Labor

2023-05-04

Neste trabalho analiso os dados de http://archive.ics.uci.edu/ml/datasets/Labor+Relations?ref=datanews.io e o artigo mencionado para o projeto está em http://ebot.gmu.edu/handle/1920/1622.

O projeto original, conforme a descrição do mesmo, teve como propósito "Software de aprendizagem de conceitos testando um método experimental para aprender descrições de conceitos em duas camadas. Os dados foram utilizados para aprender a descrição de um contrato aceitável e não aceitável. Os contratos não aceitáveis foram obtidos por meio de entrevistas com especialistas ou por meio de criação de exemplos similares, mas que não atendiam aos requisitos do contrato aceitável."

As descrições dos atributos são:

Nome	Tipo	Descrição
duration	numeric	duração do acordo
wage-increase-first-year	numeric	aumento salarial no primeiro ano do contrato
wage-increase-second-year	numeric	aumento salarial no segundo ano do contrato
wage-increase-third-year	numeric	aumento salarial no terceiro ano do contrato
cost-of-living-adjustment	{'none','tcf','tc'}	ajuda de custo de vida
working-hours	numeric	número de horas de trabalho durante a semana
pension	{'none','ret_allw', 'empl_contr'}	contribuições do empregador para o plano de pensão
standby-pay	numeric	pagamento de disponibilidade
shift-differential	numeric	diferencial de turno: suplemento para trabalho no II e III turno
education-allowance	{'yes','no'}	ajuda de custo para educação
statutory-holidays	numeric	número de feriados legais
vacation	{'below_average','average', 'generous'}	número de dias de férias remuneradas
longterm-disability-assistance	{'yes','no'}	ajuda do empregador durante a incapacidade de longo prazo do empregado
contribution-to-dental-plan	{'none','half','full'}	contribuição do empregador para o plano odontológico
bereavement-assistance	{'yes','no'}	contribuição financeira do empregador para cobrir os custos de luto
contribution-to-health-plan	{'none','half','full'}	contribuição do empregador para o plano de saúde
class	{'bad','good'}	

Limpando o ambiente de execução

```
rm(list = ls())
```

Setando o Local de trabalho

```
setwd("C:/Users/karin/OneDrive/Desktop/Mestrado/Mineração")
```

Bibliotecas

```
#install.packages("tidyverse")
#install.packages("ggplot2")
#install.packages("GGally")
#install.packages("ggcorrplot")
#install.packages("DataExplorer")
#install.packages("caret")
#install.packages("rpart.plot")
#install.packages("rpart")
#install.packages("VIM")
#install.packages("rattle")
#install.packages("sampling")
#install.packages("arules")
#install.packages("foreign")
#install.packages("zoo")
#install.packages("Hmisc")
#install.packages("corrplot")
```

Chamada das Bibliotecas

```
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.1     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.0
## v ggplot2   3.4.2     v tibble    3.2.1
## v lubridate 1.9.2     v tidyr     1.3.0
## v purrr     1.0.1
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(ggplot2)
library(GGally)
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
library(ggcorrplot)
library(readr)
library(DataExplorer)
library(doParallel)
```

```
## Carregando pacotes exigidos: foreach
##
## Attaching package: 'foreach'
##
## The following objects are masked from 'package:purrr':
##
##     accumulate, when
##
## Carregando pacotes exigidos: iterators
## Carregando pacotes exigidos: parallel
```

```
library(caret)
```

```
## Carregando pacotes exigidos: lattice
##
## Attaching package: 'caret'
##
## The following object is masked from 'package:purrr':
##
##     lift
```

```
library(rpart)
library(rattle)
```

```
## Carregando pacotes exigidos: bitops
## Rattle: A free graphical interface for data science with R.
## Version 5.5.1 Copyright (c) 2006-2021 Togaware Pty Ltd.
## Type 'rattle()' to shake, rattle, and roll your data.
```

```
library(rpart.plot)
library(RColorBrewer)
library(VIM)
```

```
## Carregando pacotes exigidos: colorspace
## Carregando pacotes exigidos: grid
## VIM is ready to use.
##
## Suggestions and bug-reports can be submitted at: https://github.com/statistikat/VIM/issues
##
## Attaching package: 'VIM'
##
## The following object is masked from 'package:rattle':
##
##     wine
##
## The following object is masked from 'package:datasets':
##
##     sleep
```

```
library(sampling)
```

```
##
## Attaching package: 'sampling'
##
## The following object is masked from 'package:caret':
##
##     cluster
```

```
library(arules)
```

```
## Carregando pacotes exigidos: Matrix
##
## Attaching package: 'Matrix'
##
## The following object is masked from 'package:bitops':
##
##     %&%
##
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
##
##
## Attaching package: 'arules'
##
## The following object is masked from 'package:dplyr':
##
##     recode
##
## The following objects are masked from 'package:base':
##
##     abbreviate, write
```

```
library(foreign)
library(zoo)
```

```
##
## Attaching package: 'zoo'
##
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(Hmisc)
```

```
##
## Attaching package: 'Hmisc'
##
## The following objects are masked from 'package:dplyr':
##
##     src, summarize
##
## The following objects are masked from 'package:base':
```

```
##
##     format.pval, units
```

