R Notebook

Parametros:

Measure = Area under the curve
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

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

Filtrando o data set

```
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            :3690   FALSE :2214   FALSE:2952
##  F1 measure                      :   0   SMOTE : 738   TRUE :738
##  G-mean                          :   0                 NA's :0
##  Matthews correlation coefficient:   0
##
##
##  tuning_measure   holdout_measure  holdout_measure_residual
##  Min.   :0.3977   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.9145   1st Qu.:0.8175   1st Qu.:0.6976
##  Median :0.9932   Median :0.9755   Median :0.8806
##  Mean   :0.9282   Mean   :0.8846   Mean   :0.8211
##  3rd Qu.:0.9997   3rd Qu.:0.9992   3rd Qu.:0.9784
##  Max.   :1.0000   Max.   :1.0000   Max.   :1.0000
##  NA's   :84       NA's   :84       NA's   :84
##  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   :84      (Other)      :3420
```

Computando as médias das iteracoes

```
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

```r
# 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$performan

# 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
```

```r
# Renomeando a variavel
df = df_tec_wide_residual

head(df)
```

```
##   ADASYN, FALSE, FALSE, classif.ksvm
## 1                          0.6462768
## 2                                 NA
## 3                          0.8346939
## 4                          0.5918367
## 5                          1.0000000
## 6                          0.8085591
##   ADASYN, FALSE, FALSE, classif.randomForest
## 1                                  0.7047856
## 2                                  0.8843849
## 3                                  0.9775510
## 4                                  0.9591837
## 5                                  1.0000000
## 6                                  0.8820066
##   ADASYN, FALSE, FALSE, classif.xgboost FALSE, FALSE, FALSE, classif.ksvm
## 1                             0.6442008                         0.6674464
## 2                             0.9069466                                NA
## 3                             0.9887755                         0.8908163
## 4                             0.9115646                         0.5000000
## 5                             1.0000000                         1.0000000
## 6                             0.8979194                         0.7755400
##   FALSE, FALSE, FALSE, classif.randomForest
## 1                                 0.6383431
## 2                                 0.9069248
## 3                                 0.9937075
## 4                                 0.9370748
## 5                                 1.0000000
## 6                                 0.9163037
##   FALSE, FALSE, FALSE, classif.xgboost FALSE, FALSE, TRUE, classif.ksvm
## 1                            0.6924366                        0.6777485
## 2                            0.9203982                        0.8300474
## 3                            0.9836735                        0.8841837
## 4                            0.9693878                        0.6751701
## 5                            1.0000000                        1.0000000
## 6                            0.8608078                        0.8093895
##   FALSE, FALSE, TRUE, classif.randomForest
## 1                                0.6600585
## 2                                       NA
## 3                                0.9904762
## 4                                0.9421769
## 5                                1.0000000
## 6                                0.8819270
##   FALSE, FALSE, TRUE, classif.xgboost FALSE, TRUE, FALSE, classif.ksvm
## 1                           0.7141033                               NA
## 2                           0.9136941                               NA
## 3                           0.9566327                        0.8908163
## 4                           0.9710884                        0.5000000
## 5                           1.0000000                        1.0000000
## 6                           0.8674663                        0.7755400
##   FALSE, TRUE, FALSE, classif.randomForest
## 1                                0.6383431
## 2                                0.9009547
## 3                                0.9911565
## 4                                0.9863946
## 5                                1.0000000
## 6                                0.9163037
##   FALSE, TRUE, FALSE, classif.xgboost SMOTE, FALSE, FALSE, classif.ksvm
## 1                           0.6997758                         0.6413060
## 2                           0.9219113                                NA
## 3                           0.9870748                         0.7391156
## 4                           0.9761905                         0.5442177
## 5                           1.0000000                         1.0000000
## 6                           0.8578192                         0.8102200
##   SMOTE, FALSE, FALSE, classif.randomForest
## 1                                 0.6820760
## 2                                        NA
## 3                                 0.9942177
## 4                                 0.9285714
## 5                                 1.0000000
## 6                                 0.8969845
##   SMOTE, FALSE, FALSE, classif.xgboost
## 1                            0.6484016
## 2                            0.9075624
## 3                            0.9850340
## 4                            0.9421769
## 5                            1.0000000
## 6                            0.8983471
```

