

Exercise Method Prediction

Gueller

7/18/2020

Overview

In this assignment, we will use data collected as part of a Human Activity Recognition (HAR) ¹ study and use machine learning to determine how effective the data collected is for prediction of the quality of how the exercise activity is being performed.

The activity is “Unilateral Dumbbell Biceps Curl”. Each participant was asked to perform 10 repetitions, 5 different methods. Essentially, one method was correct, the other four were incorrect in a defined manner. The manner in which the activity was completed is captured in the “classe” field; the definitions of the classe categorizations is shown in the Appendix.²

Executive Summary

Three models (Decision Tree, Random Forest and Generalized Boosted Method) we applied against the cleaned data. Cross-validation found that the Random Forest provided the most accurate prediction for the given data.

Setting the Environment

```
knitr::opts_chunk$set(echo=FALSE)
# Always set the seed
set.seed(1994)
# Libraries
library(caret)
library(rpart)
library(rattle)
library(randomForest)
library(knitr)
library(corrplot)
library(RColorBrewer)
library(kableExtra)
```

Loading, Cleaning and Arranging the Data

```
# locations of the data
trainURL <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv"
testURL <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv"
```

```
# read the data in
training <- read.csv(url(trainURL))
testing <- read.csv(url(testURL))
dim(testing)
```

```
## [1] 20 160
```

We will be working with the training set for now, holding out the testing set for the quiz.

Exploration of the data set using head revealed that the first 7 columns of the data are identifiers that will be removed. Using the is.na function revealed that a number of the fields were almost entirely NA (97.9%), and those will be removed as well, exploration of the data is in the Appendix.

```
# capture the exploration and print output in the Appendix
headUC <- head(training, 2)
naUC <- sapply(training, function(x) sum(is.na(x)))
# Remove Identifiers
training <- training[,-(1:7)]
dim(training)
```

```
## [1] 19622 153
```

```
# Remove NA's
remNA <- sapply(training, function(x) sum(is.na(x))) > 0
training <- training[,remNA==FALSE]
dim(training)
```

```
## [1] 19622 86
```

The near zero variance variables will be identified and removed.

```
nearZero <- nearZeroVar(training)
# nearZero
training <- training[, -nearZero]
dim(training)
```

```
## [1] 19622 53
```

The testing set will be broken into trainSet and testSet.

```
# Break the training set into training and test
inTrain <- createDataPartition(training$classe, p=0.7, list=FALSE)
trainSet <- training[inTrain,]
testSet <- training[-inTrain,]
dim(trainSet); dim(testSet)
```

```
## [1] 13737 53
```

```
## [1] 5885 53
```

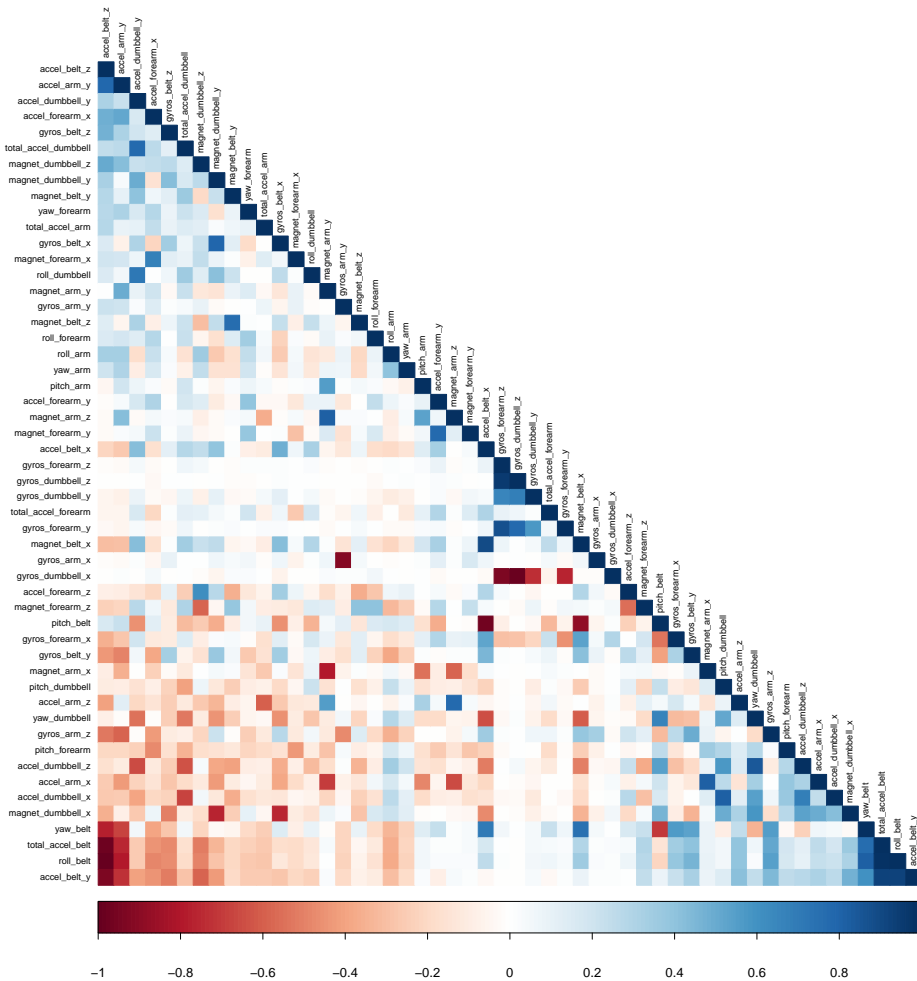
```
# set classe as.factor
trainSet$classe <- as.factor(trainSet$classe)
testSet$classe <- as.factor(testSet$classe)
```

