89 49 63 51 68 3 33 8 2 9 12  
## remission Married FamilyHx SmokingHx Sex DID  
## 1 1 0 0 never 2 A  
## 2 1 0 1 never 1 C  
## 3 1 0 0 never 1 B  
## 4 1 1 1 never 1 C  
## 5 1 0 1 former 1 B  
## 6 1 1 0 never 1 B

str(data)

## 'data.frame': 91 obs. of 17 variables:  
## $ tumorsize : num 64 41 33 86 40 89 88 24 93 66 ...  
## $ co2 : num 37 24 4 62 13 49 72 51 94 29 ...  
## $ lungcapacity: num 70 65 69 67 51 63 55 46 48 42 ...  
## $ Age : int 58 54 31 37 29 51 52 31 53 57 ...  
## $ WBC : num 25 50 96 85 92 68 61 33 83 21 ...  
## $ RBC : num 26 23 22 46 16 3 17 2 45 57 ...  
## $ IL6 : num 30 42 44 48 54 33 82 25 52 34 ...  
## $ pain : int 2 6 1 7 4 8 6 2 8 6 ...  
## $ mobility : int 3 4 3 4 8 2 3 7 6 2 ...  
## $ ntumors : int 5 5 5 4 4 9 5 4 4 4 ...  
## $ nmorphine : int 1 11 1 13 7 12 9 0 12 10 ...  
## $ remission : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...  
## $ Married : Factor w/ 2 levels "0","1": 1 1 1 2 1 2 1 2 2 2 ...  
## $ FamilyHx : Factor w/ 2 levels "0","1": 1 2 1 2 2 1 2 1 1 1 ...  
## $ SmokingHx : Factor w/ 3 levels "current","former",..: 3 3 3 3 2 3 3 2 3 2 ...  
## $ Sex : Factor w/ 2 levels "1","2": 2 1 1 1 1 1 2 2 1 2 ...  
## $ DID : Factor w/ 4 levels "A","B","C","D": 1 3 2 3 2 2 2 4 2 2 ...  
## - attr(\*, "na.action")= 'omit' Named int 18 74 85 92 95  
## ..- attr(\*, "names")= chr "18" "74" "85" "92" ...

ord\_reg <- polr(DID~., data = train, Hess = TRUE)  
summary(ord\_reg)

## Call:  
## polr(formula = DID ~ ., data = train, Hess = TRUE)  
##   
## Coefficients:  
## Value Std. Error t value  
## tumorsize -0.036328 0.021346 -1.7018  
## co2 0.024811 0.010457 2.3728  
## lungcapacity -0.004639 0.023799 -0.1949  
## Age 0.019883 0.025506 0.7796  
## WBC -0.009969 0.009949 -1.0020  
## RBC -0.008008 0.009310 -0.8602  
## IL6 0.002735 0.009220 0.2967  
## pain -0.520527 0.323135 -1.6109  
## mobility -0.225580 0.118410 -1.9051  
## ntumors -0.069828 0.099711 -0.7003  
## nmorphine 0.209614 0.137838 1.5207  
## remission1 -2.001318 0.774383 -2.5844  
## Married1 0.634784 0.478515 1.3266  
## FamilyHx1 0.140064 0.444931 0.3148  
## SmokingHxformer -0.153440 0.881237 -0.1741  
## SmokingHxnever 1.374149 1.401411 0.9805  
## Sex2 -1.217297 0.556518 -2.1873  
##   
## Intercepts:  
## Value Std. Error t value  
## A|B -5.5632 2.0484 -2.7159  
## B|C -4.0620 2.0129 -2.0180  
## C|D -2.5417 1.9736 -1.2878  
##   
## Residual Deviance: 202.2211   
## AIC: 242.2211

summary\_table <- coef(summary(ord\_reg))  
pval <- pnorm(abs(summary\_table[, "t value"]),lower.tail = FALSE)\* 2  
summary\_table <- cbind(summary\_table, "p value" = round(pval,3))  
summary\_table

## Value Std. Error t value p value  
## tumorsize -0.036328240 0.021346361 -1.7018470 0.089  
## co2 0.024811497 0.010456808 2.3727600 0.018  
## lungcapacity -0.004639073 0.023799012 -0.1949271 0.845  
## Age 0.019883395 0.025505550 0.7795713 0.436  
## WBC -0.009968964 0.009949293 -1.0019770 0.316  
## RBC -0.008008390 0.009310418 -0.8601536 0.390  
## IL6 0.002735258 0.009219500 0.2966818 0.767  
## pain -0.520527209 0.323135035 -1.6108659 0.107  
## mobility -0.225579668 0.118410492 -1.9050649 0.057  
## ntumors -0.069828092 0.099711252 -0.7003030 0.484  
## nmorphine 0.209614077 0.137838425 1.5207231 0.128  
## remission1 -2.001317949 0.774382585 -2.5844046 0.010  
## Married1 0.634783659 0.478514917 1.3265703 0.185  
## FamilyHx1 0.140064265 0.444930887 0.3148000 0.753  
## SmokingHxformer -0.153439816 0.881237263 -0.1741186 0.862  
## SmokingHxnever 1.374148778 1.401411422 0.9805463 0.327  
## Sex2 -1.217297006 0.556518428 -2.1873436 0.029  
## A|B -5.563227313 2.048406262 -2.7158808 0.007  
## B|C -4.062004698 2.012911080 -2.0179752 0.044  
## C|D -2.541677710 1.973593917 -1.2878423 0.198

Resultados:

predtrain <- predict(ord\_reg, newdata = train)  
predtest <- predict(ord\_reg, newdata = test)  
predtrain

## [1] B C C B A B A A B A B A A C A A A A B D C B B C B B A B A C A B A A C D A B  
## [39] C A A B C B C A D C C C D C B D C C C B B D C B B C C A D D D A D B C D D C  
## [77] C B C D D B C  
## Levels: A B C D

cm <- confusionMatrix(table(predtrain, train$DID))  
cmtest <- confusionMatrix(table(predtest, test$DID))  
cm$table

##   
## predtrain A B C D  
## A 9 9 2 1  
## B 7 8 6 2  
## C 4 5 9 7  
## D 1 1 4 8

cm$byClass

## Sensitivity Specificity Pos Pred Value Neg Pred Value Precision  
## Class: A 0.4285714 0.8064516 0.4285714 0.8064516 0.4285714  
## Class: B 0.3478261 0.7500000 0.3478261 0.7500000 0.3478261  
## Class: C 0.4285714 0.7419355 0.3600000 0.7931034 0.3600000  
## Class: D 0.4444444 0.9076923 0.5714286 0.8550725 0.5714286  
## Recall F1 Prevalence Detection Rate Detection Prevalence  
## Class: A 0.4285714 0.4285714 0.2530120 0.10843373 0.2530120  
## Class: B 0.3478261 0.3478261 0.2771084 0.09638554 0.2771084  
## Class: C 0.4285714 0.3913043 0.2530120 0.10843373 0.3012048  
## Class: D 0.4444444 0.5000000 0.2168675 0.09638554 0.1686747  
## Balanced Accuracy  
## Class: A 0.6175115  
## Class: B 0.5489130  
## Class: C 0.5852535  
## Class: D 0.6760684

cmtest$table

##   
## predtest A B C D  
## A 0 0 0 0  
## B 2 1 0 1  
## C 0 1 1 1  
## D 0 0 1 0

cmtest$byClass

## Sensitivity Specificity Pos Pred Value Neg Pred Value Precision Recall  
## Class: A 0.0 1.0000000 NaN 0.7500000 NA 0.0  
## Class: B 0.5 0.5000000 0.2500000 0.7500000 0.2500000 0.5  
## Class: C 0.5 0.6666667 0.3333333 0.8000000 0.3333333 0.5  
## Class: D 0.0 0.8333333 0.0000000 0.7142857 0.0000000 0.0  
## F1 Prevalence Detection Rate Detection Prevalence  
## Class: A NA 0.25 0.000 0.000  
## Class: B 0.3333333 0.25 0.125 0.500  
## Class: C 0.4000000 0.25 0.125 0.375  
## Class: D NaN 0.25 0.000 0.125  
## Balanced Accuracy  
## Class: A 0.5000000  
## Class: B 0.5000000  
## Class: C 0.5833333  
## Class: D 0.4166667

cm$overall

## Accuracy Kappa AccuracyLower AccuracyUpper AccuracyNull   
## 0.409638554 0.209062622 0.302847038 0.523072365 0.277108434   
## AccuracyPValue McnemarPValue   
## 0.006293297 0.904500810

cmtest$overall

## Accuracy Kappa AccuracyLower AccuracyUpper AccuracyNull   
## 0.25000000 0.00000000 0.03185403 0.65085579 0.25000000   
## AccuracyPValue McnemarPValue   
## 0.63291931 NaN

A acurácia e sensibilidade obtiveram valores muito baixos, sendo a maioria menos de 50%, como observado na Tabela 1. Com este método e as variáveis disponíveis, não é adequado classificar qual médico atendeu o paciente.

para a tarefa 4, foi usada a abordagem de rede neural, através da função **neuralnet** do pacote de mesmo nome. Os parâmetros o número de neurônios escondidos (vértices) foi 2.

# fit neural network  
nn=neuralnet(DID~tumorsize + co2 + lungcapacity + Age + WBC + RBC + IL6,data=train, hidden=2, act.fct = "logistic",  
 linear.output = FALSE)  
nnplot <- plot(nn)

Treino:

Predict=compute(nn,train)  
prob <- Predict$net.result  
pred <- max.col(prob, 'first')  
pred <- ifelse(pred == 1, "A", ifelse(pred==2, "B", ifelse(pred == 3, "C", "D")))  
pred <- factor(pred, levels = c("A", "B", "C", "D"))  
confusionMatrix(table(pred, train$DID))

## Confusion Matrix and Statistics  
##   
##   
## pred A B C D  
## A 2 0 0 0  
## B 0 4 1 0  
## C 19 19 20 16  
## D 0 0 0 2  
##   
## Overall Statistics  
##   
## Accuracy : 0.3373   
## 95% CI : (0.2372, 0.4495)  
## No Information Rate : 0.2771   
## P-Value [Acc > NIR] : 0.1355   
##   
## Kappa : 0.1122   
##   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: A Class: B Class: C Class: D  
## Sensitivity 0.09524 0.17391 0.9524 0.1111  
## Specificity 1.00000 0.98333 0.1290 1.0000  
## Pos Pred Value 1.00000 0.80000 0.2703 1.0000  
## Neg Pred Value 0.76543 0.75641 0.8889 0.8025  
## Prevalence 0.25301 0.27711 0.2530 0.2169  
## Detection Rate 0.02410 0.04819 0.2410 0.0241  
## Detection Prevalence 0.02410 0.06024 0.8916 0.0241  
## Balanced Accuracy 0.54762 0.57862 0.5407 0.5556

Teste:

Predict=compute(nn,test)  
prob <- Predict$net.result  
pred <- max.col(prob, 'first')  
pred <- ifelse(pred == 1, "A", ifelse(pred==2, "B", ifelse(pred == 3, "C", "D")))  
pred <- factor(pred, levels = c("A", "B", "C", "D"))  
confusionMatrix(table(pred, test$DID))

## Confusion Matrix and Statistics  
##   
##   
## pred A B C D  
## A 0 0 0 0  
## B 1 0 0 0  
## C 1 2 2 2  
## D 0 0 0 0  
##   
## Overall Statistics  
##   
## Accuracy : 0.25   
## 95% CI : (0.0319, 0.6509)  
## No Information Rate : 0.25   
## P-Value [Acc > NIR] : 0.6329   
##   
## Kappa : 0   
##   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: A Class: B Class: C Class: D  
## Sensitivity 0.00 0.0000 1.0000 0.00  
## Specificity 1.00 0.8333 0.1667 1.00  
## Pos Pred Value NaN 0.0000 0.2857 NaN  
## Neg Pred Value 0.75 0.7143 1.0000 0.75  
## Prevalence 0.25 0.2500 0.2500 0.25  
## Detection Rate 0.00 0.0000 0.2500 0.00  
## Detection Prevalence 0.00 0.1250 0.8750 0.00  
## Balanced Accuracy 0.50 0.4167 0.5833 0.50

Os resultados com redes neurais também não foram satisfatórios.