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

Importação e Visualização Geral do DataFrame

```
options(scipen = 999) #visualização dos dados sem a notação científica

labor <- read.arff("C:/Users/karin/OneDrive/Desktop/Mestrado/Mineração/labor.arff")
summary(labor)
```

```
##     duration     wage-increase-first-year wage-increase-second-year
##  Min.   :1.000   Min.   :2.000            Min.   :2.000
##  1st Qu.:2.000   1st Qu.:2.500            1st Qu.:3.000
##  Median :2.000   Median :4.000            Median :4.000
##  Mean   :2.161   Mean   :3.804            Mean   :3.972
##  3rd Qu.:3.000   3rd Qu.:4.500            3rd Qu.:4.500
##  Max.   :3.000   Max.   :7.000            Max.   :7.000
##  NA's   :1       NA's   :1                NA's   :11
##  wage-increase-third-year cost-of-living-adjustment working-hours
##  Min.   :2.000            none:22                  Min.   :27.00
##  1st Qu.:2.400            tc  : 7                  1st Qu.:37.00
##  Median :4.600            tcf : 8                  Median :38.00
##  Mean   :3.913            NA's:20                  Mean   :38.04
##  3rd Qu.:5.000                                     3rd Qu.:40.00
##  Max.   :5.100                                     Max.   :40.00
##  NA's   :42                                        NA's   :6
##        pension    standby-pay     shift-differential education-allowance
##  empl_contr:12   Min.   : 2.000   Min.   : 0.000     no  :12
##  none      :11   1st Qu.: 2.000   1st Qu.: 3.000     yes :10
##  ret_allw  : 4   Median : 8.000   Median : 4.000     NA's:35
##  NA's      :30   Mean   : 7.444   Mean   : 4.871
##                  3rd Qu.:12.000   3rd Qu.: 5.000
##                  Max.   :14.000   Max.   :25.000
##                  NA's   :48       NA's   :26
##  statutory-holidays          vacation  longterm-disability-assistance
##  Min.   : 9.00      average      :17   no  : 8
##  1st Qu.:10.00      below_average:18   yes :20
##  Median :11.00      generous     :16   NA's:29
##  Mean   :11.09      NA's         : 6
##  3rd Qu.:12.00
##  Max.   :15.00
##  NA's   :4
##  contribution-to-dental-plan bereavement-assistance contribution-to-health-plan
##  full:13                     no  : 3                full:20
##  half:15                     yes :27                half: 9
##  none: 9                     NA's:27                none: 8
##  NA's:20                                            NA's:20
##
```

```
##
##
##   class
##  bad :20
##  good:37
##
##
##
##
##
```

Número de instâncias e atributos

```
#Atributos
ncol(labor)
```

```
## [1] 17
```

```
#Instâncias
nrow(labor)
```

```
## [1] 57
```

Tipos dos atributos: 8 são númericos e 9 categóricos.