Correlation Analysis

We will look at the number of highly correlated variables to see if a Principal Component Analysis (PCA) should be done.

```
correlationMatrix <- cor(trainSet[, -53])
corrplot(correlationMatrix, order="FPC", method="color", type="lower",
         tl.cex=0.6, tl.col=rgb(0,0,0))
```

fig. 1 - Correlation Table



From the graph, there appears to be a number of \pm highly correlated variables. If PCA is performed, interpretability will be lost and will not be performed.

PCA will be explored if scalability of the select model becomes necessary.

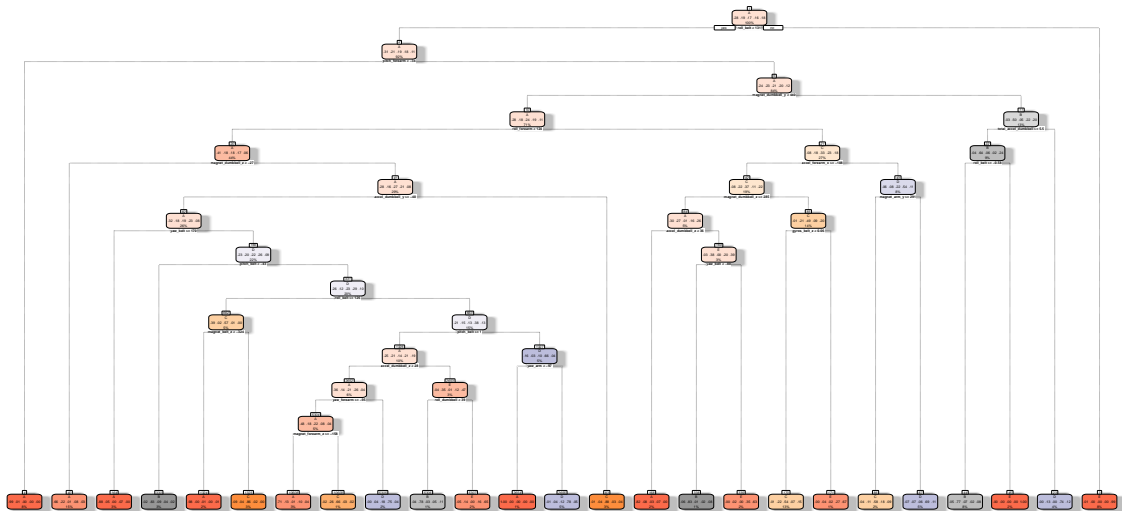
Modeling

Three models will be build with the trainSet data; Decision Tree, Random Forest and Generalized Boosted. The models will be build and cross-validated with the testSet. The model that has the highest accuracy while considering the kappa score will be used against the testing.