Vamos fazer uma tentativa com árvore de decisão:

set.seed(205650)  
indexes <- createDataPartition(data$DID, p = 0.85, list = FALSE)  
train = data[indexes, ]  
test = data[-indexes, ]  
  
model\_fit <- rpart(DID ~ ., data = train, method = "class", control = list(minsplit = 20))  
predicted\_train <- predict(model\_fit, newdata = train, "class")  
predicted\_test <- predict(model\_fit, newdata = test, "class")  
summary(model\_fit)

## Call:  
## rpart(formula = DID ~ ., data = train, method = "class", control = list(minsplit = 20))  
## n= 79   
##   
## CP nsplit rel error xerror xstd  
## 1 0.14035088 0 1.0000000 1.087719 0.06408139  
## 2 0.06140351 1 0.8596491 1.122807 0.06115715  
## 3 0.05263158 3 0.7368421 1.192982 0.05398367  
## 4 0.03508772 5 0.6315789 1.175439 0.05596798  
## 5 0.01000000 6 0.5964912 1.105263 0.06266743  
##   
## Variable importance  
## nmorphine pain lungcapacity ntumors mobility remission   
## 14 14 12 11 10 8   
## tumorsize WBC Age IL6 SmokingHx RBC   
## 5 5 5 5 5 3   
## co2   
## 3   
##   
## Node number 1: 79 observations, complexity param=0.1403509  
## predicted class=B expected loss=0.721519 P(node) =1  
## class counts: 20 22 20 17  
## probabilities: 0.253 0.278 0.253 0.215   
## left son=2 (34 obs) right son=3 (45 obs)  
## Primary splits:  
## lungcapacity < 37.5 to the right, improve=2.919980, (0 missing)  
## mobility < 3.5 to the left, improve=2.461568, (0 missing)  
## remission splits as RL, improve=2.271362, (0 missing)  
## pain < 7.5 to the right, improve=2.007448, (0 missing)  
## tumorsize < 87.5 to the right, improve=2.001306, (0 missing)  
## Surrogate splits:  
## remission splits as RL, agree=0.899, adj=0.765, (0 split)  
## SmokingHx splits as RRL, agree=0.797, adj=0.529, (0 split)  
## Age < 47 to the right, agree=0.696, adj=0.294, (0 split)  
## tumorsize < 63.5 to the right, agree=0.684, adj=0.265, (0 split)  
## IL6 < 60.5 to the left, agree=0.671, adj=0.235, (0 split)  
##   
## Node number 2: 34 observations, complexity param=0.06140351  
## predicted class=B expected loss=0.5588235 P(node) =0.4303797  
## class counts: 10 15 5 4  
## probabilities: 0.294 0.441 0.147 0.118   
## left son=4 (24 obs) right son=5 (10 obs)  
## Primary splits:  
## mobility < 3.5 to the right, improve=2.668627, (0 missing)  
## ntumors < 4.5 to the right, improve=2.598930, (0 missing)  
## Age < 39.5 to the right, improve=2.561610, (0 missing)  
## WBC < 35 to the left, improve=2.551961, (0 missing)  
## tumorsize < 31.5 to the right, improve=2.408371, (0 missing)  
## Surrogate splits:  
## lungcapacity < 68 to the left, agree=0.765, adj=0.2, (0 split)  
## Age < 53.5 to the left, agree=0.765, adj=0.2, (0 split)  
## IL6 < 12.5 to the right, agree=0.735, adj=0.1, (0 split)  
## pain < 7.5 to the left, agree=0.735, adj=0.1, (0 split)  
##   
## Node number 3: 45 observations, complexity param=0.05263158  
## predicted class=C expected loss=0.6666667 P(node) =0.5696203  
## class counts: 10 7 15 13  
## probabilities: 0.222 0.156 0.333 0.289   
## left son=6 (12 obs) right son=7 (33 obs)  
## Primary splits:  
## nmorphine < 10.5 to the right, improve=2.160606, (0 missing)  
## Age < 47.5 to the left, improve=2.128822, (0 missing)  
## RBC < 81.5 to the left, improve=1.752882, (0 missing)  
## pain < 5.5 to the right, improve=1.704885, (0 missing)  
## co2 < 39.5 to the left, improve=1.340741, (0 missing)  
## Surrogate splits:  
## pain < 5.5 to the right, agree=0.911, adj=0.667, (0 split)  
## RBC < 7.5 to the left, agree=0.778, adj=0.167, (0 split)  
## IL6 < 15.5 to the left, agree=0.778, adj=0.167, (0 split)  
## Age < 29 to the left, agree=0.756, adj=0.083, (0 split)  
## mobility < 3.5 to the left, agree=0.756, adj=0.083, (0 split)  
##   
## Node number 4: 24 observations, complexity param=0.06140351  
## predicted class=A expected loss=0.625 P(node) =0.3037975  
## class counts: 9 7 5 3  
## probabilities: 0.375 0.292 0.208 0.125   
## left son=8 (15 obs) right son=9 (9 obs)  
## Primary splits:  
## ntumors < 4.5 to the right, improve=3.166667, (0 missing)  
## pain < 4.5 to the right, improve=2.100000, (0 missing)  
## RBC < 51 to the right, improve=1.991841, (0 missing)  
## Age < 37.5 to the right, improve=1.963869, (0 missing)  
## WBC < 36 to the left, improve=1.791667, (0 missing)  
## Surrogate splits:  
## WBC < 74 to the left, agree=0.792, adj=0.444, (0 split)  
## tumorsize < 91.5 to the left, agree=0.708, adj=0.222, (0 split)  
## co2 < 93 to the left, agree=0.708, adj=0.222, (0 split)  
## RBC < 18 to the right, agree=0.708, adj=0.222, (0 split)  
## lungcapacity < 45.5 to the left, agree=0.667, adj=0.111, (0 split)  
##   
## Node number 5: 10 observations  
## predicted class=B expected loss=0.2 P(node) =0.1265823  
## class counts: 1 8 0 1  
## probabilities: 0.100 0.800 0.000 0.100   
##   
## Node number 6: 12 observations  
## predicted class=C expected loss=0.5 P(node) =0.1518987  
## class counts: 5 0 6 1  
## probabilities: 0.417 0.000 0.500 0.083   
##   
## Node number 7: 33 observations, complexity param=0.05263158  
## predicted class=D expected loss=0.6363636 P(node) =0.4177215  
## class counts: 5 7 9 12  
## probabilities: 0.152 0.212 0.273 0.364   
## left son=14 (9 obs) right son=15 (24 obs)  
## Primary splits:  
## nmorphine < 3.5 to the left, improve=1.883838, (0 missing)  
## pain < 3.5 to the left, improve=1.814394, (0 missing)  
## Age < 44.5 to the left, improve=1.697636, (0 missing)  
## Married splits as LR, improve=1.272727, (0 missing)  
## WBC < 64.5 to the right, improve=1.259394, (0 missing)  
## Surrogate splits:  
## pain < 2.5 to the left, agree=0.818, adj=0.333, (0 split)  
## mobility < 4.5 to the left, agree=0.788, adj=0.222, (0 split)  
## co2 < 5.5 to the left, agree=0.758, adj=0.111, (0 split)  
## WBC < 5 to the left, agree=0.758, adj=0.111, (0 split)  
##   
## Node number 8: 15 observations  
## predicted class=A expected loss=0.4 P(node) =0.1898734  
## class counts: 9 2 3 1  
## probabilities: 0.600 0.133 0.200 0.067   
##   
## Node number 9: 9 observations  
## predicted class=B expected loss=0.4444444 P(node) =0.1139241  
## class counts: 0 5 2 2  
## probabilities: 0.000 0.556 0.222 0.222   
##   
## Node number 14: 9 observations  
## predicted class=A expected loss=0.5555556 P(node) =0.1139241  
## class counts: 4 2 2 1  
## probabilities: 0.444 0.222 0.222 0.111   
##   
## Node number 15: 24 observations, complexity param=0.03508772  
## predicted class=D expected loss=0.5416667 P(node) =0.3037975  
## class counts: 1 5 7 11  
## probabilities: 0.042 0.208 0.292 0.458   
## left son=30 (16 obs) right son=31 (8 obs)  
## Primary splits:  
## pain < 3.5 to the right, improve=1.958333, (0 missing)  
## Age < 41.5 to the left, improve=1.776190, (0 missing)  
## RBC < 37.5 to the right, improve=1.458333, (0 missing)  
## Sex splits as RL, improve=1.388889, (0 missing)  
## mobility < 7.5 to the right, improve=1.083333, (0 missing)  
## Surrogate splits:  
## nmorphine < 4.5 to the right, agree=0.750, adj=0.250, (0 split)  
## tumorsize < 10 to the right, agree=0.708, adj=0.125, (0 split)  
## IL6 < 82.5 to the left, agree=0.708, adj=0.125, (0 split)  
## ntumors < 8.5 to the left, agree=0.708, adj=0.125, (0 split)  
## remission splits as LR, agree=0.708, adj=0.125, (0 split)  
##   
## Node number 30: 16 observations  
## predicted class=D expected loss=0.5 P(node) =0.2025316  
## class counts: 1 5 2 8  
## probabilities: 0.063 0.312 0.125 0.500   
##   
## Node number 31: 8 observations  
## predicted class=C expected loss=0.375 P(node) =0.1012658  
## class counts: 0 0 5 3  
## probabilities: 0.000 0.000 0.625 0.375

# importancia das vars  
model\_fit$variable.importance

## nmorphine pain lungcapacity ntumors mobility remission   
## 4.5340278 4.2935462 3.8055575 3.4114583 3.2673087 2.4777177   
## tumorsize WBC Age IL6 SmokingHx RBC   
## 1.7214313 1.6167228 1.5725937 1.5588096 1.5458718 1.0638047   
## co2   
## 0.9130191

#vars que foram usadas na arvore  
printcp(model\_fit)

##   
## Classification tree:  
## rpart(formula = DID ~ ., data = train, method = "class", control = list(minsplit = 20))  
##   
## Variables actually used in tree construction:  
## [1] lungcapacity mobility nmorphine ntumors pain   
##   
## Root node error: 57/79 = 0.72152  
##   
## n= 79   
##   
## CP nsplit rel error xerror xstd  
## 1 0.140351 0 1.00000 1.0877 0.064081  
## 2 0.061404 1 0.85965 1.1228 0.061157  
## 3 0.052632 3 0.73684 1.1930 0.053984  
## 4 0.035088 5 0.63158 1.1754 0.055968  
## 5 0.010000 6 0.59649 1.1053 0.062667

#Matriz de confusao  
ctable\_train <- confusionMatrix(predicted\_train, train$DID)  
ctable\_test <- confusionMatrix(predicted\_test, test$DID)  
ctable\_train$table