```r
summary(df)
```

```
##  ADASYN, FALSE, FALSE, classif.ksvm
##  Min.   :0.4185
##  1st Qu.:0.7007
##  Median :0.8743
##  Mean   :0.8388
##  3rd Qu.:0.9870
##  Max.   :1.0000
##  NA's   :5
##  ADASYN, FALSE, FALSE, classif.randomForest
##  Min.   :0.3435
##  1st Qu.:0.8832
##  Median :0.9797
##  Mean   :0.9171
##  3rd Qu.:0.9987
##  Max.   :1.0000
##  NA's   :3
##  ADASYN, FALSE, FALSE, classif.xgboost FALSE, FALSE, FALSE, classif.ksvm
##  Min.   :0.4176                        Min.   :0.3991
##  1st Qu.:0.8796                        1st Qu.:0.6745
##  Median :0.9686                        Median :0.8976
##  Mean   :0.8983                        Mean   :0.8260
##  3rd Qu.:0.9969                        3rd Qu.:0.9925
##  Max.   :1.0000                        Max.   :1.0000
##                                        NA's   :4
##  FALSE, FALSE, FALSE, classif.randomForest
##  Min.   :0.4308
##  1st Qu.:0.8920
##  Median :0.9807
##  Mean   :0.9082
##  3rd Qu.:0.9994
##  Max.   :1.0000
##  NA's   :2
##  FALSE, FALSE, FALSE, classif.xgboost FALSE, FALSE, TRUE, classif.ksvm
##  Min.   :0.4917                       Min.   :0.4769
##  1st Qu.:0.8712                       1st Qu.:0.7038
##  Median :0.9758                       Median :0.8730
##  Mean   :0.9169                       Mean   :0.8354
##  3rd Qu.:0.9942                       3rd Qu.:0.9638
##  Max.   :1.0000                       Max.   :1.0000
##
##  FALSE, FALSE, TRUE, classif.randomForest
##  Min.   :0.5018
##  1st Qu.:0.8525
##  Median :0.9751
##  Mean   :0.9054
##  3rd Qu.:0.9935
##  Max.   :1.0000
##  NA's   :3
##  FALSE, FALSE, TRUE, classif.xgboost FALSE, TRUE, FALSE, classif.ksvm
##  Min.   :0.4469                      Min.   :0.3991
##  1st Qu.:0.8532                      1st Qu.:0.6745
##  Median :0.9700                      Median :0.8953
##  Mean   :0.9045                    Mean   :0.8272
##  3rd Qu.:0.9935                    3rd Qu.:0.9925
##  Max.   :1.0000                    Max.   :1.0000
##                                    NA's   :4
##  FALSE, TRUE, FALSE, classif.randomForest
##  Min.   :0.4088
##  1st Qu.:0.8960
##  Median :0.9838
##  Mean   :0.9162
##  3rd Qu.:0.9990
##  Max.   :1.0000
##
##  FALSE, TRUE, FALSE, classif.xgboost SMOTE, FALSE, FALSE, classif.ksvm
##  Min.   :0.4611                    Min.   :0.3773
##  1st Qu.:0.8571                    1st Qu.:0.7174
##  Median :0.9777                    Median :0.8374
##  Mean   :0.9121                    Mean   :0.8315
##  3rd Qu.:0.9963                    3rd Qu.:0.9847
##  Max.   :1.0000                    Max.   :1.0000
##                                    NA's   :3
##  SMOTE, FALSE, FALSE, classif.randomForest
##  Min.   :0.4685
##  1st Qu.:0.8920
##  Median :0.9834
##  Mean   :0.9151
##  3rd Qu.:0.9985
##  Max.   :1.0000
##  NA's   :4
##  SMOTE, FALSE, FALSE, classif.xgboost
##  Min.   :0.5200
##  1st Qu.:0.8806
##  Median :0.9823
##  Mean   :0.9093
##  3rd Qu.:0.9975
##  Max.   :1.0000
##
```

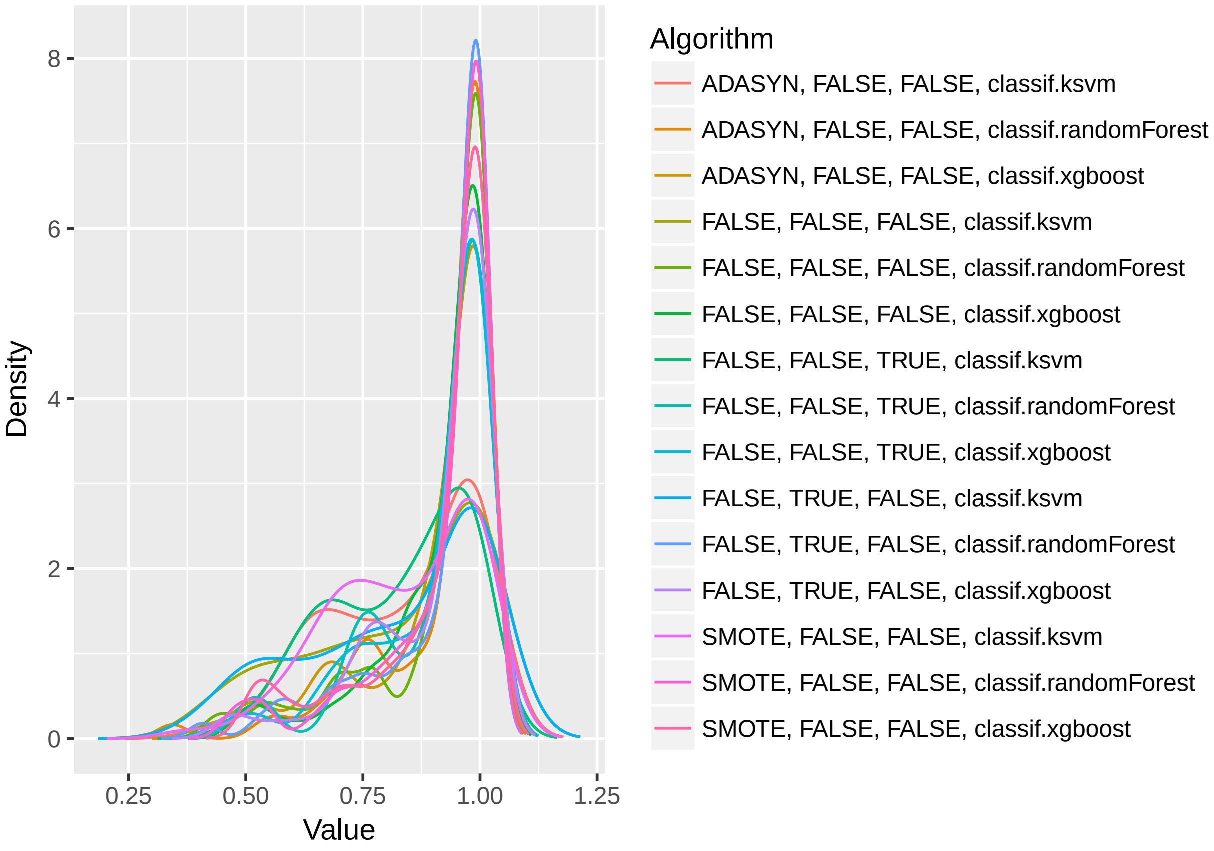
Verificando a média de cada coluna selecionada