Decision Tree

```
mdl_dt <- rpart(classe~., data=trainSet, method="class")
fancyRpartPlot(mdl_dt, sub="", palettes=c("Reds", "Greys"))
```

fig 2. - Decision Tree



```
prd_dt <- predict(mdl_dt, newdata=testSet, type="class")
cm_dt <- confusionMatrix(prd_dt, testSet$classe)
cm_dt
```

Confusion Matrix and Statistics

```
##
##           Reference
## Prediction    A    B    C    D    E
##           A 1563  227   29  105   33
##           B   41  613   52   25   65
##           C   43  208  876  103  135
##           D   20   72   69  652   70
##           E    7   19    0   79  779
##
## Overall Statistics
##
##           Accuracy : 0.7618
##           95% CI : (0.7507, 0.7726)
##           No Information Rate : 0.2845
##           P-Value [Acc > NIR] : < 2.2e-16
##
##           Kappa : 0.6971
##
## McNemar's Test P-Value : < 2.2e-16
##
## Statistics by Class:
##
##           Class: A Class: B Class: C Class: D Class: E
## Sensitivity      0.9337  0.5382  0.8538  0.6763  0.7200
## Specificity      0.9064  0.9614  0.8994  0.9531  0.9781
## Pos Pred Value   0.7987  0.7701  0.6418  0.7384  0.8812
## Neg Pred Value   0.9717  0.8966  0.9668  0.9376  0.9394
## Prevalence       0.2845  0.1935  0.1743  0.1638  0.1839
## Detection Rate   0.2656  0.1042  0.1489  0.1108  0.1324
## Detection Prevalence 0.3325  0.1353  0.2319  0.1500  0.1502
## Balanced Accuracy 0.9201  0.7498  0.8766  0.8147  0.8491
```

```
acc_dt <- round(cm_dt$overall['Accuracy'],4)
kap_dt <- round(cm_dt$overall['Kappa'],4)
```

```
set.seed(1994)
ctrl_rf <- trainControl(method="cv", number=3, verboseIter=FALSE)
mdl_rf <- train(classe~., data=trainSet, method="rf",
               trControl=ctrl_rf)
mdl_rf$finalModel
```

Random Forest

```
##
## Call:
## randomForest(x = x, y = y, mtry = param$mtry)
##           Type of random forest: classification
##           Number of trees: 500
## No. of variables tried at each split: 27
##
```

```
##          OOB estimate of  error rate: 0.68%
## Confusion matrix:
##      A      B      C      D      E class.error
## A 3897      6      2      0      1 0.002304147
## B   18 2635      5      0      0 0.008653123
## C    0   14 2372   10      0 0.010016694
## D    0    0   22 2228    2 0.010657194
## E    0    2    5    7 2511 0.005544554
```

```
prd_rf <- predict(mdl_rf, newdata=testSet)
cm_rf <- confusionMatrix(prd_rf, testSet$classe)
cm_rf
```

```
## Confusion Matrix and Statistics
```

```
##
##          Reference
## Prediction      A      B      C      D      E
##          A 1671    13      0      0      0
##          B     0 1125      8      0      0
##          C     2     1 1015    11      2
##          D     0     0     3   951      1
##          E     1     0     0     2 1079
```

```
## Overall Statistics
```

```
##
##          Accuracy : 0.9925
##          95% CI : (0.99, 0.9946)
##    No Information Rate : 0.2845
##    P-Value [Acc > NIR] : < 2.2e-16
##
##          Kappa : 0.9905
```

```
## McNemar's Test P-Value : NA
```

```
##
## Statistics by Class:
```

```
##
##          Class: A Class: B Class: C Class: D Class: E
## Sensitivity      0.9982  0.9877  0.9893  0.9865  0.9972
## Specificity      0.9969  0.9983  0.9967  0.9992  0.9994
## Pos Pred Value    0.9923  0.9929  0.9845  0.9958  0.9972
## Neg Pred Value    0.9993  0.9971  0.9977  0.9974  0.9994
## Prevalence        0.2845  0.1935  0.1743  0.1638  0.1839
## Detection Rate    0.2839  0.1912  0.1725  0.1616  0.1833
## Detection Prevalence 0.2862  0.1925  0.1752  0.1623  0.1839
## Balanced Accuracy 0.9976  0.9930  0.9930  0.9929  0.9983
```

```
acc_rf <- round(cm_rf$overall['Accuracy'],4)
kap_rf <- round(cm_rf$overall['Kappa'],4)
```

```
set.seed(1994)
ctrl_gbm <- trainControl(method="repeatedcv", number=5, repeats=1)
```

```
mdl_gbm <- train(classe~., data=trainSet, method="gbm",
  trControl=ctrl_gbm, verbose=FALSE)
mdl_gbm$finalModel
```

