R Notebook

Parametros:

Measure = F1 measure
Columns = sampling, weight_space, underbagging, learner
Performance = holdout_measure
Filter keys = imba.rate
Filter values = 0.05

```r
library("scmamp")
library(dplyr)
```

Tratamento dos dados

Carregando data set compilado

```r
ds = read.csv("/home/rodrigo/Dropbox/UNICAMP/IC/estudo_cost_learning/SummaryResults/summary_compilation

ds = filter(ds, learner != "classif.rusboost")
summary(ds)
```

```
##                  learner      weight_space
##  classif.ksvm        :17100   Mode :logical
##  classif.randomForest:17100   FALSE:41040
##  classif.rusboost    :    0   TRUE :10260
##  classif.xgboost     :17100   NA's :0
##
##
##
##                             measure        sampling     underbagging
##  Accuracy                       :10260   ADASYN:10260   Mode :logical
##  Area under the curve           :10260   FALSE :30780   FALSE:41040
##  F1 measure                     :10260   SMOTE :10260   TRUE :10260
##  G-mean                         :10260                  NA's :0
##  Matthews correlation coefficient:10260
##
##
##  tuning_measure    holdout_measure   holdout_measure_residual
##  Min.   :-0.1277   Min.   :-0.2120   Min.   :-0.4658
##  1st Qu.: 0.6911   1st Qu.: 0.4001   1st Qu.: 0.1994
##  Median : 0.9700   Median : 0.8571   Median : 0.5581
##  Mean   : 0.7903   Mean   : 0.6718   Mean   : 0.5298
##  3rd Qu.: 0.9975   3rd Qu.: 0.9900   3rd Qu.: 0.8755
##  Max.   : 1.0000   Max.   : 1.0000   Max.   : 1.0000
##  NA's   :1077      NA's   :1077      NA's   :1077
##  iteration_count                      dataset        imba.rate
##  Min.   :1       abalone                  :  900   Min.   :0.0010
##  1st Qu.:1       adult                    :  900   1st Qu.:0.0100
##  Median :2       bank                     :  900   Median :0.0300
##  Mean   :2       car                      :  900   Mean   :0.0286
##  3rd Qu.:3       cardiotocography-10clases:  900   3rd Qu.:0.0500
##  Max.   :3       cardiotocography-3clases :  900   Max.   :0.0500
##  NA's   :1077    (Other)                  :45900
```

Filtrando pela metrica

```r
ds = filter(ds, measure == params$measure)
```

Filtrando o data set

```r
if(params$filter_keys != 'NULL' && !is.null(params$filter_keys)){
  dots = paste0(params$filter_keys," == '",params$filter_values,"'")
  ds = filter_(ds, .dots = dots)
}

summary(ds)
```

```
##                  learner     weight_space
##  classif.ksvm        :1230   Mode :logical
##  classif.randomForest:1230   FALSE:2952
##  classif.rusboost    :   0   TRUE :738
##  classif.xgboost     :1230   NA's :0
##
##
##
##                             measure       sampling    underbagging
##  Accuracy                       :   0   ADASYN: 738   Mode :logical
##  Area under the curve           :   0   FALSE :2214   FALSE:2952
##  F1 measure                     :3690   SMOTE : 738   TRUE :738
##  G-mean                         :   0                 NA's :0
##  Matthews correlation coefficient:   0
##
##
##  tuning_measure   holdout_measure  holdout_measure_residual
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.00000
##  1st Qu.:0.3333   1st Qu.:0.1000   1st Qu.:0.07022
##  Median :0.8198   Median :0.5000   Median :0.32530
##  Mean   :0.6671   Mean   :0.4905   Mean   :0.39891
##  3rd Qu.:0.9848   3rd Qu.:0.8333   3rd Qu.:0.73016
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.00000
##  NA's   :51       NA's   :51       NA's   :51
##  iteration_count          dataset       imba.rate
##  Min.   :1       abalone      :  45   Min.   :0.05
##  1st Qu.:1       adult        :  45   1st Qu.:0.05
##  Median :2       annealing    :  45   Median :0.05
##  Mean   :2       arrhythmia   :  45   Mean   :0.05
##  3rd Qu.:3       balance-scale:  45   3rd Qu.:0.05
##  Max.   :3       bank         :  45   Max.   :0.05
##  NA's   :51      (Other)      :3420
```