```
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 = 0.838811280037709"
## [1] "Media da coluna ADASYN, FALSE, FALSE, classif.randomForest = 0.917130209074314"
## [1] "Media da coluna ADASYN, FALSE, FALSE, classif.xgboost = 0.898292797673017"
## [1] "Media da coluna FALSE, FALSE, FALSE, classif.ksvm = 0.826037266142787"
## [1] "Media da coluna FALSE, FALSE, FALSE, classif.randomForest = 0.908181598903527"
## [1] "Media da coluna FALSE, FALSE, FALSE, classif.xgboost = 0.916943350639857"
## [1] "Media da coluna FALSE, FALSE, TRUE, classif.ksvm = 0.835393191528209"
## [1] "Media da coluna FALSE, FALSE, TRUE, classif.randomForest = 0.90537313758152"
## [1] "Media da coluna FALSE, FALSE, TRUE, classif.xgboost = 0.904526425056321"
## [1] "Media da coluna FALSE, TRUE, FALSE, classif.ksvm = 0.827241108019741"
## [1] "Media da coluna FALSE, TRUE, FALSE, classif.randomForest = 0.91617178552635"
## [1] "Media da coluna FALSE, TRUE, FALSE, classif.xgboost = 0.912119331397177"
## [1] "Media da coluna SMOTE, FALSE, FALSE, classif.ksvm = 0.831536550259147"
## [1] "Media da coluna SMOTE, FALSE, FALSE, classif.randomForest = 0.915132907203918"
## [1] "Media da coluna SMOTE, FALSE, FALSE, classif.xgboost = 0.909290678721929"
```

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 = 243.44, 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,]                                       TRUE
##  [9,]                                       TRUE
## [10,]                                       TRUE
## [11,]                                      FALSE
## [12,]                                      FALSE
## [13,]                                       TRUE
## [14,]                                      FALSE
## [15,]                                      FALSE
##       ADASYN, FALSE, FALSE, classif.xgboost
##  [1,]                                  TRUE
##  [2,]                                 FALSE
##  [3,]                                 FALSE
##  [4,]                                 FALSE
##  [5,]                                 FALSE
##  [6,]                                 FALSE
##  [7,]                                  TRUE
##  [8,]                                 FALSE
##  [9,]                                 FALSE
## [10,]                                  TRUE
## [11,]                                 FALSE
## [12,]                                 FALSE
## [13,]                                  TRUE
## [14,]                                 FALSE
## [15,]                                 FALSE
##       FALSE, FALSE, FALSE, classif.ksvm
##  [1,]                             FALSE
##  [2,]                              TRUE
##  [3,]                             FALSE
##  [4,]                             FALSE
##  [5,]                              TRUE
##  [6,]                              TRUE
##  [7,]                             FALSE
##  [8,]                             FALSE
##  [9,]                             FALSE
## [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,]                                      TRUE
##  [8,]                                     FALSE
##  [9,]                                     FALSE
## [10,]                                      TRUE
## [11,]                                     FALSE
## [12,]                                     FALSE
## [13,]                                      TRUE
## [14,]                                     FALSE
## [15,]                                     FALSE
##       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,]                             TRUE
##  [6,]                             TRUE
##  [7,]                            FALSE
##  [8,]                             TRUE
##  [9,]                             TRUE
## [10,]                            FALSE
## [11,]                             TRUE
## [12,]                             TRUE
## [13,]                            FALSE
## [14,]                             TRUE
## [15,]                             TRUE
##       FALSE, FALSE, TRUE, classif.randomForest
##  [1,]                                    FALSE
##  [2,]                                     TRUE
##  [3,]                                    FALSE
##  [4,]                                    FALSE
##  [5,]                                    FALSE
##  [6,]                                    FALSE
##  [7,]                                     TRUE
##  [8,]                                    FALSE
##  [9,]                                    FALSE
## [10,]                                    FALSE
## [11,]                                     TRUE
## [12,]                                    FALSE
## [13,]                                     TRUE
## [14,]                                    FALSE
## [15,]                                    FALSE
##       FALSE, FALSE, TRUE, classif.xgboost FALSE, TRUE, FALSE, classif.ksvm
##  [1,]                               FALSE                            FALSE
##  [2,]                                TRUE                             TRUE
##  [3,]                               FALSE                             TRUE
##  [4,]                               FALSE                            FALSE
##  [5,]                               FALSE                             TRUE
##  [6,]                               FALSE                             TRUE
##  [7,]                                TRUE                            FALSE
##  [8,]                               FALSE                            FALSE
##  [9,]                               FALSE                            FALSE
## [10,]                               FALSE                            FALSE
## [11,]                                TRUE                             TRUE
## [12,]                               FALSE                             TRUE
## [13,]                                TRUE                            FALSE
## [14,]                               FALSE                             TRUE
## [15,]                               FALSE                             TRUE
##       FALSE, TRUE, 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
##       FALSE, TRUE, 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
##       SMOTE, 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
##       SMOTE, 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
##       SMOTE, 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
```