Generalized Boosted Method

```
## A gradient boosted model with multinomial loss function.
## 150 iterations were performed.
## There were 52 predictors of which 52 had non-zero influence.
```

```
prd_gbm <- predict(mdl_gbm, newdata=testSet)
cm_gbm <- confusionMatrix(prd_gbm, testSet$classe)
cm_gbm
```

```
## Confusion Matrix and Statistics
```

```
##
##           Reference
## Prediction   A    B    C    D    E
##           A 1643   40    0    0    0
##           B   15 1077   44    7    5
##           C    7   20  971   28    5
##           D    8    2    9  921   15
##           E    1    0    2    8 1057
```

```
## Overall Statistics
```

```
##
##           Accuracy : 0.9633
##           95% CI : (0.9582, 0.968)
##           No Information Rate : 0.2845
##           P-Value [Acc > NIR] : < 2.2e-16
##
##           Kappa : 0.9536
##
##           McNemar's Test P-Value : 1.163e-08
```

```
## Statistics by Class:
```

```
##
##           Class: A Class: B Class: C Class: D Class: E
## Sensitivity      0.9815   0.9456   0.9464   0.9554   0.9769
## Specificity      0.9905   0.9850   0.9877   0.9931   0.9977
## Pos Pred Value   0.9762   0.9382   0.9418   0.9644   0.9897
## Neg Pred Value   0.9926   0.9869   0.9887   0.9913   0.9948
## Prevalence       0.2845   0.1935   0.1743   0.1638   0.1839
## Detection Rate   0.2792   0.1830   0.1650   0.1565   0.1796
## Detection Prevalence 0.2860   0.1951   0.1752   0.1623   0.1815
## Balanced Accuracy 0.9860   0.9653   0.9670   0.9742   0.9873
```

```
acc_gbm <- round(cm_gbm$overall['Accuracy'],4)
kap_gbm <- round(cm_gbm$overall['Kappa'],4)
```

Results

```
result_dt <- as.data.frame(rbind(c(acc_dt,kap_dt)))
result_rf <- as.data.frame(rbind(c(acc_rf,kap_rf)))
result_gbm <- as.data.frame(rbind(c(acc_gbm,kap_gbm)))
tab <- rbind(result_dt, result_rf, result_gbm)
row.names(tab) <- c("Decision Tree", "Random Forest", "Generalized Boosted Method")

# Pretty output
kable(tab, "latex", caption="Results") %>%
  kable_styling(full_width=FALSE, position="left",
    latex_options=c("striped","hold_position"), stripe_index=c(2),
    stripe_color = "yellow")
```

Table 1: Results

	Accuracy	Kappa
Decision Tree	0.7618	0.6971
Random Forest	0.9925	0.9905
Generalized Boosted Method	0.9633	0.9536

Based on the Accuracy and Kappa values for each of the models, the Random Forest Model is selected. Both Accuracy and the Kappa values are higher than the others.

Appendix

Overview Information: ¹ <http://groupware.les.inf.puc-rio.br/har> , accessed 20 July 2020.

² “Qualitative Activity Recognition of Weight Lifting Exercises”, Velloso, E.; Bulling, A.; Gellersen, H.; Ugulino, W.; Fuks, H. Qualitative Activity Recognition of Weight Lifting Exercises. Proceedings of 4th International Conference in Cooperation with SIGCHI (Augmented Human '13) . Stuttgart, Germany: ACM SIGCHI, 2013. <http://groupware.les.inf.puc-rio.br/public/papers/2013.Velloso.QAR-WLE.pdf> , Accessed 20 July 2020

Dataset provided by Instructor taken from:

- <http://web.archive.org/web/20161224072740/http://groupware.les.inf.puc-rio.br/har>

Visualization

- The Classe decision tree was plotted using the R Rattle package.