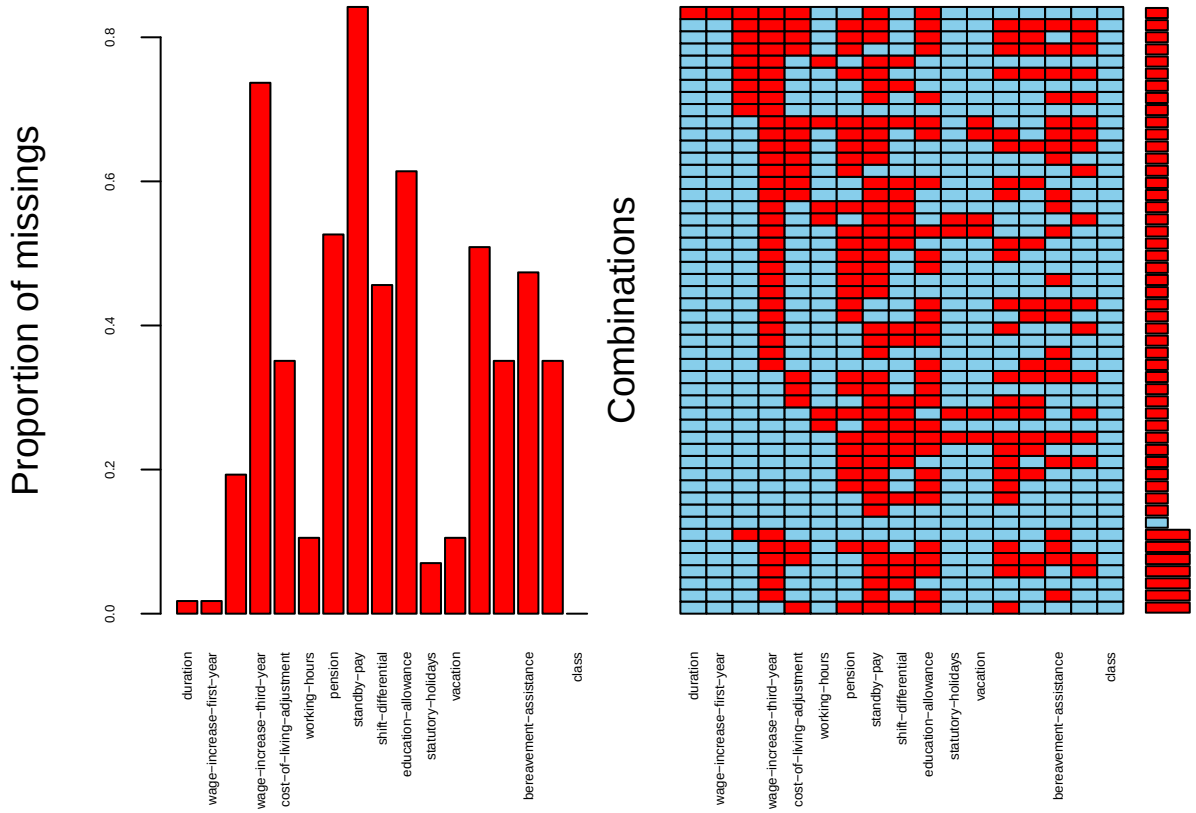
```
str(labor)
```

```
## 'data.frame':    57 obs. of  17 variables:
##  $ duration                      : num  1 2 NA 3 3 2 3 3 2 1 ...
##  $ wage-increase-first-year      : num  5 4.5 NA 3.7 4.5 2 4 6.9 3 5.7 ...
##  $ wage-increase-second-year     : num  NA 5.8 NA 4 4.5 2.5 5 4.8 7 NA ...
##  $ wage-increase-third-year      : num  NA NA NA 5 5 NA 5 2.3 NA NA ...
##  $ cost-of-living-adjustment     : Factor w/ 3 levels "none","tc","tcf": NA NA NA 2 NA NA 2 NA NA 1
##  $ working-hours                 : num  40 35 38 NA 40 35 NA 40 38 40 ...
##  $ pension                       : Factor w/ 3 levels "empl_contr","none",..: NA 3 1 NA NA NA 1 NA NA
##  $ standby-pay                   : num  NA NA NA NA NA NA NA NA 12 NA ...
##  $ shift-differential            : num  2 NA 5 NA NA 6 NA 3 25 4 ...
##  $ education-allowance           : Factor w/ 2 levels "no","yes": NA 2 NA 2 NA 2 NA NA 2 NA ...
##  $ statutory-holidays            : num  11 11 11 NA 12 12 12 12 11 11 ...
##  $ vacation                      : Factor w/ 3 levels "average","below_average",..: 1 2 3 NA 1 1 3 2
##  $ longterm-disability-assistance: Factor w/ 2 levels "no","yes": NA NA 2 NA NA NA 2 NA 2 2 ...
##  $ contribution-to-dental-plan   : Factor w/ 3 levels "full","half",..: NA 1 2 NA 2 NA 3 NA 2 1 ...
##  $ bereavement-assistance        : Factor w/ 2 levels "no","yes": 2 NA 2 2 2 NA 2 NA 2 NA ...
##  $ contribution-to-health-plan   : Factor w/ 3 levels "full","half",..: NA 1 2 NA 2 NA 2 NA NA NA ..
##  $ class                         : Factor w/ 2 levels "bad","good": 2 2 2 2 2 2 2 2 2 2 ...
```