## Reference  
## Prediction A B C D  
## A 13 4 5 2  
## B 1 13 2 3  
## C 5 0 11 4  
## D 1 5 2 8

ctable\_test$table

## Reference  
## Prediction A B C D  
## A 0 2 1 1  
## B 1 0 1 0  
## C 1 1 0 2  
## D 1 0 1 0

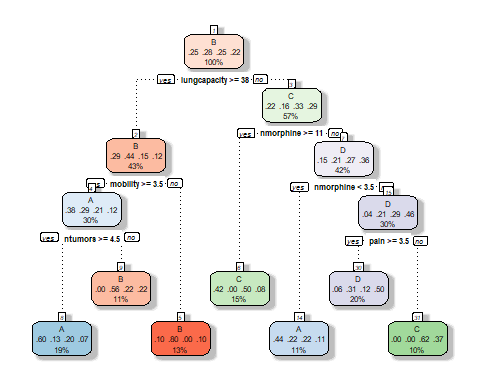
ctable\_train

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction A B C D  
## A 13 4 5 2  
## B 1 13 2 3  
## C 5 0 11 4  
## D 1 5 2 8  
##   
## Overall Statistics  
##   
## Accuracy : 0.5696   
## 95% CI : (0.4533, 0.6806)  
## No Information Rate : 0.2785   
## P-Value [Acc > NIR] : 5.527e-08   
##   
## Kappa : 0.425   
##   
## Mcnemar's Test P-Value : 0.506   
##   
## Statistics by Class:  
##   
## Class: A Class: B Class: C Class: D  
## Sensitivity 0.6500 0.5909 0.5500 0.4706  
## Specificity 0.8136 0.8947 0.8475 0.8710  
## Pos Pred Value 0.5417 0.6842 0.5500 0.5000  
## Neg Pred Value 0.8727 0.8500 0.8475 0.8571  
## Prevalence 0.2532 0.2785 0.2532 0.2152  
## Detection Rate 0.1646 0.1646 0.1392 0.1013  
## Detection Prevalence 0.3038 0.2405 0.2532 0.2025  
## Balanced Accuracy 0.7318 0.7428 0.6987 0.6708

ctable\_test

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction A B C D  
## A 0 2 1 1  
## B 1 0 1 0  
## C 1 1 0 2  
## D 1 0 1 0  
##   
## Overall Statistics  
##   
## Accuracy : 0   
## 95% CI : (0, 0.2646)  
## No Information Rate : 0.25   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : -0.3333   
##   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: A Class: B Class: C Class: D  
## Sensitivity 0.0000 0.0000 0.0000 0.0000  
## Specificity 0.5556 0.7778 0.5556 0.7778  
## Pos Pred Value 0.0000 0.0000 0.0000 0.0000  
## Neg Pred Value 0.6250 0.7000 0.6250 0.7000  
## Prevalence 0.2500 0.2500 0.2500 0.2500  
## Detection Rate 0.0000 0.0000 0.0000 0.0000  
## Detection Prevalence 0.3333 0.1667 0.3333 0.1667  
## Balanced Accuracy 0.2778 0.3889 0.2778 0.3889

fancyRpartPlot(model\_fit, yesno = 2, caption = NULL, palettes=c("Blues", "Reds", "Greens", "Purples"), tweak = 1)



dev.off()

## pdf   
## 3

Mesmo com árvore de decisão, não foi possível obter resultados satisfatórios.

# Avaliação da identificação se o paciente está em remissão

Para a identificação se o paciente está em remissão, considerando os métodos vistos até a Tarefa 3 do curso, a regressão logística se mostrou adequada, pois **remission** é uma variável dicotômica e este algoritmo funciona bem para este tipo de classificação. Além disso, é possível utilizar tanto as variáveis categóricas quanto numéricas.

## Resultados Regressão Logística binomial

A função **glm** do pacote **MASS** foi utilizada para estimar o modelo de regressão logística binomial. Após a execução do modelo quase completo (*lungcapacity* foi retirado pois o modelo não convergiu com todas as variáveis), foi aplicada a função **stepAIC**, que faz a seleção de variáveis através do *stepwise* (foi escolhido o método *backard*) tendo como critério o AIC.

set.seed(205650)  
indexes <- createDataPartition(data$remission, p = 0.85, list = FALSE)  
train = data[indexes, ]  
test = data[-indexes, ]  
  
fullMod <- glm(remission ~ DID + pain + Sex + SmokingHx + FamilyHx + Married + nmorphine + ntumors + mobility + tumorsize + co2 + Age + WBC + RBC + IL6, data = train, family = "binomial")  
summary(fullMod)

##   
## Call:  
## glm(formula = remission ~ DID + pain + Sex + SmokingHx + FamilyHx +   
## Married + nmorphine + ntumors + mobility + tumorsize + co2 +   
## Age + WBC + RBC + IL6, family = "binomial", data = train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -3.5412 -0.3047 0.0001 0.1426 1.5602   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 17.961283 7.542033 2.381 0.01724 \*   
## DIDB 0.004008 1.375559 0.003 0.99768   
## DIDC -2.964018 1.641804 -1.805 0.07102 .   
## DIDD -3.449487 1.496350 -2.305 0.02115 \*   
## pain -1.145729 0.884684 -1.295 0.19530   
## Sex2 -2.567189 1.293424 -1.985 0.04717 \*   
## SmokingHxformer 3.907616 2.490666 1.569 0.11667   
## SmokingHxnever 14.922013 5.367677 2.780 0.00544 \*\*  
## FamilyHx1 -1.675989 1.174062 -1.428 0.15343   
## Married1 2.237488 1.184193 1.889 0.05883 .   
## nmorphine 0.306452 0.357845 0.856 0.39179   
## ntumors -0.156555 0.255208 -0.613 0.53959   
## mobility -0.753191 0.349043 -2.158 0.03094 \*   
## tumorsize -0.178947 0.074182 -2.412 0.01585 \*   
## co2 0.018720 0.017009 1.101 0.27107   
## Age 0.020766 0.050505 0.411 0.68095   
## WBC -0.022548 0.020180 -1.117 0.26383   
## RBC -0.036203 0.024708 -1.465 0.14286   
## IL6 -0.050502 0.028761 -1.756 0.07910 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 109.505 on 78 degrees of freedom  
## Residual deviance: 39.802 on 60 degrees of freedom  
## AIC: 77.802  
##   
## Number of Fisher Scoring iterations: 8

stepMod <- stepAIC(fullMod, direction = "backward")

## Start: AIC=77.8  
## remission ~ DID + pain + Sex + SmokingHx + FamilyHx + Married +   
## nmorphine + ntumors + mobility + tumorsize + co2 + Age +   
## WBC + RBC + IL6  
##   
## Df Deviance AIC  
## - Age 1 39.972 75.972  
## - ntumors 1 40.198 76.198  
## - nmorphine 1 40.625 76.625  
## - co2 1 41.069 77.069  
## - WBC 1 41.172 77.172  
## <none> 39.802 77.802  
## - pain 1 41.803 77.803  
## - FamilyHx 1 42.103 78.103  
## - RBC 1 42.990 78.990  
## - IL6 1 44.163 80.163  
## - Married 1 44.508 80.508  
## - DID 3 48.706 80.706  
## - Sex 1 45.217 81.217  
## - mobility 1 47.442 83.442  
## - tumorsize 1 54.519 90.519  
## - SmokingHx 2 73.814 107.814  
##   
## Step: AIC=75.97  
## remission ~ DID + pain + Sex + SmokingHx + FamilyHx + Married +   
## nmorphine + ntumors + mobility + tumorsize + co2 + WBC +   
## RBC + IL6  
##   
## Df Deviance AIC  
## - ntumors 1 40.356 74.356  
## - nmorphine 1 40.865 74.865  
## - co2 1 41.133 75.133  
## - WBC 1 41.390 75.390  
## <none> 39.972 75.972  
## - pain 1 41.992 75.992  
## - RBC 1 43.151 77.151  
## - FamilyHx 1 43.255 77.255  
## - IL6 1 44.232 78.232  
## - Married 1 44.631 78.631  
## - DID 3 48.711 78.711  
## - Sex 1 45.261 79.261  
## - mobility 1 47.619 81.619  
## - tumorsize 1 54.545 88.545  
## - SmokingHx 2 76.840 108.840  
##   
## Step: AIC=74.36  
## remission ~ DID + pain + Sex + SmokingHx + FamilyHx + Married +   
## nmorphine + mobility + tumorsize + co2 + WBC + RBC + IL6  
##   
## Df Deviance AIC  
## - nmorphine 1 41.259 73.259  
## - co2 1 41.298 73.298  
## - WBC 1 41.977 73.977  
## <none> 40.356 74.356  
## - pain 1 42.528 74.528  
## - RBC 1 43.422 75.422  
## - FamilyHx 1 43.654 75.654  
## - IL6 1 44.266 76.266  
## - Married 1 44.685 76.685  
## - DID 3 48.939 76.939  
## - Sex 1 45.571 77.571  
## - mobility 1 47.735 79.735  
## - tumorsize 1 54.910 86.910  
## - SmokingHx 2 77.359 107.359  
##   
## Step: AIC=73.26  
## remission ~ DID + pain + Sex + SmokingHx + FamilyHx + Married +   
## mobility + tumorsize + co2 + WBC + RBC + IL6  
##   
## Df Deviance AIC  
## - co2 1 42.293 72.293  
## - WBC 1 42.645 72.645  
## <none> 41.259 73.259  
## - pain 1 43.492 73.492  
## - FamilyHx 1 43.793 73.793  
## - RBC 1 43.837 73.837  
## - IL6 1 44.573 74.573  
## - Married 1 44.883 74.883  
## - DID 3 48.942 74.942  
## - Sex 1 45.932 75.932  
## - mobility 1 48.069 78.069  
## - tumorsize 1 55.879 85.879  
## - SmokingHx 2 78.025 106.025  
##   
## Step: AIC=72.29  
## remission ~ DID + pain + Sex + SmokingHx + FamilyHx + Married +   
## mobility + tumorsize + WBC + RBC + IL6  
##   
## Df Deviance AIC  
## - WBC 1 43.579 71.579  
## <none> 42.293 72.293  
## - RBC 1 44.324 72.324  
## - pain 1 44.570 72.570  
## - FamilyHx 1 44.662 72.662  
## - DID 3 49.786 73.786  
## - IL6 1 46.024 74.024  
## - Married 1 46.024 74.024  
## - Sex 1 46.494 74.494  
## - mobility 1 49.525 77.525  
## - tumorsize 1 55.883 83.883  
## - SmokingHx 2 78.714 104.714  
##   
## Step: AIC=71.58  
## remission ~ DID + pain + Sex + SmokingHx + FamilyHx + Married +   
## mobility + tumorsize + RBC + IL6  
##   
## Df Deviance AIC  
## - RBC 1 44.983 70.983  
## <none> 43.579 71.579  
## - pain 1 45.681 71.681  
## - DID 3 50.182 72.182  
## - FamilyHx 1 46.707 72.707  
## - Sex 1 46.861 72.861  
## - Married 1 47.274 73.274  
## - IL6 1 48.069 74.069  
## - mobility 1 51.324 77.324  
## - tumorsize 1 57.093 83.093  
## - SmokingHx 2 79.638 103.638  
##   
## Step: AIC=70.98  
## remission ~ DID + pain + Sex + SmokingHx + FamilyHx + Married +   
## mobility + tumorsize + IL6  
##   
## Df Deviance AIC  
## - pain 1 46.409 70.409  
## <none> 44.983 70.983  
## - Sex 1 47.648 71.648  
## - FamilyHx 1 47.761 71.761  
## - Married 1 47.808 71.808  
## - DID 3 51.852 71.852  
## - IL6 1 48.807 72.807  
## - mobility 1 52.631 76.631  
## - tumorsize 1 58.609 82.609  
## - SmokingHx 2 81.022 103.022  
##   
## Step: AIC=70.41  
## remission ~ DID + Sex + SmokingHx + FamilyHx + Married + mobility +   
## tumorsize + IL6  
##   
## Df Deviance AIC  
## - Married 1 48.226 70.226  
## - DID 3 52.359 70.359  
## <none> 46.409 70.409  
## - FamilyHx 1 48.937 70.937  
## - IL6 1 48.956 70.956  
## - Sex 1 49.059 71.059  
## - mobility 1 56.478 78.478  
## - tumorsize 1 65.121 87.121  
## - SmokingHx 2 81.731 101.731  
##   
## Step: AIC=70.23  
## remission ~ DID + Sex + SmokingHx + FamilyHx + mobility + tumorsize +   
## IL6  
##   
## Df Deviance AIC  
## - DID 3 53.400 69.400  
## - Sex 1 50.115 70.115  
## <none> 48.226 70.226  
## - FamilyHx 1 50.641 70.641  
## - IL6 1 51.771 71.771  
## - mobility 1 57.563 77.563  
## - tumorsize 1 68.627 88.627  
## - SmokingHx 2 85.722 103.722  
##   
## Step: AIC=69.4  
## remission ~ Sex + SmokingHx + FamilyHx + mobility + tumorsize +   
## IL6  
##   
## Df Deviance AIC  
## - Sex 1 55.345 69.345  
## <none> 53.400 69.400  
## - FamilyHx 1 56.234 70.234  
## - IL6 1 57.153 71.153  
## - mobility 1 64.708 78.708  
## - tumorsize 1 79.200 93.200  
## - SmokingHx 2 95.376 107.376  
##   
## Step: AIC=69.35  
## remission ~ SmokingHx + FamilyHx + mobility + tumorsize + IL6  
##   
## Df Deviance AIC  
## <none> 55.345 69.345  
## - FamilyHx 1 57.903 69.903  
## - IL6 1 59.891 71.891  
## - mobility 1 66.033 78.033  
## - tumorsize 1 79.302 91.302  
## - SmokingHx 2 97.395 107.395