“classe” Definitions

Class A - Exactly according to the specification

Class B - Throwing the elbows to the front

Class C - Lifting the dumbbell only halfway

Class D - Lowering the dumbbell only halfway

Class E - Throwing the hips to the front


```
print("Head Information Before Cleaning:")
```

Before Cleaning Head and is.na's

```
## [1] "Head Information Before Cleaning:"
```

```
headUC
```

```
## X user_name raw_timestamp_part_1 raw_timestamp_part_2 cvtd_timestamp
## 1 1 carlitos 1323084231 788290 05/12/2011 11:23
## 2 2 carlitos 1323084231 808298 05/12/2011 11:23
## new_window num_window roll_belt pitch_belt yaw_belt total_accel_belt
## 1 no 11 1.41 8.07 -94.4 3
## 2 no 11 1.41 8.07 -94.4 3
## kurtosis_roll_belt kurtosis_pitch_belt kurtosis_yaw_belt skewness_roll_belt
## 1
## 2
## skewness_roll_belt.1 skewness_yaw_belt max_roll_belt max_pitch_belt
## 1 NA NA
## 2 NA NA
## max_yaw_belt min_roll_belt min_pitch_belt min_yaw_belt amplitude_roll_belt
## 1 NA NA NA
## 2 NA NA NA
## amplitude_pitch_belt amplitude_yaw_belt var_total_accel_belt avg_roll_belt
## 1 NA NA NA
## 2 NA NA NA
## stddev_roll_belt var_roll_belt avg_pitch_belt stddev_pitch_belt
## 1 NA NA NA NA
## 2 NA NA NA NA
## var_pitch_belt avg_yaw_belt stddev_yaw_belt var_yaw_belt gyros_belt_x
## 1 NA NA NA NA 0.00
## 2 NA NA NA NA 0.02
## gyros_belt_y gyros_belt_z accel_belt_x accel_belt_y accel_belt_z
## 1 0 -0.02 -21 4 22
## 2 0 -0.02 -22 4 22
## magnet_belt_x magnet_belt_y magnet_belt_z roll_arm pitch_arm yaw_arm
## 1 -3 599 -313 -128 22.5 -161
## 2 -7 608 -311 -128 22.5 -161
## total_accel_arm var_accel_arm avg_roll_arm stddev_roll_arm var_roll_arm
## 1 34 NA NA NA NA
## 2 34 NA NA NA NA
## avg_pitch_arm stddev_pitch_arm var_pitch_arm avg_yaw_arm stddev_yaw_arm
## 1 NA NA NA NA NA
## 2 NA NA NA NA NA
## var_yaw_arm gyros_arm_x gyros_arm_y gyros_arm_z accel_arm_x accel_arm_y
## 1 NA 0.00 0.00 -0.02 -288 109
## 2 NA 0.02 -0.02 -0.02 -290 110
## accel_arm_z magnet_arm_x magnet_arm_y magnet_arm_z kurtosis_roll_arm
## 1 -123 -368 337 516
## 2 -125 -369 337 513
## kurtosis_pitch_arm kurtosis_yaw_arm skewness_roll_arm skewness_pitch_arm
```

```

## 1
## 2
## skewness_yaw_arm max_roll_arm max_picth_arm max_yaw_arm min_roll_arm
## 1 NA NA NA NA
## 2 NA NA NA NA
## min_pitch_arm min_yaw_arm amplitude_roll_arm amplitude_pitch_arm
## 1 NA NA NA NA
## 2 NA NA NA NA
## amplitude_yaw_arm roll_dumbbell pitch_dumbbell yaw_dumbbell
## 1 NA 13.05217 -70.49400 -84.87394
## 2 NA 13.13074 -70.63751 -84.71065
## kurtosis_roll_dumbbell kurtosis_picth_dumbbell kurtosis_yaw_dumbbell
## 1
## 2
## skewness_roll_dumbbell skewness_pitch_dumbbell skewness_yaw_dumbbell
## 1
## 2
## max_roll_dumbbell max_picth_dumbbell max_yaw_dumbbell min_roll_dumbbell
## 1 NA NA NA NA
## 2 NA NA NA NA
## min_pitch_dumbbell min_yaw_dumbbell amplitude_roll_dumbbell
## 1 NA NA NA
## 2 NA NA NA
## amplitude_pitch_dumbbell amplitude_yaw_dumbbell total_accel_dumbbell
## 1 NA 37
## 2 NA 37
## var_accel_dumbbell avg_roll_dumbbell stddev_roll_dumbbell var_roll_dumbbell
## 1 NA NA NA NA
## 2 NA NA NA NA
## avg_pitch_dumbbell stddev_pitch_dumbbell var_pitch_dumbbell avg_yaw_dumbbell
## 1 NA NA NA NA
## 2 NA NA NA NA
## stddev_yaw_dumbbell var_yaw_dumbbell gyros_dumbbell_x gyros_dumbbell_y
## 1 NA NA 0 -0.02
## 2 NA NA 0 -0.02
## gyros_dumbbell_z accel_dumbbell_x accel_dumbbell_y accel_dumbbell_z
## 1 0 -234 47 -271
## 2 0 -233 47 -269
## magnet_dumbbell_x magnet_dumbbell_y magnet_dumbbell_z roll_forearm
## 1 -559 293 -65 28.4
## 2 -555 296 -64 28.3
## pitch_forearm yaw_forearm kurtosis_roll_forearm kurtosis_picth_forearm
## 1 -63.9 -153
## 2 -63.9 -153
## kurtosis_yaw_forearm skewness_roll_forearm skewness_pitch_forearm
## 1
## 2
## skewness_yaw_forearm max_roll_forearm max_picth_forearm max_yaw_forearm
## 1 NA NA
## 2 NA NA
## min_roll_forearm min_pitch_forearm min_yaw_forearm amplitude_roll_forearm
## 1 NA NA NA
## 2 NA NA NA
## amplitude_pitch_forearm amplitude_yaw_forearm total_accel_forearm

