Assignment 6

Luis Nicolas Luarte Rodriguez

```
# load libs
if (!require("pacman")) install.packages("pacman")
pacman::p_load(ggplot2, knitr, kableExtra, tidyverse, My.stepwise, reshape2,
    GGally, ggpubr, gtsummary, formatR, gt, olsrr, stargazer, boot, nnet,
    mlogit, xtable, generics, caret, rlist)
```

```
# load the dataset
setwd("/home/nicoluarte/uni/PHD/stat_course")
dataSet <- as_tibble(read.csv("ENS.csv", sep = ",", header = TRUE))
dataSet <- dataSet[-c(1), ]
head(dataSet)
```

1. Using the ENS 2016-17 database, build a multiclass logistic regression model to predict the cardiovascular risk obtained through the RCV CHILENO RECODIFICADO (low risk = 1, moderate risk = 2 and high risk = 3).

a. Use a multiclass logistic regression model with all predictors and present the results including p values

```
# recode variables and drop rows with at least 1 NA
dataFinal <- dataSet %>% type_convert(cols(Sexo = col_factor(), Glucosa = col_double(),
    Colesterol_HDL = col_double(), Colesterol_Total = col_double(), m2p11a_PAS = col_double(),
    m2p11a_PAD = col_double(), m7p3 = col_double(), ta3 = col_double(),
    IMC = col_double(), RCV_CHILENO_RECODIFICADO = col_factor())) %>% drop_na()
```

```
# fit model
mdlData <- mlogit.data(dataFinal, shape = "wide", choice = "RCV_CHILENO_RECODIFICADO")
mdl <- mlogit(RCV_CHILENO_RECODIFICADO ~ 1 | Edad + Sexo + Glucosa + Colesterol_HDL +
    Colesterol_Total + m2p11a_PAS + m2p11a_PAD + m7p3 + ta3 + IMC, data = mdlData,
    reflevel = "1")
summary(mdl)
```

Call: mlogit(formula = RCV_CHILENO_RECODIFICADO ~ 1 | Edad + Sexo + Glucosa + Colesterol_HDL + Colesterol_Total + m2p11a_PAS + m2p11a_PAD + m7p3 + ta3 + IMC, data = mdlData, reflevel = "1", method = "nr")

Frequencies of alternatives:choice 1 2 3 0.48201 0.24389 0.27410

nr method 7 iterations, 0h:0m:0s g'(-H)^-1g = 1.56E-06 successive function values within tolerance limits

Coefficients : Estimate Std. Error z-value Pr(>|z|)
(Intercept):2 -9.3805844 1.0325919 -9.0845 < 2.2e-16 ***(Intercept):3 -15.7116122 1.1539686 -13.6153 < 2.2e-16*** Edad:2 0.0294123 0.0046391 6.3401 2.296e-10 ***Edad:3 0.0801552 0.0056632 14.1536 < 2.2e-16*** Sexo1:2 -0.6445139 0.1453492 -4.4342 9.240e-06 ***Sexo1:3 -0.7937301 0.1624881 -4.8848 1.035e-06*** Glucosa:2 0.0394638 0.0070585 5.5910 2.258e-08 ***Glucosa:3 0.0805565 0.0071005 11.3453 < 2.2e-16*** Colesterol_HDL:2 -0.1030027 0.0070061 -14.7018 < 2.2e-16 ***Colesterol_HDL:3 -0.0779299 0.0070249 -11.0933 < 2.2e-16*** Colesterol_Total:2 0.0101533 0.0018266 5.5587 2.719e-08 ***Colesterol_Total:3 0.0098652 0.0019767 4.9907 6.016e-07*** m2p11a_PAS:2 0.0168723 0.0065644 2.5703 0.0101621 * m2p11a_PAS:3 0.0148855 0.0066516 2.2379 0.0252281 *