stepMod$anova

## Stepwise Model Path   
## Analysis of Deviance Table  
##   
## Initial Model:  
## remission ~ DID + pain + Sex + SmokingHx + FamilyHx + Married +   
## nmorphine + ntumors + mobility + tumorsize + co2 + Age +   
## WBC + RBC + IL6  
##   
## Final Model:  
## remission ~ SmokingHx + FamilyHx + mobility + tumorsize + IL6  
##   
##   
## Step Df Deviance Resid. Df Resid. Dev AIC  
## 1 60 39.80220 77.80220  
## 2 - Age 1 0.1698026 61 39.97200 75.97200  
## 3 - ntumors 1 0.3836302 62 40.35563 74.35563  
## 4 - nmorphine 1 0.9032755 63 41.25891 73.25891  
## 5 - co2 1 1.0336032 64 42.29251 72.29251  
## 6 - WBC 1 1.2863181 65 43.57883 71.57883  
## 7 - RBC 1 1.4037116 66 44.98254 70.98254  
## 8 - pain 1 1.4259936 67 46.40853 70.40853  
## 9 - Married 1 1.8177086 68 48.22624 70.22624  
## 10 - DID 3 5.1736382 71 53.39988 69.39988  
## 11 - Sex 1 1.9454991 72 55.34538 69.34538

O modelo final, após o *stepwise backard*, contém as seguintes variáveis: **SmokingHx, FamilyHx, mobility, tumorsize e IL6**. As variáveis consideradas significativas (p-valor < 0.05) foram: **SmokingHx, mobility, tumorsize, DID, Sex, Married**.

Os resultados obtidos foram:

Treino: 82,37% de acuracia, 75,71% de sensibilidade e 89,74% de especificidade. Teste: 83,33% de acuracia, 100% sensibilidade, 66,67% de especificidade.

predStepTrain <- ifelse(predict(stepMod, newdata = train) >= 0.5, 1, 0)  
cm <- confusionMatrix(table(predStepTrain, train$remission), positive = "1")  
cm$byClass

## Sensitivity Specificity Pos Pred Value   
## 0.7500000 0.8974359 0.8823529   
## Neg Pred Value Precision Recall   
## 0.7777778 0.8823529 0.7500000   
## F1 Prevalence Detection Rate   
## 0.8108108 0.5063291 0.3797468   
## Detection Prevalence Balanced Accuracy   
## 0.4303797 0.8237179

kable(cm$byClass)

|  |  |
| --- | --- |
|  | x |
| Sensitivity | 0.7500000 |
| Specificity | 0.8974359 |
| Pos Pred Value | 0.8823529 |
| Neg Pred Value | 0.7777778 |
| Precision | 0.8823529 |
| Recall | 0.7500000 |
| F1 | 0.8108108 |
| Prevalence | 0.5063291 |
| Detection Rate | 0.3797468 |
| Detection Prevalence | 0.4303797 |
| Balanced Accuracy | 0.8237179 |

predStepTest <- ifelse(predict(stepMod, newdata = test) >= 0.5, 1, 0)  
cm <- confusionMatrix(table(predStepTest, test$remission), positive = "1")  
cm

## Confusion Matrix and Statistics  
##   
##   
## predStepTest 0 1  
## 0 4 0  
## 1 2 6  
##   
## Accuracy : 0.8333   
## 95% CI : (0.5159, 0.9791)  
## No Information Rate : 0.5   
## P-Value [Acc > NIR] : 0.01929   
##   
## Kappa : 0.6667   
##   
## Mcnemar's Test P-Value : 0.47950   
##   
## Sensitivity : 1.0000   
## Specificity : 0.6667   
## Pos Pred Value : 0.7500   
## Neg Pred Value : 1.0000   
## Prevalence : 0.5000   
## Detection Rate : 0.5000   
## Detection Prevalence : 0.6667   
## Balanced Accuracy : 0.8333   
##   
## 'Positive' Class : 1   
##

Portanto, conclui-se que sim, pode-se identificar se o paciente está em remissão com aproximadamente 80% de acurácia. Apesar de uma acurácia relativamente boa, não considera-se um bom modelo, pois, neste caso, a especificidade tem grande importância (queremos identificar corretamente quem não está em remissão para definir um novo tratamento) e obteve um resultado pouco satisfatório.

Considerando os métodos apresentados para a Tarefa 4, utilizaremos SVM, que também é um bom algoritmo para classificação de variáveis dicotômicas. Contudo, ele é adequado para variáveis numéricas, portanto, utilizaremos apenas estas variáveis na análise.

SVM para as variáveis numéricas:

dataSVM <- na.omit(dataCont)  
dataSVM <- cbind(remission = data[,"remission"], dataSVM)  
head(dataSVM)

## remission tumorsize co2 lungcapacity Age WBC RBC IL6  
## 1 1 64 37 70 58 25 26 30  
## 2 1 41 24 65 54 50 23 42  
## 3 1 33 4 69 31 96 22 44  
## 4 1 86 62 67 37 85 46 48  
## 5 1 40 13 51 29 92 16 54  
## 6 1 89 49 63 51 68 3 33

summary(dataSVM)

## remission tumorsize co2 lungcapacity Age   
## 0:45 Min. : 1.00 Min. : 1.00 Min. : 1.00 Min. :23.00   
## 1:46 1st Qu.:22.50 1st Qu.:24.50 1st Qu.:21.50 1st Qu.:35.00   
## Median :46.00 Median :48.00 Median :34.00 Median :41.00   
## Mean :46.91 Mean :48.24 Mean :34.87 Mean :41.67   
## 3rd Qu.:70.50 3rd Qu.:71.50 3rd Qu.:49.50 3rd Qu.:48.00   
## Max. :95.00 Max. :96.00 Max. :70.00 Max. :71.00   
## WBC RBC IL6   
## Min. : 1.00 Min. : 1.0 Min. : 1.00   
## 1st Qu.:24.50 1st Qu.:23.5 1st Qu.:23.50   
## Median :47.00 Median :46.0 Median :46.00   
## Mean :47.81 Mean :46.0 Mean :45.89   
## 3rd Qu.:70.50 3rd Qu.:68.5 3rd Qu.:68.50   
## Max. :96.00 Max. :91.0 Max. :90.00

set.seed(205650)  
indexes <- createDataPartition(dataSVM$remission, p = 0.85, list = FALSE)  
train = dataSVM[indexes, ]  
test = dataSVM[-indexes, ]  
  
svmfit = svm(remission ~ ., data = train, kernel ="linear", cost = 10, scale = FALSE )  
summary(svmfit)

##   
## Call:  
## svm(formula = remission ~ ., data = train, kernel = "linear", cost = 10,   
## scale = FALSE)  
##   
##   
## Parameters:  
## SVM-Type: C-classification   
## SVM-Kernel: linear   
## cost: 10   
##   
## Number of Support Vectors: 27  
##   
## ( 12 15 )  
##   
##   
## Number of Classes: 2   
##   
## Levels:   
## 0 1

# indice dos support vectors:  
svmfit$index

## [1] 5 10 16 24 25 31 32 33 34 36 37 38 45 47 48 50 51 53 63 64 65 66 67 72 77  
## [26] 78 79

svmfit$coefs

## [,1]  
## [1,] 10.000000  
## [2,] 10.000000  
## [3,] 4.934157  
## [4,] 10.000000  
## [5,] 10.000000  
## [6,] 10.000000  
## [7,] 10.000000  
## [8,] 10.000000  
## [9,] 10.000000  
## [10,] 10.000000  
## [11,] 8.373840  
## [12,] 10.000000  
## [13,] -4.060741  
## [14,] -1.151661  
## [15,] -10.000000  
## [16,] -1.893499  
## [17,] -2.939604  
## [18,] -10.000000  
## [19,] -10.000000  
## [20,] -10.000000  
## [21,] -10.000000  
## [22,] -10.000000  
## [23,] -10.000000  
## [24,] -10.000000  
## [25,] -4.395122  
## [26,] -10.000000  
## [27,] -8.867371

dataSVM[svmfit$index,]

## remission tumorsize co2 lungcapacity Age WBC RBC IL6  
## 5 1 40 13 51 29 92 16 54  
## 10 1 66 29 42 57 21 57 34  
## 16 1 22 17 58 53 69 59 36  
## 25 1 68 47 41 36 45 50 60  
## 26 1 67 75 53 30 81 32 89  
## 32 1 92 73 50 59 89 34 32  
## 33 1 28 26 58 45 56 12 4  
## 34 1 46 40 66 48 8 20 21  
## 35 1 3 1 62 32 95 52 62  
## 37 1 83 88 57 32 54 61 57  
## 38 1 78 84 35 42 36 74 39  
## 39 1 56 67 26 39 9 30 67  
## 46 1 54 78 38 58 58 72 47  
## 48 0 10 32 11 36 15 84 19  
## 49 0 79 33 2 41 31 90 1  
## 51 0 55 77 29 36 76 24 38  
## 52 0 63 93 34 38 60 79 73  
## 54 0 91 68 25 71 90 81 78  
## 64 0 75 43 15 44 32 11 71  
## 65 0 62 20 6 45 94 48 14  
## 66 0 9 44 24 34 41 83 85  
## 67 0 37 42 17 35 77 80 11  
## 68 0 31 63 9 51 35 78 35  
## 73 0 44 36 30 35 52 70 26  
## 79 0 58 28 35 41 55 4 9  
## 80 0 8 3 27 38 2 29 81  
## 81 0 29 19 24 23 88 6 81

# betas:  
#beta0 = svmfit$rho  
#svmfit$coefs  
#c(svmfit$coefs)  
#betas = c(svmfit$coefs)%\*%as.matrix(train[svmfit$index,])  
#coefs <- cbind(intercept = beta0, betas)  
  
  
#formula <- paste(round(coefs, 1), colnames(coefs), sep = "\*", collapse = " ")  
#formula  
  
# matrix de confus?o:  
confusionMatrix(table(predict(svmfit), train$remission), positive="1")

## Confusion Matrix and Statistics  
##   
##   
## 0 1  
## 0 34 3  
## 1 5 37  
##   
## Accuracy : 0.8987   
## 95% CI : (0.8102, 0.9553)  
## No Information Rate : 0.5063   
## P-Value [Acc > NIR] : 1.072e-13   
##   
## Kappa : 0.7973   
##   
## Mcnemar's Test P-Value : 0.7237   
##   
## Sensitivity : 0.9250   
## Specificity : 0.8718   
## Pos Pred Value : 0.8810   
## Neg Pred Value : 0.9189   
## Prevalence : 0.5063   
## Detection Rate : 0.4684   
## Detection Prevalence : 0.5316   
## Balanced Accuracy : 0.8984   
##   
## 'Positive' Class : 1   
##

SVM ficou com 89,87% de acuracia, 92,50% de sensibilidade e 87,18% de especificidade.