Computando as médias das iteracoes

```r
ds = group_by(ds, learner , weight_space , measure , sampling , underbagging , dataset , imba.rate)
ds = summarise(ds, tuning_measure = mean(tuning_measure), holdout_measure = mean(holdout_measure),
               holdout_measure_residual = mean(holdout_measure_residual))

ds = as.data.frame(ds)
```

Criando dataframe

```
# Dividindo o ds em n, um para cada técnica
splited_df = ds %>% group_by_at(.vars = params$columns) %>% do(vals = as.data.frame(.)) %>% select(vals

# Juntando cada uma das partes horizontalmente em um data set
df_tec_wide = do.call("cbind", splited_df)

# Renomeando duplicacao de nomes
colnames(df_tec_wide) = make.unique(colnames(df_tec_wide))

# Selecionando apenas as medidas da performance escolhida
df_tec_wide_residual = select(df_tec_wide, matches(paste("^", params$performance, "$|", params$performa

# Renomeando colunas
new_names = NULL
for(i in (1:length(splited_df))){
  id = toString(sapply(splited_df[[i]][1, params$columns], as.character))
  new_names = c(new_names, id)
}
colnames(df_tec_wide_residual) = new_names

# Verificando a dimensao do df
dim(df_tec_wide_residual)
```

```
## [1] 82 15
```

```
# Renomeando a variavel
df = df_tec_wide_residual

head(df)
```

```
##   ADASYN, FALSE, FALSE, classif.ksvm
## 1                         0.11482128
## 2                         0.19240139
## 3                         0.42592593
## 4                         0.00000000
## 5                         1.00000000
## 6                         0.03333333
##   ADASYN, FALSE, FALSE, classif.randomForest
## 1                                  0.1101246
## 2                                         NA
## 3                                  0.7575092
## 4                                  0.2222222
## 5                                  1.0000000
## 6                                  0.3091016
##   ADASYN, FALSE, FALSE, classif.xgboost FALSE, FALSE, FALSE, classif.ksvm
## 1                            0.08612787                         0.1529369
## 2                            0.47906509                         0.2891669
## 3                            0.81663435                         0.3285714
## 4                            0.71111111                         0.0000000
## 5                            1.00000000                         1.0000000
## 6                            0.28633208                         0.2166694
##   FALSE, FALSE, FALSE, classif.randomForest
## 1                                 0.0000000
## 2                                 0.4757640
## 3                                 0.7964052
## 4                                 0.5777778
## 5                                 1.0000000
## 6                                 0.2060694
##   FALSE, FALSE, FALSE, classif.xgboost FALSE, FALSE, TRUE, classif.ksvm
## 1                           0.01960784                        0.1331912
## 2                           0.48854872                        0.2629873
## 3                           0.60916861                        0.3997910
## 4                           0.55555556                        0.1582959
## 5                           1.00000000                        0.9523810
## 6                           0.28650352                        0.1695055
##   FALSE, FALSE, TRUE, classif.randomForest
## 1                                0.1516397
## 2                                       NA
## 3                                0.4494949
## 4                                0.4219577
## 5                                0.9047619
## 6                                0.2964672
##   FALSE, FALSE, TRUE, classif.xgboost FALSE, TRUE, FALSE, classif.ksvm
## 1                           0.1572054                        0.1263895
## 2                           0.3164026                        0.3003710
## 3                           0.4472222                        0.3444444
## 4                           0.4656085                        0.0000000
## 5                           0.7555556                        1.0000000
## 6                           0.2999118                        0.1472620
##   FALSE, TRUE, FALSE, classif.randomForest
## 1                                0.0000000
## 2                                       