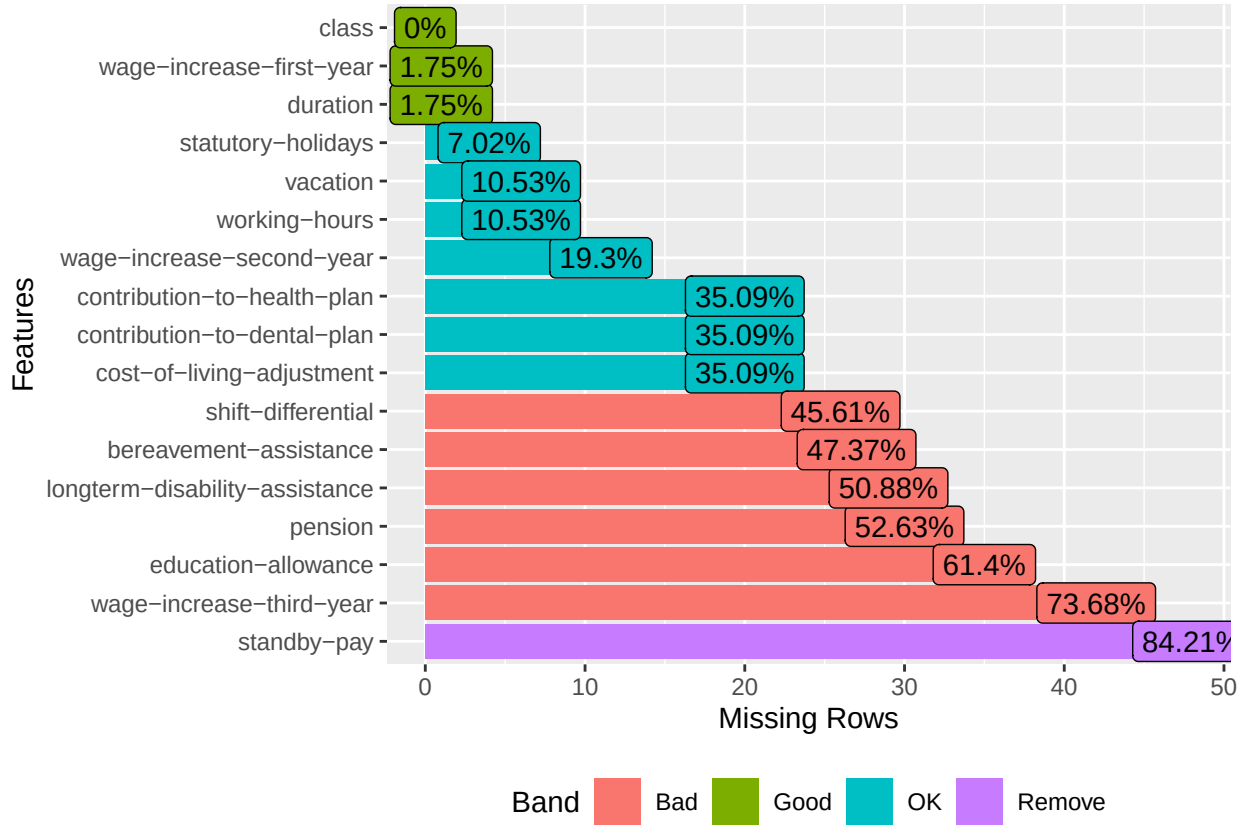
Verificação de dados Missing

```
ppData <- labor
missPlotData <- aggr(ppData, numbers = TRUE, sortvars = TRUE, labels = names(ppData), cex.axis = 0.4, g
```

```
## Warning in plot.aggr(res, ...): not enough vertical space to display
## frequencies (too many combinations)
```




```
plot_missing(labor)
```