Para visualizar, vamos escolher as duas variáveis consideradas mais importantes pelo modelo:

w <- t(svmfit$coefs) %\*% svmfit$SV # weight vectors  
w

## tumorsize co2 lungcapacity Age WBC RBC  
## [1,] -0.01107499 -0.001460239 0.1539888 0.07397403 -0.02334463 0.007611741  
## IL6  
## [1,] -0.01938388

w <- apply(w, 2, function(v){sqrt(sum(v^2))}) # weight  
w <- sort(w, decreasing = T)  
w

## lungcapacity Age WBC IL6 tumorsize RBC   
## 0.153988835 0.073974031 0.023344630 0.019383876 0.011074991 0.007611741   
## co2   
## 0.001460239

svmfit2 = svm(remission ~ lungcapacity + Age, data = train, kernel ="linear", cost = 10, scale = FALSE )  
head(train)

## remission tumorsize co2 lungcapacity Age WBC RBC IL6  
## 1 1 64 37 70 58 25 26 30  
## 2 1 41 24 65 54 50 23 42  
## 3 1 33 4 69 31 96 22 44  
## 4 1 86 62 67 37 85 46 48  
## 5 1 40 13 51 29 92 16 54  
## 6 1 89 49 63 51 68 3 33

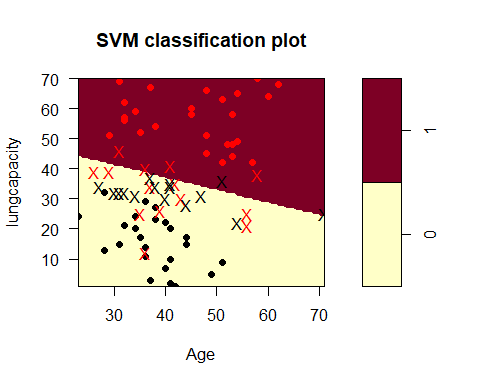
summary(svmfit2)

##   
## Call:  
## svm(formula = remission ~ lungcapacity + Age, data = train, kernel = "linear",   
## cost = 10, scale = FALSE)  
##   
##   
## Parameters:  
## SVM-Type: C-classification   
## SVM-Kernel: linear   
## cost: 10   
##   
## Number of Support Vectors: 29  
##   
## ( 14 15 )  
##   
##   
## Number of Classes: 2   
##   
## Levels:   
## 0 1

dat = data.frame(lungcapacity = train$lungcapacity, Age = train$Age, remission = as.factor( train$remission ))  
dat

## lungcapacity Age remission  
## 1 70 58 1  
## 2 65 54 1  
## 3 69 31 1  
## 4 67 37 1  
## 5 51 29 1  
## 6 63 51 1  
## 7 46 31 1  
## 8 48 53 1  
## 9 42 57 1  
## 10 39 29 1  
## 11 49 54 1  
## 12 45 48 1  
## 13 52 35 1  
## 14 58 53 1  
## 15 41 41 1  
## 16 39 26 1  
## 17 44 53 1  
## 18 68 62 1  
## 19 64 60 1  
## 20 51 48 1  
## 21 54 38 1  
## 22 42 51 1  
## 23 60 45 1  
## 24 21 56 1  
## 25 25 35 1  
## 26 58 45 1  
## 27 66 48 1  
## 28 62 32 1  
## 29 48 52 1  
## 30 57 32 1  
## 31 35 42 1  
## 32 26 39 1  
## 33 25 56 1  
## 34 34 37 1  
## 35 56 32 1  
## 36 40 36 1  
## 37 30 43 1  
## 38 12 36 1  
## 39 38 58 1  
## 40 59 34 1  
## 41 11 36 0  
## 42 2 41 0  
## 43 3 37 0  
## 44 29 36 0  
## 45 34 38 0  
## 46 10 41 0  
## 47 25 71 0  
## 48 34 41 0  
## 49 1 42 0  
## 50 23 38 0  
## 51 17 44 0  
## 52 21 32 0  
## 53 30 40 0  
## 54 5 49 0  
## 55 20 41 0  
## 56 15 44 0  
## 57 24 34 0  
## 58 17 35 0  
## 59 9 51 0  
## 60 32 32 0  
## 61 7 40 0  
## 62 31 34 0  
## 63 34 27 0  
## 64 22 54 0  
## 65 32 28 0  
## 66 35 41 0  
## 67 27 38 0  
## 68 24 23 0  
## 69 13 28 0  
## 70 14 36 0  
## 71 22 40 0  
## 72 37 37 0  
## 73 32 31 0  
## 74 20 34 0  
## 75 28 44 0  
## 76 15 31 0  
## 77 36 51 0  
## 78 32 30 0  
## 79 31 47 0

# limites de decis?o:  
plot(svmfit2, dat, grid = 100, svSymbol = "X", dataSymbol = 19)



O SVM apresentou resultados satisfatórios, porém, foram utilizadas apenas a variáveis contínuas. Dessa forma, vamos avaliar o desfecho remissão com Árvore de Decisão:

set.seed(205650)  
indexes <- createDataPartition(data$remission, p = 0.85, list = FALSE)  
train = data[indexes, ]  
test = data[-indexes, ]  
  
model\_fit <- rpart(remission ~ ., data = train, method = "class", control = list(minsplit = 20))  
predicted\_train <- predict(model\_fit, newdata = train, "class")  
predicted\_test <- predict(model\_fit, newdata = test, "class")  
summary(model\_fit)

## Call:  
## rpart(formula = remission ~ ., data = train, method = "class",   
## control = list(minsplit = 20))  
## n= 79   
##   
## CP nsplit rel error xerror xstd  
## 1 0.79487179 0 1.0000000 1.2051282 0.11187828  
## 2 0.01282051 1 0.2051282 0.2307692 0.07240893  
## 3 0.01000000 3 0.1794872 0.2820513 0.07889921  
##   
## Variable importance  
## lungcapacity SmokingHx IL6 Age tumorsize co2   
## 37 17 12 12 8 7   
## mobility WBC   
## 3 2   
##   
## Node number 1: 79 observations, complexity param=0.7948718  
## predicted class=1 expected loss=0.4936709 P(node) =1  
## class counts: 39 40  
## probabilities: 0.494 0.506   
## left son=2 (47 obs) right son=3 (32 obs)  
## Primary splits:  
## lungcapacity < 37.5 to the left, improve=26.217080, (0 missing)  
## IL6 < 62.5 to the right, improve=13.295150, (0 missing)  
## SmokingHx splits as LLR, improve=10.734350, (0 missing)  
## RBC < 76.5 to the right, improve= 9.493671, (0 missing)  
## Age < 44.5 to the left, improve= 5.357241, (0 missing)  
## Surrogate splits:  
## SmokingHx splits as LLR, agree=0.785, adj=0.469, (0 split)  
## Age < 44.5 to the left, agree=0.722, adj=0.312, (0 split)  
## IL6 < 62.5 to the right, agree=0.722, adj=0.312, (0 split)  
## tumorsize < 75.5 to the left, agree=0.671, adj=0.188, (0 split)  
## co2 < 86 to the left, agree=0.671, adj=0.188, (0 split)  
##   
## Node number 2: 47 observations, complexity param=0.01282051  
## predicted class=0 expected loss=0.1702128 P(node) =0.5949367  
## class counts: 39 8  
## probabilities: 0.830 0.170   
## left son=4 (16 obs) right son=5 (31 obs)  
## Primary splits:  
## WBC < 58.5 to the right, improve=1.405628, (0 missing)  
## RBC < 29.5 to the left, improve=1.405628, (0 missing)  
## mobility < 4.5 to the right, improve=1.341461, (0 missing)  
## pain < 2.5 to the right, improve=1.098024, (0 missing)  
## DID splits as RLLL, improve=1.074576, (0 missing)  
## Surrogate splits:  
## RBC < 10.5 to the left, agree=0.745, adj=0.250, (0 split)  
## tumorsize < 84.5 to the right, agree=0.702, adj=0.125, (0 split)  
## lungcapacity < 35.5 to the right, agree=0.702, adj=0.125, (0 split)  
## IL6 < 87.5 to the right, agree=0.702, adj=0.125, (0 split)  
## pain < 7.5 to the right, agree=0.681, adj=0.063, (0 split)  
##   
## Node number 3: 32 observations  
## predicted class=1 expected loss=0 P(node) =0.4050633  
## class counts: 0 32  
## probabilities: 0.000 1.000   
##   
## Node number 4: 16 observations  
## predicted class=0 expected loss=0 P(node) =0.2025316  
## class counts: 16 0  
## probabilities: 1.000 0.000   
##   
## Node number 5: 31 observations, complexity param=0.01282051  
## predicted class=0 expected loss=0.2580645 P(node) =0.3924051  
## class counts: 23 8  
## probabilities: 0.742 0.258   
## left son=10 (22 obs) right son=11 (9 obs)  
## Primary splits:  
## mobility < 5.5 to the right, improve=2.244705, (0 missing)  
## RBC < 75.5 to the right, improve=1.966206, (0 missing)  
## WBC < 41.5 to the left, improve=1.775730, (0 missing)  
## lungcapacity < 24.5 to the left, improve=1.484413, (0 missing)  
## Age < 34.5 to the left, improve=1.436185, (0 missing)  
## Surrogate splits:  
## tumorsize < 46.5 to the left, agree=0.774, adj=0.222, (0 split)  
## Age < 30.5 to the right, agree=0.742, adj=0.111, (0 split)  
## IL6 < 6.5 to the right, agree=0.742, adj=0.111, (0 split)  
## nmorphine < 4.5 to the right, agree=0.742, adj=0.111, (0 split)  
##   
## Node number 10: 22 observations  
## predicted class=0 expected loss=0.1363636 P(node) =0.278481  
## class counts: 19 3  
## probabilities: 0.864 0.136   
##   
## Node number 11: 9 observations  
## predicted class=1 expected loss=0.4444444 P(node) =0.1139241  
## class counts: 4 5  
## probabilities: 0.444 0.556

# importancia das vars  
model\_fit$variable.importance

## lungcapacity SmokingHx IL6 Age tumorsize co2   
## 26.39277864 12.28925397 8.61795116 8.44224766 5.59022845 4.91570159   
## mobility WBC RBC nmorphine pain   
## 2.24470512 1.40562800 0.35140700 0.24941168 0.08785175

#vars que foram usadas na arvore  
printcp(model\_fit)

##   
## Classification tree:  
## rpart(formula = remission ~ ., data = train, method = "class",   
## control = list(minsplit = 20))  
##   
## Variables actually used in tree construction:  
## [1] lungcapacity mobility WBC   
##   
## Root node error: 39/79 = 0.49367  
##   
## n= 79   
##   
## CP nsplit rel error xerror xstd  
## 1 0.794872 0 1.00000 1.20513 0.111878  
## 2 0.012821 1 0.20513 0.23077 0.072409  
## 3 0.010000 3 0.17949 0.28205 0.078899

#Matriz de confusao  
ctable\_train <- confusionMatrix(predicted\_train, as.factor(train$remission), positive = "1")  
ctable\_test <- confusionMatrix(predicted\_test, as.factor(test$remission), positive = "1")  
ctable\_train$table