NA
## 3                                0.7124871
## 4                                0.6555556
## 5                                1.0000000
## 6                                0.2060694
##   FALSE, TRUE, FALSE, classif.xgboost SMOTE, FALSE, FALSE, classif.ksvm
## 1                           0.0000000                         0.1426888
## 2                           0.4985972                         0.2291537
## 3                           0.5185185                         0.3619529
## 4                           0.7222222                         0.0000000
## 5                           1.0000000                         0.7666667
## 6                           0.2487880                         0.0932914
##   SMOTE, FALSE, FALSE, classif.randomForest
## 1                                 0.1244320
## 2                                        NA
## 3                                 0.8328431
## 4                                 0.3888889
## 5                                 1.0000000
## 6                                 0.2516727
##   SMOTE, FALSE, FALSE, classif.xgboost
## 1                            0.1571411
## 2                            0.4810402
## 3                            0.8633987
## 4                            0.7777778
## 5                            1.0000000
## 6                            0.2902838
```

```r
summary(df)
```

```
##  ADASYN, FALSE, FALSE, classif.ksvm
##  Min.   :0.0000
##  1st Qu.:0.0000
##  Median :0.2222
##  Mean   :0.3346
##  3rd Qu.:0.6667
##  Max.   :1.0000
##  NA's   :1
##  ADASYN, FALSE, FALSE, classif.randomForest
##  Min.   :0.0000
##  1st Qu.:0.3723
##  Median :0.6825
##  Mean   :0.5985
##  3rd Qu.:0.8792
##  Max.   :1.0000
##  NA's   :5
##  ADASYN, FALSE, FALSE, classif.xgboost FALSE, FALSE, FALSE, classif.ksvm
##  Min.   :0.0000                        Min.   :0.0000
##  1st Qu.:0.3803                        1st Qu.:0.0000
##  Median :0.7144                        Median :0.2056
##  Mean   :0.6182                        Mean   :0.3031
##  3rd Qu.:0.8746                        3rd Qu.:0.5500
##  Max.   :1.0000                        Max.   :1.0000
##
##  FALSE, FALSE, FALSE, classif.randomForest
##  Min.   :0.0000
##  1st Qu.:0.2222
##  Median :0.6000
##  Mean   :0.5554
##  3rd Qu.:0.8849
##  Max.   :1.0000
##  NA's   :1
##  FALSE, FALSE, FALSE, classif.xgboost FALSE, FALSE, TRUE, classif.ksvm
##  Min.   :0.0000                       Min.   :0.06515
##  1st Qu.:0.2494                       1st Qu.:0.17906
##  Median :0.6647                       Median :0.39119
##  Mean   :0.5784                       Mean   :0.43139
##  3rd Qu.:0.8476                       3rd Qu.:0.62527
##  Max.   :1.0000                       Max.   :0.98030
##
##  FALSE, FALSE, TRUE, classif.randomForest
##  Min.   :0.06886
##  1st Qu.:0.24500
##  Median :0.46174
##  Mean   :0.49703
##  3rd Qu.:0.74651
##  Max.   :1.00000
##  NA's   :3
##  FALSE, FALSE, TRUE, classif.xgboost FALSE, TRUE, FALSE, classif.ksvm
##  Min.   :0.07637                     Min.   :0.0000
##  1st Qu.:0.25815                     1st Qu.:0.0000
##  Median :0.41770                     Median :0.1865
##  Mean   :0.47140                     Mean   :0.2977
##  3rd Qu.:0.67242                     3rd Qu.:0.5303
##  Max.   :1.00000                     Max.   :1.0000
##
##  FALSE, TRUE, FALSE, classif.randomForest
##  Min.   :0.0000
##  1st Qu.:0.2222
##  Median :0.6556
##  Mean   :0.5623
##  3rd Qu.:0.8801
##  Max.   :1.0000
##  NA's   :3
##  FALSE, TRUE, FALSE, classif.xgboost SMOTE, FALSE, FALSE, classif.ksvm
##  Min.   :0.0000                      Min.   :0.00000
##  1st Qu.:0.1944                      1st Qu.:0.01353
##  Median :0.6741                      Median :0.26698
##  Mean   :0.5596                      Mean   :0.33683
##  3rd Qu.:0.8617                      3rd Qu.:0.64701
##  Max.   :1.0000                      Max.   :1.00000
##
##  SMOTE, FALSE, FALSE, classif.randomForest
##  Min.   :0.0000
##  1st Qu.:0.2801
##  Median :0.6974
##  Mean   :0.6011
##  3rd Qu.:0.9047
##  Max.   :1.0000
##  NA's   :4
##  SMOTE, FALSE, FALSE, classif.xgboost
##  Min.   :0.0000
##  1st Qu.:0.3967
##  Median :0.7111
##  Mean   :0.6259
##  3rd Qu.:0.9211
##  Max.   :1.0000
##
```