Retirada das colunas com missing acima de 30%

Foram retiradas essas colunas tendo em vista que ao fazer a imputação geraria um desbalanceamento das classes comprometendo a confiabilidade do modelo

```
myvars <- names(labor) %in% c("shift-differential","bereavement-assistance","longterm-disability-assista
labor <- labor[!myvars]
#str(labor)
summary(labor)
```

```
##     duration     wage-increase-first-year wage-increase-second-year
##  Min.   :1.000   Min.   :2.000            Min.   :2.000
##  1st Qu.:2.000   1st Qu.:2.500            1st Qu.:3.000
##  Median :2.000   Median :4.000            Median :4.000
##  Mean   :2.161   Mean   :3.804            Mean   :3.972
##  3rd Qu.:3.000   3rd Qu.:4.500            3rd Qu.:4.500
##  Max.   :3.000   Max.   :7.000            Max.   :7.000
##  NA's   :1       NA's   :1                NA's   :11
##  working-hours   statutory-holidays          vacation   class
##  Min.   :27.00   Min.   : 9.00      average      :17   bad :20
##  1st Qu.:37.00   1st Qu.:10.00      below_average:18   good:37
##  Median :38.00   Median :11.00      generous     :16
##  Mean   :38.04   Mean   :11.09      NA's         : 6
##  3rd Qu.:40.00   3rd Qu.:12.00
##  Max.   :40.00   Max.   :15.00
##  NA's   :6       NA's   :4
```

Imputação nos campos Missing

```
labor$duration <- impute(labor$duration, median)
labor$`wage-increase-first-year` <- impute(labor$`wage-increase-first-year`, median)
labor$`wage-increase-second-year`<- impute(labor$`wage-increase-second-year`, median)
labor$`working-hours`<- impute(labor$`working-hours`,median)
labor$`statutory-holidays`<- impute(labor$`statutory-holidays`,median)
labor$vacation <- impute(labor$vacation, mode)

summary(labor)
```