m2p11a_PAD:2 0.0400090 0.0114609 3.4909 0.0004814 ***m2p11a_PAD:3 0.0250097 0.0121021 2.0666 0.0387763***
m7p3:2 -0.1122808 0.0731644 -1.5346 0.1248731
m7p3:3 -0.0308528 0.0778916 -0.3961 0.6920319
ta3:2 -0.0484431 0.0530239 -0.9136 0.3609226
ta3:3 -0.1057411 0.0620387 -1.7044 0.0882993 .
IMC:2 0.0897137 0.0142285 6.3052 2.878e-10 *IMC:3 0.0826014 0.0158205 5.2211 1.778e-07* ** — Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -1524.5 McFadden R^2: 0.35542 Likelihood ratio test : chisq = 1681.1 (p.value = < 2.22e-16)

```
stargazer(summary(mdl)$CoefTable, title = "Model results")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Nov 30, 2020 - 12:38:30 AM

Table 1: Model results

	Estimate	Std. Error	z-value	Pr(>|z|)
(Intercept):2	-9.381	1.033	-9.085	0
(Intercept):3	-15.712	1.154	-13.615	0
Edad:2	0.029	0.005	6.340	0
Edad:3	0.080	0.006	14.154	0
Sexo1:2	-0.645	0.145	-4.434	0.00001
Sexo1:3	-0.794	0.162	-4.885	0.00000
Glucosa:2	0.039	0.007	5.591	0.00000
Glucosa:3	0.081	0.007	11.345	0
Colesterol_HDL:2	-0.103	0.007	-14.702	0
Colesterol_HDL:3	-0.078	0.007	-11.093	0
Colesterol_Total:2	0.010	0.002	5.559	0.00000
Colesterol_Total:3	0.010	0.002	4.991	0.00000
m2p11a_PAS:2	0.017	0.007	2.570	0.010
m2p11a_PAS:3	0.015	0.007	2.238	0.025
m2p11a_PAD:2	0.040	0.011	3.491	0.0005
m2p11a_PAD:3	0.025	0.012	2.067	0.039
m7p3:2	-0.112	0.073	-1.535	0.125
m7p3:3	-0.031	0.078	-0.396	0.692
ta3:2	-0.048	0.053	-0.914	0.361
ta3:3	-0.106	0.062	-1.704	0.088
IMC:2	0.090	0.014	6.305	0
IMC:3	0.083	0.016	5.221	0.00000

b. Interpret the results obtained. What are the variables that have a significant contribution to the model? Was this result expected?

First the full model coefficients are significantly different from the null model (Log-likelihood = -1524.5; $\chi^2 = 1681.1$; p = < 0.001). Predictive ability of the model is modest with McFadden $R^2 = 0.355$. As for coefficients, all were significant in the model, except for smoking and alcohol consumption. Coeffients are reported with coronary risk score '1' = low risk as reference. Considering age, we see a positive estimate for class '2' = moderate and class '3' = high risk, meaning that being older is more predictive of a higher coronary risk. More precisely, being 1 year older, increases the multinomial log-odds for coronary risk, relative to '1', would decrease by 0.029 and 0.08, for class '2' and '3', respectively. For sex, the comparison is female (Sexo1) against male (Sexo2), with negative coefficients meaning that females are less likely to have a moderate or high coronary risk score, compared to males and holding all other variables constant. Similarly, higher glucose increase the log-odds of moderate and higher risks. HDL cholesterol, unlike total cholesterol, may be

a protective factor against cardiovascular risk, as higher levels predicts less log-odds for classes '2' and '3'. Increases in both type of blood pressure are related to greater cardiovascular risk. Finally higher BMI is also associated with higher risk of cardiovascular risk. Most of the coefficients estimates were not surprising as they are typically related to cardiovascular health, however, alcohol consumption not reaching significance did surprise me, as I would expect that increasing alcohol consumption would lead to a worse health status, including cardiovascular health. Perhaps, alcohol consumption effects are made effective via blood pressure, so controlling for blood pressure renders alcohol consumption not significant.

c. Calculate the probability of belonging to each of the categories predicted by the model for a subject with the following characteristics: Age: 46, Sex: female, Basal glucose: 95, Cholesterol: 194, HDL: 47, PAS: 128, PAD: 76, Daily alcohol consumption: No, Gr. Salt: 10, Smoking category: Ex-smoker (<6 months), BMI: 28. Interpret the result.