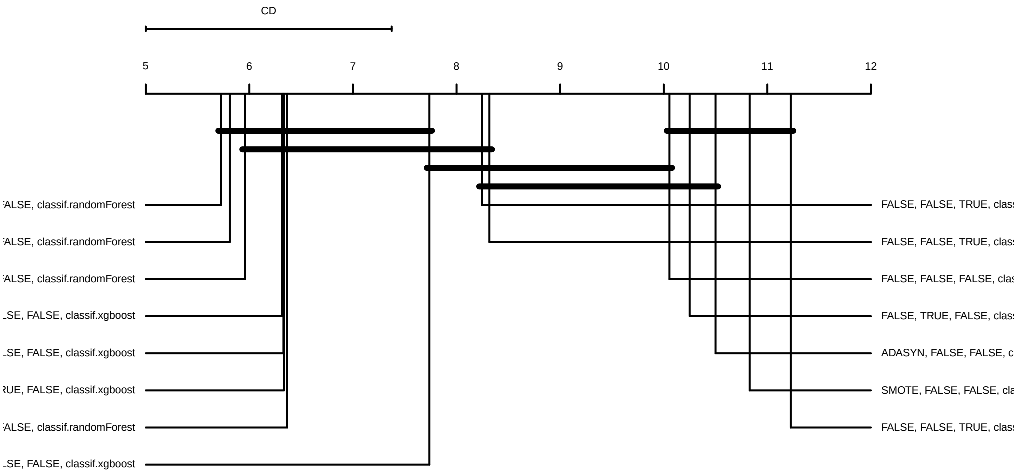
Plotando os ranks

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

```
##         ADASYN, FALSE, FALSE, classif.ksvm
##                                  10.500000
## ADASYN, FALSE, FALSE, classif.randomForest
##                                   5.810976
##      ADASYN, FALSE, FALSE, classif.xgboost
##                                   7.737805
##          FALSE, FALSE, FALSE, classif.ksvm
##                                  10.054878
##  FALSE, FALSE, FALSE, classif.randomForest
##                                   6.365854
##       FALSE, FALSE, FALSE, classif.xgboost
##                                   6.329268
##           FALSE, FALSE, TRUE, classif.ksvm
##                                  11.225610
##   FALSE, FALSE, TRUE, classif.randomForest
##                                   8.317073
##        FALSE, FALSE, TRUE, classif.xgboost
##                                   8.243902
##           FALSE, TRUE, FALSE, classif.ksvm
##                                  10.250000
##   FALSE, TRUE, FALSE, classif.randomForest
##                                   5.725610
##        FALSE, TRUE, FALSE, classif.xgboost
##                                   6.335366
##          SMOTE, FALSE, FALSE, classif.ksvm
##                                  10.829268
##  SMOTE, FALSE, FALSE, classif.randomForest
##                                   5.957317
##       SMOTE, FALSE, FALSE, classif.xgboost
##                                   6.317073
```

Plotando grafico de Critical Diference

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


CD
ALSE, classif.randomForest
ALSE, classif.randomForest
ALSE, classif.randomForest
.SE, FALSE, classif.xgboost
.SE, FALSE, classif.xgboost
RUE, FALSE, classif.xgboost
ALSE, classif.randomForest
.SE, FALSE, classif.xgboost
FALSE, FALSE, TRUE, clas
FALSE, FALSE, TRUE, clas
FALSE, FALSE, FALSE, clas
FALSE, TRUE, FALSE, clas
ADASYN, FALSE, FALSE, c
SMOTE, FALSE, FALSE, cla
FALSE, FALSE, TRUE, clas