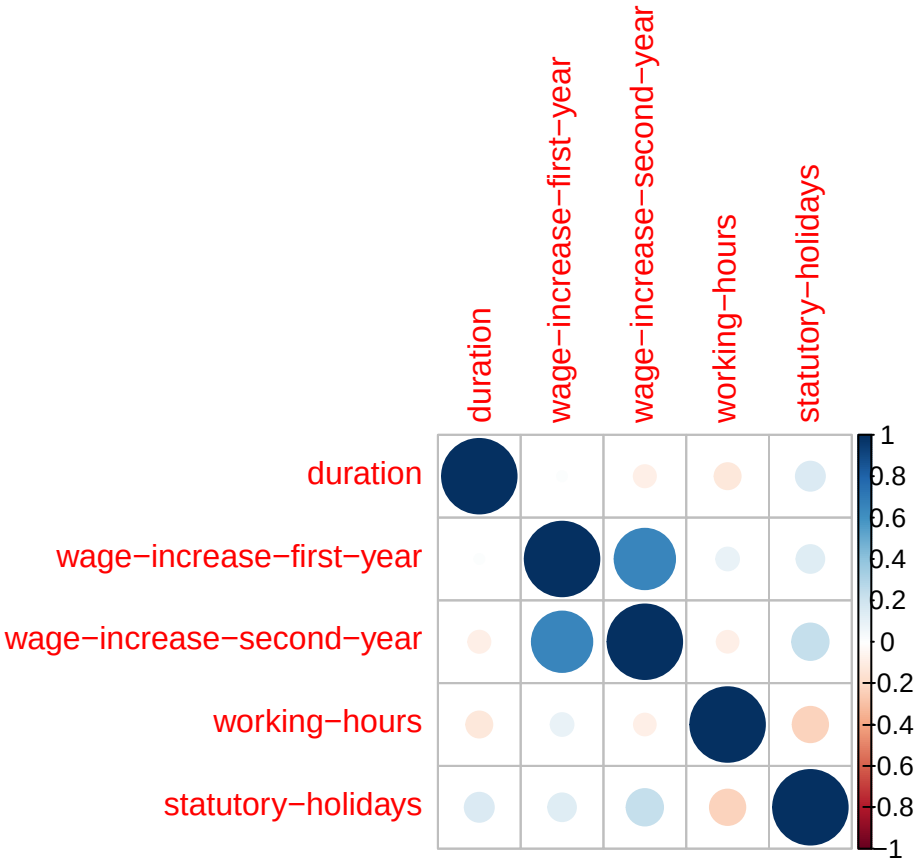
```
##
##  1 values imputed to 2
##
##
##  1 values imputed to 4
##
##
##  11 values imputed to 4
##
##
##  6 values imputed to 38
##
##
##  4 values imputed to 11
##
##
##  6 values imputed to below_average
```

```
##     duration     wage-increase-first-year wage-increase-second-year
##  Min.   :1.000   Min.   :2.000            Min.   :2.000
##  1st Qu.:2.000   1st Qu.:2.500            1st Qu.:3.500
##  Median :2.000   Median :4.000            Median :4.000
##  Mean   :2.158   Mean   :3.807            Mean   :3.977
##  3rd Qu.:3.000   3rd Qu.:4.500            3rd Qu.:4.500
##  Max.   :3.000   Max.   :7.000            Max.   :7.000
##  working-hours   statutory-holidays          vacation   class
##  Min.   :27.00   Min.   : 9.00      average      :17   bad :20
##  1st Qu.:37.00   1st Qu.:10.00      below_average:24   good:37
##  Median :38.00   Median :11.00      generous     :16
##  Mean   :38.04   Mean   :11.09
##  3rd Qu.:40.00   3rd Qu.:12.00
##  Max.   :40.00   Max.   :15.00
```

Correlação Criei uma nova base de dados sem as colunas com dados categóricos para plotar a matriz de correlação

```
myvars <- names(labor) %in% c("vacation","class")
labor_corr <- labor[!myvars]

corrplot(cor(labor_corr), method = "circle")
```



Análise Geral dos Dados

```
ggpairs(labor, columns = 1:7, ggplot2::aes(colour=class))
```