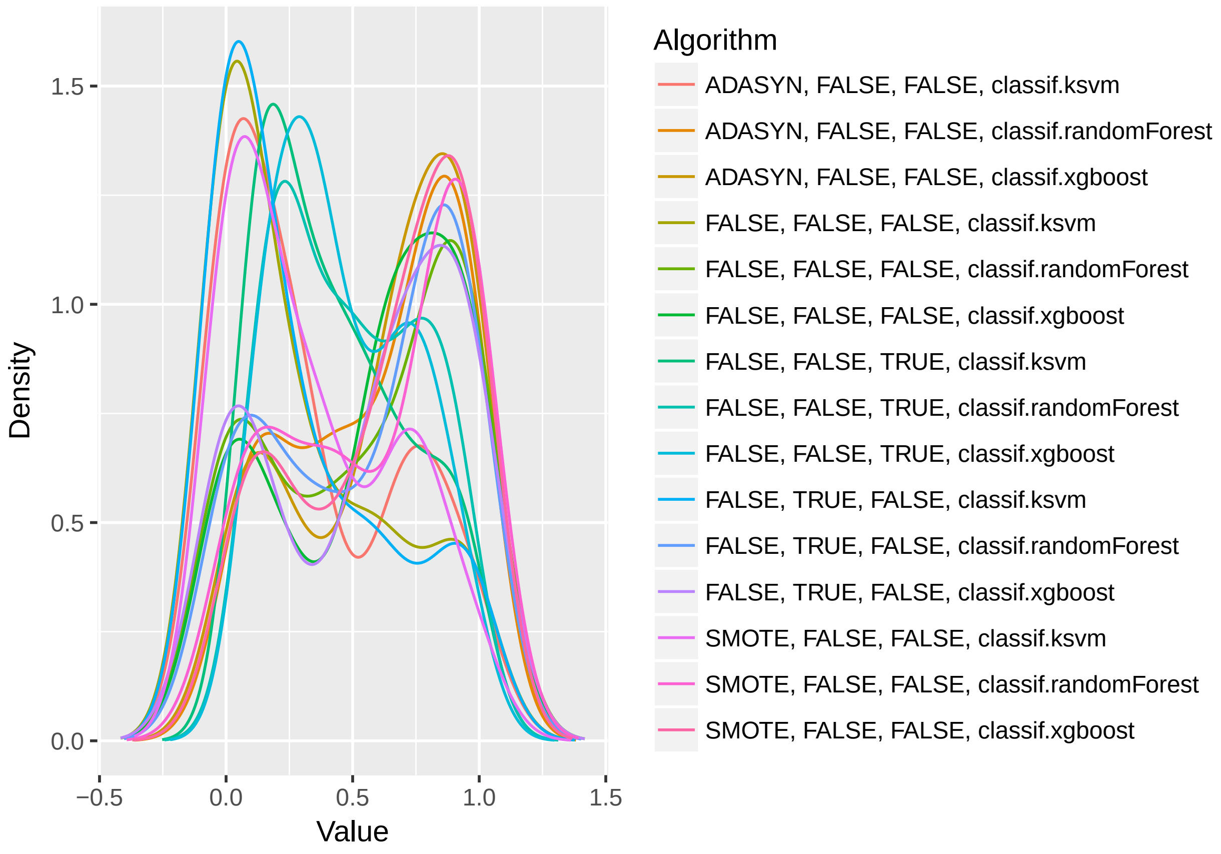
Verificando a média de cada coluna selecionada

```
for(i in (1:dim(df)[2])){
  print(paste("Media da coluna ", colnames(df)[i], " = ", mean(df[,i], na.rm = TRUE), sep=""))
}
```

```
## [1] "Media da coluna ADASYN, FALSE, FALSE, classif.ksvm = 0.334574489595268"
## [1] "Media da coluna ADASYN, FALSE, FALSE, classif.randomForest = 0.598483955405256"
## [1] "Media da coluna ADASYN, FALSE, FALSE, classif.xgboost = 0.618226612863882"
## [1] "Media da coluna FALSE, FALSE, FALSE, classif.ksvm = 0.303114132133715"
## [1] "Media da coluna FALSE, FALSE, FALSE, classif.randomForest = 0.555409954365357"
## [1] "Media da coluna FALSE, FALSE, FALSE, classif.xgboost = 0.578430313726661"
## [1] "Media da coluna FALSE, FALSE, TRUE, classif.ksvm = 0.43138644517747"
## [1] "Media da coluna FALSE, FALSE, TRUE, classif.randomForest = 0.497026798711374"
## [1] "Media da coluna FALSE, FALSE, TRUE, classif.xgboost = 0.471402297071855"
## [1] "Media da coluna FALSE, TRUE, FALSE, classif.ksvm = 0.297729388148578"
## [1] "Media da coluna FALSE, TRUE, FALSE, classif.randomForest = 0.562325380993278"
## [1] "Media da coluna FALSE, TRUE, FALSE, classif.xgboost = 0.559558409430232"
## [1] "Media da coluna SMOTE, FALSE, FALSE, classif.ksvm = 0.336827506230979"
## [1] "Media da coluna SMOTE, FALSE, FALSE, classif.randomForest = 0.601108904055108"
## [1] "Media da coluna SMOTE, FALSE, FALSE, classif.xgboost = 0.625878252923021"
```