## Reference  
## Prediction 0 1  
## 0 35 3  
## 1 4 37

ctable\_test$table

## Reference  
## Prediction 0 1  
## 0 6 0  
## 1 0 6

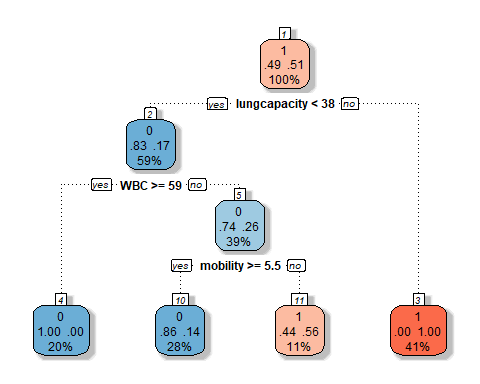
ctable\_train

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 35 3  
## 1 4 37  
##   
## Accuracy : 0.9114   
## 95% CI : (0.8259, 0.9636)  
## No Information Rate : 0.5063   
## P-Value [Acc > NIR] : 1.201e-14   
##   
## Kappa : 0.8227   
##   
## Mcnemar's Test P-Value : 1   
##   
## Sensitivity : 0.9250   
## Specificity : 0.8974   
## Pos Pred Value : 0.9024   
## Neg Pred Value : 0.9211   
## Prevalence : 0.5063   
## Detection Rate : 0.4684   
## Detection Prevalence : 0.5190   
## Balanced Accuracy : 0.9112   
##   
## 'Positive' Class : 1   
##

ctable\_test

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 6 0  
## 1 0 6  
##   
## Accuracy : 1   
## 95% CI : (0.7354, 1)  
## No Information Rate : 0.5   
## P-Value [Acc > NIR] : 0.0002441   
##   
## Kappa : 1   
##   
## Mcnemar's Test P-Value : NA   
##   
## Sensitivity : 1.0   
## Specificity : 1.0   
## Pos Pred Value : 1.0   
## Neg Pred Value : 1.0   
## Prevalence : 0.5   
## Detection Rate : 0.5   
## Detection Prevalence : 0.5   
## Balanced Accuracy : 1.0   
##   
## 'Positive' Class : 1   
##

fancyRpartPlot(model\_fit, yesno = 2, caption = NULL, palettes=c("Blues", "Reds"), tweak = 1)



dev.off()

## pdf   
## 3

Observamos uma acurácia de 91% no treino e 100% no teste. São resultados muito satisfatórios, porém é importante ressaltar que o banco possui 91 observações, ou seja, apenas 12 observações foram utilizadas para o teste.

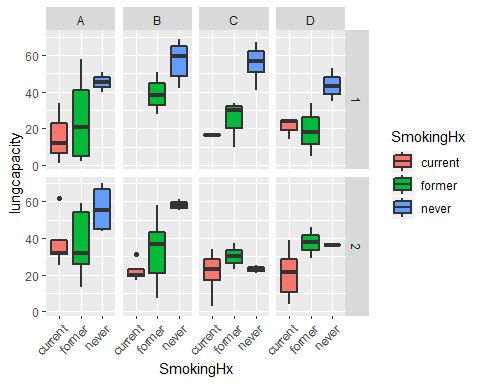
# Avaliação da identificação se o paciente é fumante

Regressao logistica ordinal

head(data)

## tumorsize co2 lungcapacity Age WBC RBC IL6 pain mobility ntumors nmorphine  
## 1 64 37 70 58 25 26 30 2 3 5 1  
## 2 41 24 65 54 50 23 42 6 4 5 11  
## 3 33 4 69 31 96 22 44 1 3 5 1  
## 4 86 62 67 37 85 46 48 7 4 4 13  
## 5 40 13 51 29 92 16 54 4 8 4 7  
## 6 89 49 63 51 68 3 33 8 2 9 12  
## remission Married FamilyHx SmokingHx Sex DID  
## 1 1 0 0 never 2 A  
## 2 1 0 1 never 1 C  
## 3 1 0 0 never 1 B  
## 4 1 1 1 never 1 C  
## 5 1 0 1 former 1 B  
## 6 1 1 0 never 1 B

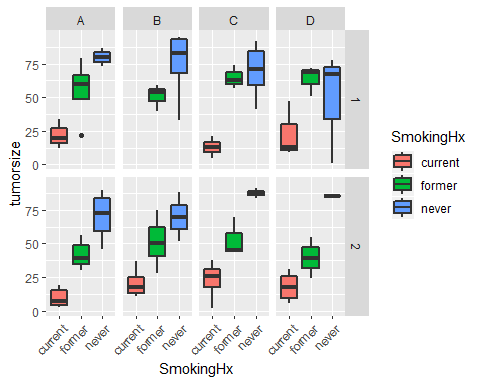
ggplot(data, aes(x = SmokingHx, y = lungcapacity, fill = SmokingHx)) + geom\_boxplot(size = .75) + facet\_grid(Sex ~ DID, margins = FALSE) + theme(axis.text.x = element\_text(angle = 45, hjust = 1, vjust = 1))



head(data)

## tumorsize co2 lungcapacity Age WBC RBC IL6 pain mobility ntumors nmorphine  
## 1 64 37 70 58 25 26 30 2 3 5 1  
## 2 41 24 65 54 50 23 42 6 4 5 11  
## 3 33 4 69 31 96 22 44 1 3 5 1  
## 4 86 62 67 37 85 46 48 7 4 4 13  
## 5 40 13 51 29 92 16 54 4 8 4 7  
## 6 89 49 63 51 68 3 33 8 2 9 12  
## remission Married FamilyHx SmokingHx Sex DID  
## 1 1 0 0 never 2 A  
## 2 1 0 1 never 1 C  
## 3 1 0 0 never 1 B  
## 4 1 1 1 never 1 C  
## 5 1 0 1 former 1 B  
## 6 1 1 0 never 1 B

ggplot(data, aes(x = SmokingHx, y = tumorsize, fill = SmokingHx)) + geom\_boxplot(size = .75) + facet\_grid(Sex ~ DID, margins = FALSE) + theme(axis.text.x = element\_text(angle = 45, hjust = 1, vjust = 1))



head(data)

## tumorsize co2 lungcapacity Age WBC RBC IL6 pain mobility ntumors nmorphine  
## 1 64 37 70 58 25 26 30 2 3 5 1  
## 2 41 24 65 54 50 23 42 6 4 5 11  
## 3 33 4 69 31 96 22 44 1 3 5 1  
## 4 86 62 67 37 85 46 48 7 4 4 13  
## 5 40 13 51 29 92 16 54 4 8 4 7  
## 6 89 49 63 51 68 3 33 8 2 9 12  
## remission Married FamilyHx SmokingHx Sex DID  
## 1 1 0 0 never 2 A  
## 2 1 0 1 never 1 C  
## 3 1 0 0 never 1 B  
## 4 1 1 1 never 1 C  
## 5 1 0 1 former 1 B  
## 6 1 1 0 never 1 B

str(data)

## 'data.frame': 91 obs. of 17 variables:  
## $ tumorsize : num 64 41 33 86 40 89 88 24 93 66 ...  
## $ co2 : num 37 24 4 62 13 49 72 51 94 29 ...  
## $ lungcapacity: num 70 65 69 67 51 63 55 46 48 42 ...  
## $ Age : int 58 54 31 37 29 51 52 31 53 57 ...  
## $ WBC : num 25 50 96 85 92 68 61 33 83 21 ...  
## $ RBC : num 26 23 22 46 16 3 17 2 45 57 ...  
## $ IL6 : num 30 42 44 48 54 33 82 25 52 34 ...  
## $ pain : int 2 6 1 7 4 8 6 2 8 6 ...  
## $ mobility : int 3 4 3 4 8 2 3 7 6 2 ...  
## $ ntumors : int 5 5 5 4 4 9 5 4 4 4 ...  
## $ nmorphine : int 1 11 1 13 7 12 9 0 12 10 ...  
## $ remission : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 2 2 2 2 ...  
## $ Married : Factor w/ 2 levels "0","1": 1 1 1 2 1 2 1 2 2 2 ...  
## $ FamilyHx : Factor w/ 2 levels "0","1": 1 2 1 2 2 1 2 1 1 1 ...  
## $ SmokingHx : Factor w/ 3 levels "current","former",..: 3 3 3 3 2 3 3 2 3 2 ...  
## $ Sex : Factor w/ 2 levels "1","2": 2 1 1 1 1 1 2 2 1 2 ...  
## $ DID : Factor w/ 4 levels "A","B","C","D": 1 3 2 3 2 2 2 4 2 2 ...  
## - attr(\*, "na.action")= 'omit' Named int 18 74 85 92 95  
## ..- attr(\*, "names")= chr "18" "74" "85" "92" ...

set.seed(205650)  
indexes <- createDataPartition(data$SmokingHx, p = 0.85, list = FALSE)  
train = data[indexes, ]  
test = data[-indexes, ]  
  
ord\_reg <- polr(SmokingHx~tumorsize+pain+mobility+ntumors+Married+Sex+DID+IL6+co2+lungcapacity+nmorphine, data = train, Hess = TRUE)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

summary(ord\_reg)

## Call:  
## polr(formula = SmokingHx ~ tumorsize + pain + mobility + ntumors +   
## Married + Sex + DID + IL6 + co2 + lungcapacity + nmorphine,   
## data = train, Hess = TRUE)  
##   
## Coefficients:  
## Value Std. Error t value  
## tumorsize 0.162436 0.03937 4.12548  
## pain -0.171914 0.64929 -0.26477  
## mobility 0.003893 0.20752 0.01876  
## ntumors -0.068765 0.20812 -0.33040  
## Married1 -0.388719 0.89364 -0.43498  
## Sex2 -0.894084 1.02418 -0.87298  
## DIDB 0.348756 1.18245 0.29494  
## DIDC 0.229029 1.16977 0.19579  
## DIDD -0.200678 1.30639 -0.15361  
## IL6 0.004919 0.01791 0.27462  
## co2 0.014119 0.02027 0.69664  
## lungcapacity 0.180015 0.04477 4.02118  
## nmorphine 0.101087 0.29489 0.34279  
##   
## Intercepts:  
## Value Std. Error t value  
## current|former 10.4014 3.5750 2.9095  
## former|never 17.5455 4.6788 3.7500  
##   
## Residual Deviance: 47.55333   
## AIC: 77.55333

summary\_table <- coef(summary(ord\_reg))  
pval <- pnorm(abs(summary\_table[, "t value"]),lower.tail = FALSE)\* 2  
summary\_table <- cbind(summary\_table, "p value" = round(pval,3))  
summary\_table

## Value Std. Error t value p value  
## tumorsize 0.162436083 0.03937387 4.12547919 0.000  
## pain -0.171914347 0.64929050 -0.26477262 0.791  
## mobility 0.003892824 0.20751714 0.01875905 0.985  
## ntumors -0.068764896 0.20812338 -0.33040448 0.741  
## Married1 -0.388719228 0.89364081 -0.43498375 0.664  
## Sex2 -0.894083776 1.02417560 -0.87297898 0.383  
## DIDB 0.348756358 1.18245091 0.29494363 0.768  
## DIDC 0.229029463 1.16977343 0.19578959 0.845  
## DIDD -0.200677621 1.30638985 -0.15361235 0.878  
## IL6 0.004918791 0.01791121 0.27462075 0.784  
## co2 0.014118539 0.02026648 0.69664474 0.486  
## lungcapacity 0.180015280 0.04476683 4.02117549 0.000  
## nmorphine 0.101087255 0.29489478 0.34279092 0.732  
## current|former 10.401419427 3.57502201 2.90947004 0.004  
## former|never 17.545541256 4.67884890 3.74996963 0.000

vars com p-value < 0.05: tumorsize, lungcapacity

ord\_reg\_sg <- polr(SmokingHx~tumorsize+lungcapacity, data = train, Hess = TRUE)  
summary(ord\_reg\_sg)

## Call:  
## polr(formula = SmokingHx ~ tumorsize + lungcapacity, data = train,   
## Hess = TRUE)  
##   
## Coefficients:  
## Value Std. Error t value  
## tumorsize 0.1714 0.03489 4.911  
## lungcapacity 0.1670 0.03734 4.471  
##   
## Intercepts:  
## Value Std. Error t value  
## current|former 10.7058 2.2569 4.7435  
## former|never 17.1553 3.4741 4.9380  
##   
## Residual Deviance: 49.77701   
## AIC: 57.77701

summary\_table <- coef(summary(ord\_reg\_sg))  
pval <- pnorm(abs(summary\_table[, "t value"]),lower.tail = FALSE)\* 2  
summary\_table <- cbind(summary\_table, "p value" = round(pval,3))  
summary\_table