```

```
## 1          NA          36
## 2          NA          36
##   var_accel_forearm avg_roll_forearm stddev_roll_forearm var_roll_forearm
## 1          NA          NA          NA          NA
## 2          NA          NA          NA          NA
##   avg_pitch_forearm stddev_pitch_forearm var_pitch_forearm avg_yaw_forearm
## 1          NA          NA          NA          NA
## 2          NA          NA          NA          NA
##   stddev_yaw_forearm var_yaw_forearm gyros_forearm_x gyros_forearm_y
## 1          NA          NA          0.03          0
## 2          NA          NA          0.02          0
##   gyros_forearm_z accel_forearm_x accel_forearm_y accel_forearm_z
## 1         -0.02         192         203         -215
## 2         -0.02         192         203         -216
##   magnet_forearm_x magnet_forearm_y magnet_forearm_z classe
## 1          -17         654         476          A
## 2          -18         661         473          A
```

```
print("is.na Information Before Cleaning")
```

```
## [1] "is.na Information Before Cleaning"
```

```
naUC
```

```
##           X           user_name   raw_timestamp_part_1
##           0           0           0
##   raw_timestamp_part_2   cvtd_timestamp   new_window
##           0           0           0
##           num_window           roll_belt           pitch_belt
##           0           0           0
##           yaw_belt           total_accel_belt           kurtosis_roll_belt
##           0           0           0
##   kurtosis_picth_belt   kurtosis_yaw_belt   skewness_roll_belt
##           0           0           0
##   skewness_roll_belt.1   skewness_yaw_belt           max_roll_belt
##           0           0           19216
##           max_picth_belt           max_yaw_belt           min_roll_belt
##           19216           0           19216
##           min_pitch_belt           min_yaw_belt   amplitude_roll_belt
##           19216           0           19216
##   amplitude_pitch_belt   amplitude_yaw_belt   var_total_accel_belt
##           19216           0           19216
##           avg_roll_belt           stddev_roll_belt           var_roll_belt
##           19216           19216           19216
##           avg_pitch_belt           stddev_pitch_belt           var_pitch_belt
##           19216           19216           19216
##           avg_yaw_belt           stddev_yaw_belt           var_yaw_belt
##           19216           19216           19216
##           gyros_belt_x           gyros_belt_y           gyros_belt_z
##           0           0           0
##           accel_belt_x           accel_belt_y           accel_belt_z
##           0           0           0
##           magnet_belt_x           magnet_belt_y           magnet_belt_z
```