Fazendo teste de normalidade

```
plotDensities(data = na.omit(df))
```



Testando as diferencas

```
friedmanTest(df)
```

```
##
##  Friedman's rank sum test
##
## data:  df
## Friedman's chi-squared = 273.03, df = 14, p-value < 2.2e-16
```

Testando as diferencas par a par

```r
test <- nemenyiTest (df, alpha=0.05)
abs(test$diff.matrix) > test$statistic
```

```
##       ADASYN, FALSE, FALSE, classif.ksvm
##  [1,]                              FALSE
##  [2,]                               TRUE
##  [3,]                               TRUE
##  [4,]                              FALSE
##  [5,]                               TRUE
##  [6,]                               TRUE
##  [7,]                              FALSE
##  [8,]                              FALSE
##  [9,]                              FALSE
## [10,]                              FALSE
## [11,]                               TRUE
## [12,]                               TRUE
## [13,]                              FALSE
## [14,]                               TRUE
## [15,]                               TRUE
##       ADASYN, FALSE, FALSE, classif.randomForest
##  [1,]                                       TRUE
##  [2,]                                      FALSE
##  [3,]                                      FALSE
##  [4,]                                       TRUE
##  [5,]                                      FALSE
##  [6,]                                      FALSE
##  [7,]                                       TRUE
##  [8,]                                      FALSE
##  [9,]                                      FALSE
## [10,]                                       TRUE
## [11,]                                      FALSE
## [12,]                                      FALSE
## [13,]                                       TRUE
## [14,]                                      FALSE
## [15,]                                      FALSE
##       ADASYN, FALSE, FALSE, classif.xgboost
##  [1,]                                  TRUE
##  [2,]                                 FALSE
##  [3,]                                 FALSE
##  [4,]                                  TRUE
##  [5,]                                 FALSE
##  [6,]                                 FALSE
##  [7,]                                  TRUE
##  [8,]                                  TRUE
##  [9,]                                  TRUE
## [10,]                                  TRUE
## [11,]                                 FALSE
## [12,]                                 FALSE
## [13,]                                  TRUE
## [14,]                                 FALSE
## [15,]                                 FALSE
##         FALSE, FALSE, FALSE, classif.ksvm
##  [1,]                             FALSE
##  [2,]                              TRUE
##  [3,]                              TRUE
##  [4,]                             FALSE
##  [5,]                              TRUE
##  [6,]                              TRUE
##  [7,]                             FALSE
##  [8,]                              TRUE
##  [9,]                              TRUE
## [10,]                             FALSE
## [11,]                              TRUE
## [12,]                              TRUE
## [13,]                             FALSE
## [14,]                              TRUE
## [15,]                              TRUE
##         FALSE, FALSE, FALSE, classif.randomForest
##  [1,]                                      TRUE
##  [2,]                                     FALSE
##  [3,]                                     FALSE
##  [4,]                                      TRUE
##  [5,]                                     FALSE
##  [6,]                                     FALSE
##  [7,]                                     FALSE
##  [8,]                                     FALSE
##  [9,]                                     FALSE
## [10,]                                      TRUE
## [11,]                                     FALSE
## [12,]                                     FALSE
## [13,]                                      TRUE
## [14,]                                     FALSE
## [15,]                                      TRUE
##         FALSE, FALSE, FALSE, classif.xgboost
##  [1,]                                 TRUE
##  [2,]                                FALSE
##  [3,]                                FALSE
##  [4,]                                 TRUE
##  [5,]                                FALSE
##  [6,]                                FALSE
##  [7,]                                 TRUE
##  [8,]                                FALSE
##  [9,]                                FALSE
## [10,]                                 TRUE
## [11,]                                FALSE
## [12,]                                FALSE
## [13,]                                 TRUE
## [14,]                                FALSE
## [15,]                                FALSE
##         FALSE, FALSE, TRUE, classif.ksvm
##  [1,]                            FALSE
##  [2,]                             TRUE
##  [3,]                             TRUE
##  [4,]                            FALSE
##  [5,]                            FALSE
##  [6,]                             TRUE
##  [7,]                            FALSE
##  [8,]                            FALSE
##  [9,]                            FALSE
## [10,]                            FALSE
## [11,]                            FALSE
## [12,]                            FALSE
## [13,]                            FALSE
## [14,]                             TRUE
## [15,]                             TRUE
##       FALSE, FALSE, TRUE, classif.randomForest
##  [1,]                                    FALSE
##  [2,]                                    FALSE
##  [3,]                                     TRUE
##  [4,]                                     TRUE
##  [5,]                                    FALSE
##  [6,]                                    FALSE
##  [7,]                                    FALSE
##  [8,]                                    FALSE
##  [9,]                                    FALSE
## [10,]                                     TRUE
## [11,]                                    FALSE
## [12,]                                    FALSE
## [13,]                                    FALSE
## [14,]                                     TRUE
## [15,]                                     TRUE
##       FALSE, FALSE, TRUE, classif.xgboost FALSE, TRUE, FALSE, classif.ksvm
##  [1,]                               FALSE                            FALSE
##  [2,]                               FALSE                             TRUE
##  [3,]                                TRUE                             TRUE
##  [4,]                                TRUE                            FALSE
##  [5,]                               FALSE                             TRUE
##  [6,]                               FALSE                             TRUE
##  [7,]                               FALSE                            FALSE
##  [8,]                               FALSE                             TRUE
##  [9,]                               FALSE                             TRUE
## [10,]                                TRUE                            FALSE
## [11,]                               FALSE                             TRUE
## [12,]                               FALSE                             TRUE
## [13,]                               FALSE                            FALSE
## [14,]                                TRUE                             TRUE
## [15,]                                TRUE                             TRUE
##       FALSE, TRUE, FALSE, classif.randomForest
##  [1,]                                     TRUE
##  [2,]                                    FALSE
##  [3,]                                    FALSE
##  [4,]                                     TRUE
##  [5,]                                    FALSE
##  [6,]                                    FALSE
##  [7,]                                    FALSE
##  [8,]                                    FALSE
##  [9,]                                    FALSE
## [10,]                                     TRUE
## [11,]                                    FALSE
## [12,]                                    FALSE
## [13,]                                     TRUE
## [14,]                                    FALSE
## [15,]                                     TRUE
##        FALSE, TRUE, FALSE, classif.xgboost
##  [1,]                                TRUE
##  [2,]                               FALSE
##  [3,]                               FALSE
##  [4,]                                TRUE
##  [5,]                               FALSE
##  [6,]                               FALSE
##  [7,]                               FALSE
##  [8,]                               FALSE
##  [9,]                               FALSE
## [10,]                                TRUE
## [11,]                               FALSE
## [12,]                               FALSE
## [13,]                                TRUE
## [14,]                               FALSE
## [15,]                                TRUE
##        SMOTE, FALSE, FALSE, classif.ksvm
##  [1,]                             FALSE
##  [2,]                              TRUE
##  [3,]                              TRUE
##  [4,]                             FALSE
##  [5,]                              TRUE
##  [6,]                              TRUE
##  [7,]                             FALSE
##  [8,]                             FALSE
##  [9,]                             FALSE
## [10,]                             FALSE
## [11,]                              TRUE
## [12,]                              TRUE
## [13,]                             FALSE
## [14,]                              TRUE
## [15,]                              TRUE
##        SMOTE, FALSE, FALSE, classif.randomForest
##  [1,]                                      TRUE
##  [2,]                                     FALSE
##  [3,]                                     FALSE
##  [4,]                                      TRUE
##  [5,]                                     FALSE
##  [6,]                                     FALSE
##  [7,]                                      TRUE
##  [8,]                                      TRUE
##  [9,]                                      TRUE
## [10,]                                      TRUE
## [11,]                                     FALSE
## [12,]                                     FALSE
## [13,]                                      TRUE
## [14,]                                     FALSE
## [15,]                                     FALSE
##        SMOTE, FALSE, FALSE, classif.xgboost
##  [1,]                                 TRUE
##  [2,]                               FALSE
##  [3,]                               FALSE
##  [4,]                                TRUE
##  [5,]                                TRUE
##  [6,]                               FALSE
##  [7,]                                TRUE
##  [8,]                                TRUE
##  [9,]                                TRUE
## [10,]                                TRUE
## [11,]                                TRUE
## [12,]                                TRUE
## [13,]                                TRUE
## [14,]                               FALSE
## [15,]                               FALSE
```