## Value Std. Error t value p value  
## tumorsize 0.1713530 0.03488902 4.911374 0  
## lungcapacity 0.1669734 0.03734320 4.471320 0  
## current|former 10.7057743 2.25691851 4.743536 0  
## former|never 17.1553156 3.47411084 4.938045 0

Comparando:

predsig<- predict(ord\_reg\_sg, newdata = train)  
pred <- predict(ord\_reg, newdata = train)  
confusionMatrix(table(predsig, train$SmokingHx))

## Confusion Matrix and Statistics  
##   
##   
## predsig current former never  
## current 24 1 1  
## former 3 24 0  
## never 0 2 24  
##   
## Overall Statistics  
##   
## Accuracy : 0.9114   
## 95% CI : (0.8259, 0.9636)  
## No Information Rate : 0.3418   
## P-Value [Acc > NIR] : <2e-16   
##   
## Kappa : 0.8671   
##   
## Mcnemar's Test P-Value : 0.2615   
##   
## Statistics by Class:  
##   
## Class: current Class: former Class: never  
## Sensitivity 0.8889 0.8889 0.9600  
## Specificity 0.9615 0.9423 0.9630  
## Pos Pred Value 0.9231 0.8889 0.9231  
## Neg Pred Value 0.9434 0.9423 0.9811  
## Prevalence 0.3418 0.3418 0.3165  
## Detection Rate 0.3038 0.3038 0.3038  
## Detection Prevalence 0.3291 0.3418 0.3291  
## Balanced Accuracy 0.9252 0.9156 0.9615

confusionMatrix(table(pred, train$SmokingHx))

## Confusion Matrix and Statistics  
##   
##   
## pred current former never  
## current 25 1 1  
## former 2 23 0  
## never 0 3 24  
##   
## Overall Statistics  
##   
## Accuracy : 0.9114   
## 95% CI : (0.8259, 0.9636)  
## No Information Rate : 0.3418   
## P-Value [Acc > NIR] : <2e-16   
##   
## Kappa : 0.8671   
##   
## Mcnemar's Test P-Value : 0.2276   
##   
## Statistics by Class:  
##   
## Class: current Class: former Class: never  
## Sensitivity 0.9259 0.8519 0.9600  
## Specificity 0.9615 0.9615 0.9444  
## Pos Pred Value 0.9259 0.9200 0.8889  
## Neg Pred Value 0.9615 0.9259 0.9808  
## Prevalence 0.3418 0.3418 0.3165  
## Detection Rate 0.3165 0.2911 0.3038  
## Detection Prevalence 0.3418 0.3165 0.3418  
## Balanced Accuracy 0.9437 0.9067 0.9522

predtest <- predict(ord\_reg, newdata = test)  
confusionMatrix(table(predtest, test$SmokingHx))

## Confusion Matrix and Statistics  
##   
##   
## predtest current former never  
## current 4 0 0  
## former 0 2 1  
## never 0 2 3  
##   
## Overall Statistics  
##   
## Accuracy : 0.75   
## 95% CI : (0.4281, 0.9451)  
## No Information Rate : 0.3333   
## P-Value [Acc > NIR] : 0.003856   
##   
## Kappa : 0.625   
##   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: current Class: former Class: never  
## Sensitivity 1.0000 0.5000 0.7500  
## Specificity 1.0000 0.8750 0.7500  
## Pos Pred Value 1.0000 0.6667 0.6000  
## Neg Pred Value 1.0000 0.7778 0.8571  
## Prevalence 0.3333 0.3333 0.3333  
## Detection Rate 0.3333 0.1667 0.2500  
## Detection Prevalence 0.3333 0.2500 0.4167  
## Balanced Accuracy 1.0000 0.6875 0.7500

Árvore:

set.seed(205650)  
indexes <- createDataPartition(data$SmokingHx, p = 0.85, list = FALSE)  
train = data[indexes, ]  
test = data[-indexes, ]  
model\_fit <- rpart(SmokingHx ~ ., data = train, method = "class", control = list(minsplit = 20))  
predicted\_train <- predict(model\_fit, newdata = train, "class")  
predicted\_test <- predict(model\_fit, newdata = test, "class")  
summary(model\_fit)

## Call:  
## rpart(formula = SmokingHx ~ ., data = train, method = "class",   
## control = list(minsplit = 20))  
## n= 79   
##   
## CP nsplit rel error xerror xstd  
## 1 0.40384615 0 1.0000000 1.2884615 0.06134952  
## 2 0.30769231 1 0.5961538 0.8461538 0.08490700  
## 3 0.05769231 2 0.2884615 0.3846154 0.07432315  
## 4 0.03846154 3 0.2307692 0.3846154 0.07432315  
## 5 0.01000000 4 0.1923077 0.4038462 0.07551040  
##   
## Variable importance  
## tumorsize lungcapacity co2 remission mobility WBC   
## 26 23 11 10 9 5   
## RBC IL6 nmorphine Age pain   
## 5 4 3 3 1   
##   
## Node number 1: 79 observations, complexity param=0.4038462  
## predicted class=current expected loss=0.6582278 P(node) =1  
## class counts: 27 27 25  
## probabilities: 0.342 0.342 0.316   
## left son=2 (31 obs) right son=3 (48 obs)  
## Primary splits:  
## tumorsize < 34.5 to the left, improve=16.564360, (0 missing)  
## lungcapacity < 39.5 to the left, improve=13.974150, (0 missing)  
## remission splits as LR, improve= 8.497252, (0 missing)  
## mobility < 8.5 to the right, improve= 5.904259, (0 missing)  
## co2 < 45.5 to the left, improve= 5.058909, (0 missing)  
## Surrogate splits:  
## mobility < 8.5 to the right, agree=0.772, adj=0.419, (0 split)  
## co2 < 22.5 to the left, agree=0.747, adj=0.355, (0 split)  
## lungcapacity < 25.5 to the left, agree=0.671, adj=0.161, (0 split)  
## nmorphine < 4.5 to the left, agree=0.658, adj=0.129, (0 split)  
## WBC < 4.5 to the left, agree=0.646, adj=0.097, (0 split)  
##   
## Node number 2: 31 observations, complexity param=0.05769231  
## predicted class=current expected loss=0.1935484 P(node) =0.3924051  
## class counts: 25 4 2  
## probabilities: 0.806 0.129 0.065   
## left son=4 (24 obs) right son=5 (7 obs)  
## Primary splits:  
## lungcapacity < 41 to the left, improve=6.1935480, (0 missing)  
## remission splits as LR, improve=2.1935480, (0 missing)  
## tumorsize < 21.5 to the left, improve=1.7988120, (0 missing)  
## IL6 < 57.5 to the right, improve=1.1409170, (0 missing)  
## WBC < 25.5 to the left, improve=0.9935484, (0 missing)  
## Surrogate splits:  
## co2 < 87 to the left, agree=0.839, adj=0.286, (0 split)  
## WBC < 91.5 to the left, agree=0.839, adj=0.286, (0 split)  
## tumorsize < 3.5 to the right, agree=0.806, adj=0.143, (0 split)  
## mobility < 3.5 to the right, agree=0.806, adj=0.143, (0 split)  
## nmorphine < 0.5 to the right, agree=0.806, adj=0.143, (0 split)  
##   
## Node number 3: 48 observations, complexity param=0.3076923  
## predicted class=former expected loss=0.5208333 P(node) =0.6075949  
## class counts: 2 23 23  
## probabilities: 0.042 0.479 0.479   
## left son=6 (18 obs) right son=7 (30 obs)  
## Primary splits:  
## lungcapacity < 34.5 to the left, improve=11.586110, (0 missing)  
## remission splits as LR, improve= 9.991152, (0 missing)  
## tumorsize < 66.5 to the left, improve= 9.621032, (0 missing)  
## nmorphine < 8.5 to the left, improve= 5.352314, (0 missing)  
## RBC < 64.5 to the right, improve= 4.868939, (0 missing)  
## Surrogate splits:  
## remission splits as LR, agree=0.938, adj=0.833, (0 split)  
## RBC < 77.5 to the right, agree=0.771, adj=0.389, (0 split)  
## IL6 < 60.5 to the right, agree=0.750, adj=0.333, (0 split)  
## tumorsize < 64.5 to the left, agree=0.708, adj=0.222, (0 split)  
## Age < 46.5 to the left, agree=0.708, adj=0.222, (0 split)  
##   
## Node number 4: 24 observations  
## predicted class=current expected loss=0 P(node) =0.3037975  
## class counts: 24 0 0  
## probabilities: 1.000 0.000 0.000   
##   
## Node number 5: 7 observations  
## predicted class=former expected loss=0.4285714 P(node) =0.08860759  
## class counts: 1 4 2  
## probabilities: 0.143 0.571 0.286   
##   
## Node number 6: 18 observations  
## predicted class=former expected loss=0.1111111 P(node) =0.2278481  
## class counts: 2 16 0  
## probabilities: 0.111 0.889 0.000   
##   
## Node number 7: 30 observations, complexity param=0.03846154  
## predicted class=never expected loss=0.2333333 P(node) =0.3797468  
## class counts: 0 7 23  
## probabilities: 0.000 0.233 0.767   
## left son=14 (12 obs) right son=15 (18 obs)  
## Primary splits:  
## tumorsize < 66.5 to the left, improve=4.900000, (0 missing)  
## co2 < 39.5 to the left, improve=4.828571, (0 missing)  
## pain < 4.5 to the left, improve=3.346970, (0 missing)  
## nmorphine < 8 to the left, improve=2.669841, (0 missing)  
## DID splits as LLRR, improve=1.633333, (0 missing)  
## Surrogate splits:  
## co2 < 43.5 to the left, agree=0.867, adj=0.667, (0 split)  
## lungcapacity < 58 to the right, agree=0.700, adj=0.250, (0 split)  
## WBC < 60 to the left, agree=0.700, adj=0.250, (0 split)  
## pain < 4.5 to the left, agree=0.667, adj=0.167, (0 split)  
## mobility < 7.5 to the right, agree=0.667, adj=0.167, (0 split)  
##   
## Node number 14: 12 observations  
## predicted class=former expected loss=0.4166667 P(node) =0.1518987  
## class counts: 0 7 5  
## probabilities: 0.000 0.583 0.417   
##   
## Node number 15: 18 observations  
## predicted class=never expected loss=0 P(node) =0.2278481  
## class counts: 0 0 18  
## probabilities: 0.000 0.000 1.000

# importancia das vars  
model\_fit$variable.importance

## tumorsize lungcapacity co2 remission mobility WBC   
## 24.9238470 21.6763310 10.9139291 9.6550926 8.6478051 4.5975881   
## RBC IL6 nmorphine Age pain   
## 4.5057099 3.8620370 3.0221298 2.5746914 0.8166667

#vars que foram usadas na arvore  
printcp(model\_fit)

##   
## Classification tree:  
## rpart(formula = SmokingHx ~ ., data = train, method = "class",   
## control = list(minsplit = 20))  
##   
## Variables actually used in tree construction:  
## [1] lungcapacity tumorsize   
##   
## Root node error: 52/79 = 0.65823  
##   
## n= 79   
##   
## CP nsplit rel error xerror xstd  
## 1 0.403846 0 1.00000 1.28846 0.061350  
## 2 0.307692 1 0.59615 0.84615 0.084907  
## 3 0.057692 2 0.28846 0.38462 0.074323  
## 4 0.038462 3 0.23077 0.38462 0.074323  
## 5 0.010000 4 0.19231 0.40385 0.075510

#Matriz de confusao  
ctable\_train <- confusionMatrix(predicted\_train, as.factor(train$SmokingHx))  
ctable\_test <- confusionMatrix(predicted\_test, as.factor(test$SmokingHx))  
ctable\_train$table