##	0	0	0
##	roll_arm	pitch_arm	yaw_arm
##	0	0	0
##	total_accel_arm	var_accel_arm	avg_roll_arm
##	0	19216	19216
##	stddev_roll_arm	var_roll_arm	avg_pitch_arm
##	19216	19216	19216
##	stddev_pitch_arm	var_pitch_arm	avg_yaw_arm
##	19216	19216	19216
##	stddev_yaw_arm	var_yaw_arm	gyros_arm_x
##	19216	19216	0
##	gyros_arm_y	gyros_arm_z	accel_arm_x
##	0	0	0
##	accel_arm_y	accel_arm_z	magnet_arm_x
##	0	0	0
##	magnet_arm_y	magnet_arm_z	kurtosis_roll_arm
##	0	0	0
##	kurtosis_pitch_arm	kurtosis_yaw_arm	skewness_roll_arm
##	0	0	0
##	skewness_pitch_arm	skewness_yaw_arm	max_roll_arm
##	0	0	19216
##	max_pitch_arm	max_yaw_arm	min_roll_arm
##	19216	19216	19216
##	min_pitch_arm	min_yaw_arm	amplitude_roll_arm
##	19216	19216	19216
##	amplitude_pitch_arm	amplitude_yaw_arm	roll_dumbbell
##	19216	19216	0
##	pitch_dumbbell	yaw_dumbbell	kurtosis_roll_dumbbell
##	0	0	0
##	kurtosis_pitch_dumbbell	kurtosis_yaw_dumbbell	skewness_roll_dumbbell
##	0	0	0
##	skewness_pitch_dumbbell	skewness_yaw_dumbbell	max_roll_dumbbell
##	0	0	19216
##	max_pitch_dumbbell	max_yaw_dumbbell	min_roll_dumbbell
##	19216	0	19216
##	min_pitch_dumbbell	min_yaw_dumbbell	amplitude_roll_dumbbell
##	19216	0	19216
##	amplitude_pitch_dumbbell	amplitude_yaw_dumbbell	total_accel_dumbbell
##	19216	0	0
##	var_accel_dumbbell	avg_roll_dumbbell	stddev_roll_dumbbell
##	19216	19216	19216
##	var_roll_dumbbell	avg_pitch_dumbbell	stddev_pitch_dumbbell
##	19216	19216	19216
##	var_pitch_dumbbell	avg_yaw_dumbbell	stddev_yaw_dumbbell
##	19216	19216	19216
##	var_yaw_dumbbell	gyros_dumbbell_x	gyros_dumbbell_y
##	19216	0	0
##	gyros_dumbbell_z	accel_dumbbell_x	accel_dumbbell_y
##	0	0	0
##	accel_dumbbell_z	magnet_dumbbell_x	magnet_dumbbell_y
##	0	0	0
##	magnet_dumbbell_z	roll_forearm	pitch_forearm
##	0	0	0
##	yaw_forearm	kurtosis_roll_forearm	kurtosis_pitch_forearm

##	0	0	0
##	kurtosis_yaw_forearm	skewness_roll_forearm	skewness_pitch_forearm
##	0	0	0
##	skewness_yaw_forearm	max_roll_forearm	max_pitch_forearm
##	0	19216	19216
##	max_yaw_forearm	min_roll_forearm	min_pitch_forearm
##	0	19216	19216
##	min_yaw_forearm	amplitude_roll_forearm	amplitude_pitch_forearm
##	0	19216	19216
##	amplitude_yaw_forearm	total_accel_forearm	var_accel_forearm
##	0	0	19216
##	avg_roll_forearm	stddev_roll_forearm	var_roll_forearm
##	19216	19216	19216
##	avg_pitch_forearm	stddev_pitch_forearm	var_pitch_forearm
##	19216	19216	19216
##	avg_yaw_forearm	stddev_yaw_forearm	var_yaw_forearm
##	19216	19216	19216
##	gyros_forearm_x	gyros_forearm_y	gyros_forearm_z
##	0	0	0
##	accel_forearm_x	accel_forearm_y	accel_forearm_z
##	0	0	0
##	magnet_forearm_x	magnet_forearm_y	magnet_forearm_z
##	0	0	0
##	classe		
##	0		