```
newData <- data.frame(
              Edad=46,
              Sexo=factor(1, levels=c(1,2)),
              Glucosa=95,
              Colesterol_Total=194,
              Colesterol_HDL=47,
              m2p11a_PAS=128,
              m2p11a_PAD=76,
              m7p3=1,
              ta3=1,
              IMC=28,
              # not relevant, only for reshaping
              RCV_CHILENO_RECODIFICADO=factor(3, levels=c(1,2,3))

)
newDataL <- mlogit.data(newData, choice="RCV_CHILENO_RECODIFICADO", shape="wide")
probs <- data.frame(Risk = c('low', 'moderate', 'high'),
       Probability = c(predict(mdl, newDataL)))
stargazer(probs, summary=FALSE, title='Category probability')
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Nov 30, 2020 - 12:38:30 AM

Table 2: Category probability

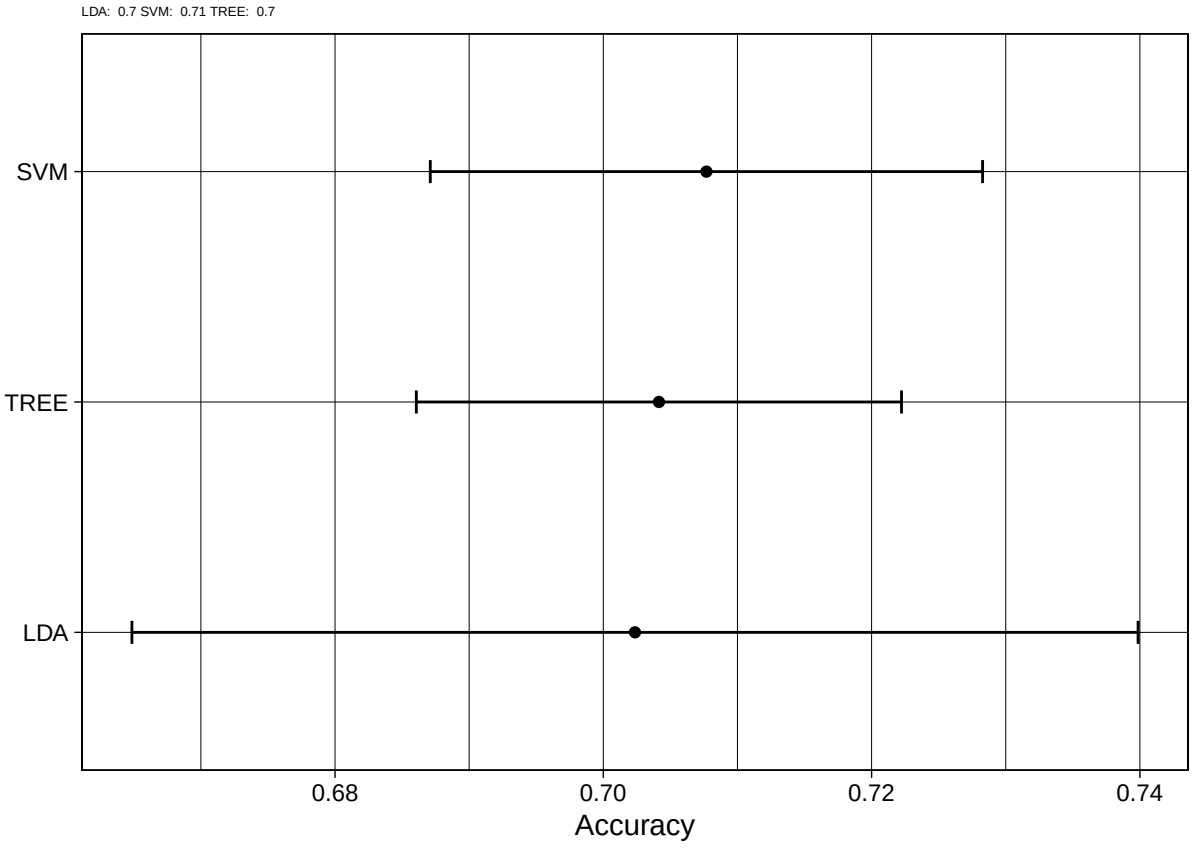
	Risk	Probability
1	low	0.297
2	moderate	0.444
3	high	0.259

Given such values the most likely class is '2', meaning a person with such characteristics would be categorized, by the model, as 'moderate' cardiovascular risk. Which is expected as most positive coefficients were elevated in this simulated person.