## Reference  
## Prediction current former never  
## current 24 0 0  
## former 3 27 7  
## never 0 0 18

ctable\_test$table

## Reference  
## Prediction current former never  
## current 3 0 0  
## former 1 3 3  
## never 0 1 1

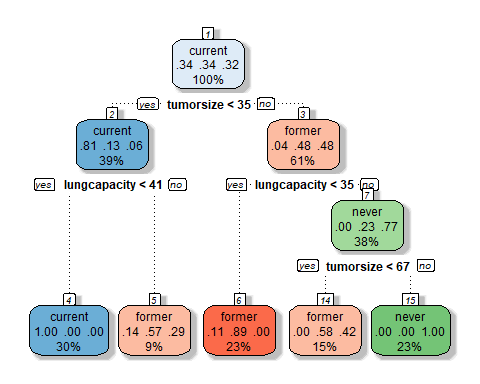
ctable\_train

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction current former never  
## current 24 0 0  
## former 3 27 7  
## never 0 0 18  
##   
## Overall Statistics  
##   
## Accuracy : 0.8734   
## 95% CI : (0.7795, 0.9376)  
## No Information Rate : 0.3418   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.8094   
##   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: current Class: former Class: never  
## Sensitivity 0.8889 1.0000 0.7200  
## Specificity 1.0000 0.8077 1.0000  
## Pos Pred Value 1.0000 0.7297 1.0000  
## Neg Pred Value 0.9455 1.0000 0.8852  
## Prevalence 0.3418 0.3418 0.3165  
## Detection Rate 0.3038 0.3418 0.2278  
## Detection Prevalence 0.3038 0.4684 0.2278  
## Balanced Accuracy 0.9444 0.9038 0.8600

ctable\_test

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction current former never  
## current 3 0 0  
## former 1 3 3  
## never 0 1 1  
##   
## Overall Statistics  
##   
## Accuracy : 0.5833   
## 95% CI : (0.2767, 0.8483)  
## No Information Rate : 0.3333   
## P-Value [Acc > NIR] : 0.06645   
##   
## Kappa : 0.375   
##   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: current Class: former Class: never  
## Sensitivity 0.7500 0.7500 0.25000  
## Specificity 1.0000 0.5000 0.87500  
## Pos Pred Value 1.0000 0.4286 0.50000  
## Neg Pred Value 0.8889 0.8000 0.70000  
## Prevalence 0.3333 0.3333 0.33333  
## Detection Rate 0.2500 0.2500 0.08333  
## Detection Prevalence 0.2500 0.5833 0.16667  
## Balanced Accuracy 0.8750 0.6250 0.56250

fancyRpartPlot(model\_fit, yesno = 2, caption = NULL, palettes=c("Blues", "Reds", "Greens"), tweak = 1)



dev.off()

## pdf   
## 3

## Remissao + vars definidas

Utilize um modelo apropriado para analisar a relação entre o desfecho remissão e as variáveis sexo, histórico familiar, estado civil e DID.

tarefa 3 - Reg logistica tarefa 4 - tree

model\_fit <- rpart(remission~Sex+FamilyHx+Married+DID, data = data, method = "class", control = list(minsplit = 20))  
predicted\_train <- predict(model\_fit, newdata = data, "class")  
predicted\_test <- predict(model\_fit, newdata = data, "class")  
summary(model\_fit)

## Call:  
## rpart(formula = remission ~ Sex + FamilyHx + Married + DID, data = data,   
## method = "class", control = list(minsplit = 20))  
## n= 91   
##   
## CP nsplit rel error xerror xstd  
## 1 0.28888889 0 1.0000000 1.1777778 0.1045434  
## 2 0.02222222 1 0.7111111 0.8222222 0.1041273  
## 3 0.01000000 5 0.6222222 0.8888889 0.1052160  
##   
## Variable importance  
## DID Sex FamilyHx Married   
## 55 34 9 2   
##   
## Node number 1: 91 observations, complexity param=0.2888889  
## predicted class=1 expected loss=0.4945055 P(node) =1  
## class counts: 45 46  
## probabilities: 0.495 0.505   
## left son=2 (43 obs) right son=3 (48 obs)  
## Primary splits:  
## DID splits as RRLL, improve=4.001288000, (0 missing)  
## Sex splits as RL, improve=2.470115000, (0 missing)  
## FamilyHx splits as RL, improve=0.049094870, (0 missing)  
## Married splits as LR, improve=0.008019008, (0 missing)  
## Surrogate splits:  
## Sex splits as LR, agree=0.538, adj=0.023, (0 split)  
## FamilyHx splits as RL, agree=0.538, adj=0.023, (0 split)  
##   
## Node number 2: 43 observations, complexity param=0.02222222  
## predicted class=0 expected loss=0.3488372 P(node) =0.4725275  
## class counts: 28 15  
## probabilities: 0.651 0.349   
## left son=4 (22 obs) right son=5 (21 obs)  
## Primary splits:  
## Sex splits as RL, improve=2.51323900, (0 missing)  
## FamilyHx splits as LR, improve=0.17836200, (0 missing)  
## DID splits as --RL, improve=0.17836200, (0 missing)  
## Married splits as LR, improve=0.03488372, (0 missing)  
## Surrogate splits:  
## FamilyHx splits as LR, agree=0.628, adj=0.238, (0 split)  
## Married splits as RL, agree=0.535, adj=0.048, (0 split)  
## DID splits as --RL, agree=0.535, adj=0.048, (0 split)  
##   
## Node number 3: 48 observations, complexity param=0.02222222  
## predicted class=1 expected loss=0.3541667 P(node) =0.5274725  
## class counts: 17 31  
## probabilities: 0.354 0.646   
## left son=6 (28 obs) right son=7 (20 obs)  
## Primary splits:  
## Sex splits as RL, improve=0.744047600, (0 missing)  
## FamilyHx splits as RL, improve=0.245046600, (0 missing)  
## Married splits as RL, improve=0.053571430, (0 missing)  
## DID splits as RL--, improve=0.003550725, (0 missing)  
## Surrogate splits:  
## Married splits as RL, agree=0.646, adj=0.15, (0 split)  
##   
## Node number 4: 22 observations  
## predicted class=0 expected loss=0.1818182 P(node) =0.2417582  
## class counts: 18 4  
## probabilities: 0.818 0.182   
##   
## Node number 5: 21 observations, complexity param=0.02222222  
## predicted class=1 expected loss=0.4761905 P(node) =0.2307692  
## class counts: 10 11  
## probabilities: 0.476 0.524   
## left son=10 (9 obs) right son=11 (12 obs)  
## Primary splits:  
## DID splits as --RL, improve=0.19841270, (0 missing)  
## FamilyHx splits as RL, improve=0.04761905, (0 missing)  
## Married splits as LR, improve=0.04761905, (0 missing)  
## Surrogate splits:  
## FamilyHx splits as LR, agree=0.619, adj=0.111, (0 split)  
##   
## Node number 6: 28 observations, complexity param=0.02222222  
## predicted class=1 expected loss=0.4285714 P(node) =0.3076923  
## class counts: 12 16  
## probabilities: 0.429 0.571   
## left son=12 (14 obs) right son=13 (14 obs)  
## Primary splits:  
## DID splits as RL--, improve=1.142857000, (0 missing)  
## FamilyHx splits as RL, improve=0.285714300, (0 missing)  
## Married splits as RL, improve=0.005952381, (0 missing)  
## Surrogate splits:  
## FamilyHx splits as RL, agree=0.571, adj=0.143, (0 split)  
##   
## Node number 7: 20 observations  
## predicted class=1 expected loss=0.25 P(node) =0.2197802  
## class counts: 5 15  
## probabilities: 0.250 0.750   
##   
## Node number 10: 9 observations  
## predicted class=0 expected loss=0.4444444 P(node) =0.0989011  
## class counts: 5 4  
## probabilities: 0.556 0.444   
##   
## Node number 11: 12 observations  
## predicted class=1 expected loss=0.4166667 P(node) =0.1318681  
## class counts: 5 7  
## probabilities: 0.417 0.583   
##   
## Node number 12: 14 observations  
## predicted class=0 expected loss=0.4285714 P(node) =0.1538462  
## class counts: 8 6  
## probabilities: 0.571 0.429   
##   
## Node number 13: 14 observations  
## predicted class=1 expected loss=0.2857143 P(node) =0.1538462  
## class counts: 4 10  
## probabilities: 0.286 0.714

# importancia das vars  
model\_fit$variable.importance

## DID Sex FamilyHx Married   
## 5.4622363 3.3503395 0.8767545 0.2312852

#vars que foram usadas na arvore  
printcp(model\_fit)

##   
## Classification tree:  
## rpart(formula = remission ~ Sex + FamilyHx + Married + DID, data = data,   
## method = "class", control = list(minsplit = 20))  
##   
## Variables actually used in tree construction:  
## [1] DID Sex  
##   
## Root node error: 45/91 = 0.49451  
##   
## n= 91   
##   
## CP nsplit rel error xerror xstd  
## 1 0.288889 0 1.00000 1.17778 0.10454  
## 2 0.022222 1 0.71111 0.82222 0.10413  
## 3 0.010000 5 0.62222 0.88889 0.10522

#Matriz de confusao  
ctable\_train <- confusionMatrix(predicted\_train, as.factor(data$remission))  
ctable\_test <- confusionMatrix(predicted\_test, as.factor(data$remission))  
ctable\_train$table

## Reference  
## Prediction 0 1  
## 0 31 14  
## 1 14 32

ctable\_test$table

## Reference  
## Prediction 0 1  
## 0 31 14  
## 1 14 32

ctable\_train

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 31 14  
## 1 14 32  
##   
## Accuracy : 0.6923   
## 95% CI : (0.5868, 0.7849)  
## No Information Rate : 0.5055   
## P-Value [Acc > NIR] : 0.0002321   
##   
## Kappa : 0.3845   
##   
## Mcnemar's Test P-Value : 1.0000000   
##   
## Sensitivity : 0.6889   
## Specificity : 0.6957   
## Pos Pred Value : 0.6889   
## Neg Pred Value : 0.6957   
## Prevalence : 0.4945   
## Detection Rate : 0.3407   
## Detection Prevalence : 0.4945   
## Balanced Accuracy : 0.6923   
##   
## 'Positive' Class : 0   
##

ctable\_test

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 31 14  
## 1 14 32  
##   
## Accuracy : 0.6923   
## 95% CI : (0.5868, 0.7849)  
## No Information Rate : 0.5055   
## P-Value [Acc > NIR] : 0.0002321   
##   
## Kappa : 0.3845   
##   
## Mcnemar's Test P-Value : 1.0000000   
##   
## Sensitivity : 0.6889   
## Specificity : 0.6957   
## Pos Pred Value : 0.6889   
## Neg Pred Value : 0.6957   
## Prevalence : 0.4945   
## Detection Rate : 0.3407   
## Detection Prevalence : 0.4945   
## Balanced Accuracy : 0.6923   
##   
## 'Positive' Class : 0   
##

#install.packages('rattle')  
library(rattle)  
library(RColorBrewer)  
png('arvore\_cannabis.png')  
fancyRpartPlot(model\_fit, yesno = 2, caption = NULL, palettes=c("Blues", "Reds"), tweak = 1)  
dev.off()

## png   
## 2

<https://towardsdatascience.com/do-you-know-how-to-choose-the-right-machine-learning-algorithm-among-7-different-types-295d0b0c7f60>

<https://medium.com/datadriveninvestor/choosing-the-best-algorithm-for-your-classification-model-7c632c78f38f>

<http://jmlr.csail.mit.edu/papers/v3/guyon03a.html>

<https://towardsdatascience.com/implementing-and-interpreting-ordinal-logistic-regression-1ee699274cf5>

<https://www.datacamp.com/community/tutorials/neural-network-models-r>

<https://www.theanalysisfactor.com/logistic-regression-models-for-multinomial-and-ordinal-variables/>