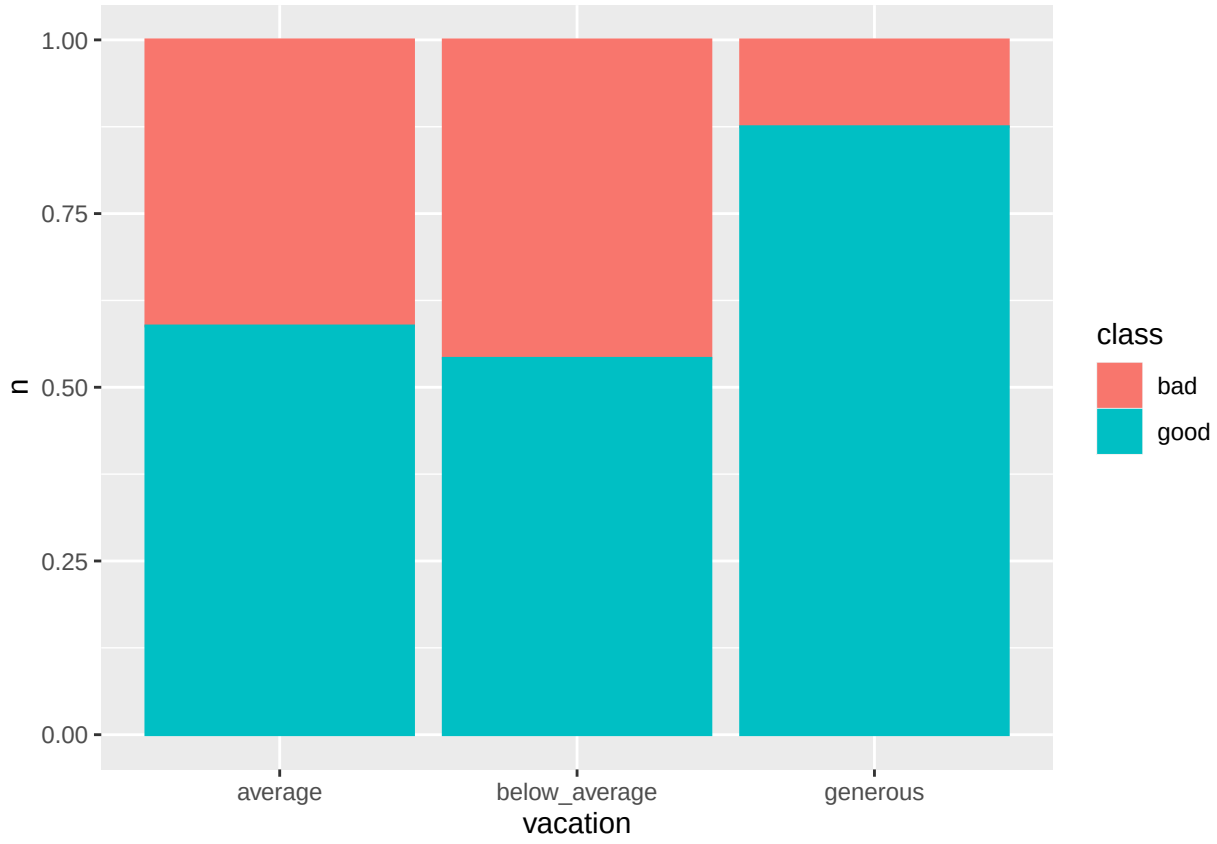
```
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
```

```
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
## Don’t know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.
```

```
## Don't know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Don't know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Don't know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Don't know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Don't know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Don't know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Don't know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## Don't know how to automatically pick scale for object of type <impute>.
## Defaulting to continuous.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

Relação entre o tempo de férias e a classe

```
count_data = labor %>% group_by(vacation, class) %>% count()
ggplot(count_data, aes(x = vacation, y = n, color = class, fill = class)) +
    geom_bar(position = "fill",stat = "identity")
```



Separação entre treino e teste

```
set.seed(123)
partition <- createDataPartition(labor$class, p=0.75, list = FALSE)

train.set <- labor[partition,]
test.set <- labor[-partition,]

test.set
```

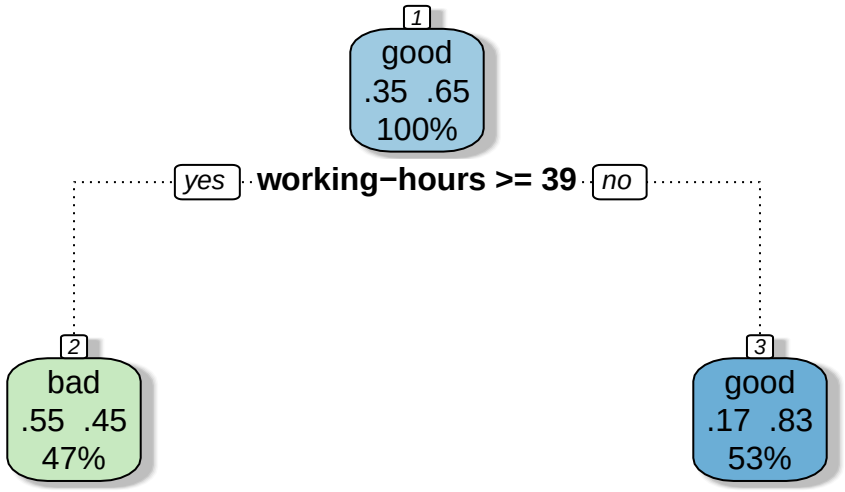
```
##    duration wage-increase-first-year wage-increase-second-year working-hours
## 1         1                      5.0                       4.0            40
## 2         2                      4.5                       5.8            35
## 3         2                      4.0                       4.0            38
## 6         2                      2.0                       2.5            35
## 13        2                      3.5                       4.0            40
## 24        2                      4.5                       4.0            40
## 27        2                      4.5                       4.5            38
## 32        3                      3.0                       2.0            40
## 33        2                      2.5                       2.5            38
## 38        1                      2.8                       4.0            38
```