2. Build a model based on discriminant analysis, another on support vector machines and another on trees to predict the cardiovascular risk obtained by the classification described above (low risk = 1, moderate risk = 2 and high risk = 3)

a. Use the Matlab Classification Learner App (or similar) to build the three models with all the predictors using 5-fold cross validation and report the accuracy obtained.

```r
# discriminant analysis
head(dataFinal)
# center and scale data
preproc <- c("center", "scale")
# setup cross validation
control <- trainControl(method = "cv", number = 5)
# setup accuracy as metric
performance_metric <- "Accuracy"
LDAmdl <- train(RCV_CHILENO_RECODIFICADO ~ ., data = dataFinal %>% select(-c(IdEncuesta)),
    metric = performance_metric, trControl = control, preProcess = preproc,
    method = "lda")
# support vector machine parameter C was optimized to render maximum
# accuracy
SVMmdl <- train(RCV_CHILENO_RECODIFICADO ~ ., data = dataFinal %>% select(-c(IdEncuesta)),
    metric = performance_metric, trControl = control, preProcess = preproc,
    method = "svmLinear", tuneGrid = expand.grid(C = seq(0.1, 2, length = 50)))
# tree model complexity parameter optimized to render maximum accuracy
TREEmdl <- train(RCV_CHILENO_RECODIFICADO ~ ., data = dataFinal %>% select(-c(IdEncuesta)),
    metric = performance_metric, trControl = control, preProcess = preproc,
    method = "rpart", tuneLength = 10)
results <- resamples(list(LDA = LDAmdl, SVM = SVMmdl, TREE = TREEmdl))
values <- as.data.frame(summary(results)$statistics$Accuracy)$Mean
title <- paste("LDA: ", round(values[1], 2), "SVM: ", round(values[2],
    2), "TREE: ", round(values[3], 2), sep = " ")
ggplot(results) + labs(y = "Accuracy", title = title) + theme_linedraw() +
    theme(plot.title = element_text(size = 5))
```



```
ggsave("accuracy.png", plot = last_plot(), width = 7, height = 7, units = c("cm"))
```

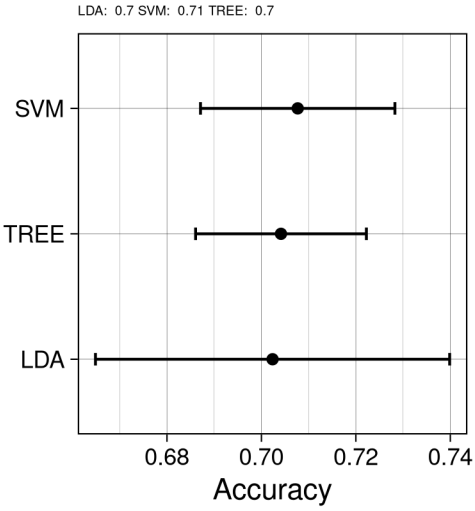


Figure 1: Accuracy values for the three models

b. Obtain the confusion matrix and the percentage of true positives and false negatives for each class (for more than two classes this definition is quite ambiguous so just report the percentage of data classified correctly and the percentage of data classified incorrectly for each class). Compare the results of the three models.

```
treeCM <- confusionMatrix(TREEmdl)$table
svmCM <- confusionMatrix(SVMmdl)$table
ldaCM <- confusionMatrix(LDAmdl)$table
cmList <- list(TREE = treeCM, SVM = svmCM, LDA = ldaCM)

# correctly classified data per class
goodPerc <- 1:3 %>% map(function(x) list(cmList[[x]]["1", "1"], cmList[[x]]["2",
    "2"], cmList[[x]]["3", "3"])) %>% list.stack() %>% rename(Low = V1,
    Moderate = V2, High = V3) %>% mutate(Model = names(cmList)) %>% relocate(Model)

# incorrectly classified data per class
badPerc <- 1:3 %>% map(function(x) list(sum(cmList[[x]]["1", c("2", "3")]),
    sum(cmList[[x]]["2", c("1", "3")]), sum(cmList[[x]]["3", c("1", "2")]))) %>%
    list.stack() %>% rename(Low = V1, Moderate = V2, High = V3) %>% mutate(Model = names(cmList)) %>%
    relocate(Model)

stargazer(goodPerc, summary = FALSE, title = "Correct classification per Model per Class")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Nov 30, 2020 - 12:38:59 AM

```
stargazer(badPerc, summary = FALSE, title = "Incorrect classification per Model per Class")
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Nov 30, 2020 - 12:38:59 AM

Table 3: Correct classification per Model per Class

	Model	Low	Moderate	High
1	TREE	41.804	11.950	16.659
2	SVM	41.271	11.239	18.259
3	LDA	40.782	12.483	16.970

Table 4: Incorrect classification per Model per Class

	Model	Low	Moderate	High
1	TREE	13.816	10.440	5.331
2	SVM	12.039	8.041	9.151
3	LDA	11.906	9.951	7.908