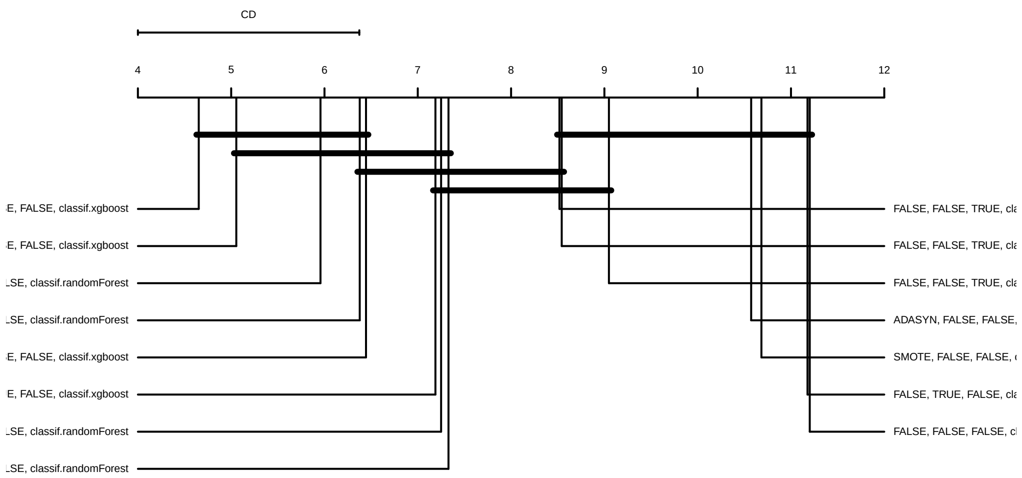
Plotando os ranks

```r
print(colMeans(rankMatrix(df)))
```

```
##         ADASYN, FALSE, FALSE, classif.ksvm 
##                                  10.573171 
## ADASYN, FALSE, FALSE, classif.randomForest 
##                                   6.378049 
##      ADASYN, FALSE, FALSE, classif.xgboost 
##                                   5.054878 
##          FALSE, FALSE, FALSE, classif.ksvm 
##                                  11.201220 
##  FALSE, FALSE, FALSE, classif.randomForest 
##                                   7.250000 
##       FALSE, FALSE, FALSE, classif.xgboost 
##                                   6.445122 
##           FALSE, FALSE, TRUE, classif.ksvm 
##                                   9.048780 
##   FALSE, FALSE, TRUE, classif.randomForest 
##                                   8.518293 
##        FALSE, FALSE, TRUE, classif.xgboost 
##                                   8.542683 
##           FALSE, TRUE, FALSE, classif.ksvm 
##                                  11.176829 
##   FALSE, TRUE, FALSE, classif.randomForest 
##                                   7.329268 
##        FALSE, TRUE, FALSE, classif.xgboost 
##                                   7.189024 
##          SMOTE, FALSE, FALSE, classif.ksvm 
##                                  10.682927 
##  SMOTE, FALSE, FALSE, classif.randomForest 
##                                   5.957317 
##       SMOTE, FALSE, FALSE, classif.xgboost 
##                                   4.652439
```

Plotando grafico de Critical Diference

```
result = tryCatch({
    plotCD(df, alpha=0.05, cex = 0.35)
}, error = function(e) {})
```


CD
E, FALSE, classif.xgboost
E, FALSE, classif.xgboost
.SE, classif.randomForest
.SE, classif.randomForest
E, FALSE, classif.xgboost
E, FALSE, classif.xgboost
.SE, classif.randomForest
.SE, classif.randomForest
FALSE, FALSE, TRUE, cla
FALSE, FALSE, TRUE, cla
FALSE, FALSE, TRUE, cla
ADASYN, FALSE, FALSE,
SMOTE, FALSE, FALSE,
FALSE, TRUE, FALSE, cla
FALSE, FALSE, FALSE, cl