```
## 42         2                    2.0                    3.0           38
## 52         3                    2.0                    3.0           38
## 53         3                    3.5                    4.0           35
## 57         3                    6.0                    6.0           35
##    statutory-holidays      vacation class
## 1                  11       average  good
## 2                  11 below_average  good
## 3                  11      generous  good
## 6                  12       average  good
## 13                 10 below_average   bad
## 24                 10      generous  good
## 27                 10 below_average  good
## 32                 10 below_average   bad
## 33                 10       average   bad
## 38                  9 below_average   bad
## 42                 12      generous   bad
## 52                 11 below_average  good
## 53                 13      generous  good
## 57                  9      generous  good
```

Modelo e plot da Árvore

```
labor_tree <- rpart(class~., data=train.set, method = "class", control=rpart.control(minsplit=20, minbu

fancyRpartPlot(labor_tree, caption = NULL)
```



Predições

```
predictions <- predict(labor_tree,test.set)

predictions
```

```
##         bad     good
## 1  0.550000 0.450000
## 2  0.173913 0.826087
## 3  0.173913 0.826087
## 6  0.173913 0.826087
## 13 0.550000 0.450000
## 24 0.550000 0.450000
## 27 0.173913 0.826087
## 32 0.550000 0.450000
## 33 0.173913 0.826087
## 38 0.173913 0.826087
## 42 0.173913 0.826087
## 52 0.173913 0.826087
## 53 0.173913 0.826087
## 57 0.173913 0.826087
```

Como a estrutura de predictions está dessa forma

bad	good
0.550000	0.450000

Eu fiz uma função para colocar em uma lista o nome da coluna com o maior valor. E após uso a função factor para que fique igual a test.set$Age o que me permite usar a confusionMatrix

```
maior_coluna <- function(dados) {
  idx <- max.col(dados)
  nomes <- colnames(dados)
  resultado <- lapply(1:nrow(dados), function(i) nomes[idx[i]])
  return(resultado)
}

predicao <- maior_coluna(data.frame(predictions))

predicao_class <- factor(make.names(predicao))

str(predicao_class)
```

```
##  Factor w/ 2 levels "bad","good": 1 2 2 2 1 1 2 1 2 2 ...
```

```
str(test.set$class)
```

```
##  Factor w/ 2 levels "bad","good": 2 2 2 2 1 2 2 1 1 1 ...
```

Matriz de Confusão

```
cm <- confusionMatrix(predicao_class, test.set$class, mode = "everything")
cm
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction bad good
##       bad    2    2
##       good   3    7
##
##                Accuracy : 0.6429
##                  95% CI : (0.3514, 0.8724)
##     No Information Rate : 0.6429
##     P-Value [Acc > NIR] : 0.6188
##
##                   Kappa : 0.186
##
##  Mcnemar's Test P-Value : 1.0000
##
##             Sensitivity : 0.4000
##             Specificity : 0.7778
##          Pos Pred Value : 0.5000
##          Neg Pred Value : 0.7000
##               Precision : 0.5000
##                  Recall : 0.4000
##                      F1 : 0.4444
##              Prevalence : 0.3571
##          Detection Rate : 0.1429
##    Detection Prevalence : 0.2857
##       Balanced Accuracy : 0.5889
##
##        'Positive' Class : bad
##
```

```
str(train.set$class)
```

```
##  Factor w/ 2 levels "bad","good": 2 2 2 2 2 2 2 2 2 2 ...
```

```
str(predicao_class)
```

```
##  Factor w/ 2 levels "bad","good": 1 2 2 2 1 1 2 1 2 2 ...
```