practical-machine-learning-Project

Pankaj Singh

December 21, 2016

## Final Project Report - Practical Machine Learning Course

These are the files produced during a homework assignment of Coursera's MOOC Practical Machine Learning from Johns Hopkins University. For more information about the several MOOCs comprised in this Specialization, please visit: <https://www.coursera.org/specialization/jhudatascience/>

The scripts have been solely produced, tested and executed on MAC OS X 10.9.4, and RStudio Version 0.98.976.

# Developer: Diogo Aurelio

GitHub repo: <https://github.com/diogoaurelio/Practical-Machine-Learning/tree/master/Course%20Project>

RPubs: <http://rpubs.com/dpaurelio/29426>

Background

Using devices such as Jawbone Up, Nike FuelBand, and Fitbit it is now possible to collect a large amount of data about personal activity relatively inexpensively. These type of devices are part of the quantified self movement - a group of enthusiasts who take measurements about themselves regularly to improve their health, to find patterns in their behavior, or because they are tech geeks. One thing that people regularly do is quantify how much of a particular activity they do, but they rarely quantify how well they do it. In this project, your goal will be to use data from accelerometers on the belt, forearm, arm, and dumbell of 6 participants. They were asked to perform barbell lifts correctly and incorrectly in 5 different ways. More information is available from the website here: <http://groupware.les.inf.puc-rio.br/har> (see the section on the Weight Lifting Exercise Dataset).

# Data

The training data for this project are available here:

<https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv>

The test data are available here:

<https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv>

The data for this project come from this source: <http://groupware.les.inf.puc-rio.br/har>. If you use the document you create for this class for any purpose please cite them as they have been very generous in allowing their data to be used for this kind of assignment.

# What you should submit

The goal of your project is to predict the manner in which they did the exercise. This is the "classe" variable in the training set. You may use any of the other variables to predict with. You should create a report describing how you built your model, how you used cross validation, what you think the expected out of sample error is, and why you made the choices you did. You will also use your prediction model to predict 20 different test cases.

# Peer Review Portion

Your submission for the Peer Review portion should consist of a link to a Github repo with your R markdown and compiled HTML file describing your analysis. Please constrain the text of the writeup to < 2000 words and the number of figures to be less than 5. It will make it easier for the graders if you submit a repo with a gh-pages branch so the HTML page can be viewed online (and you always want to make it easy on graders :-).

# Course Project Prediction Quiz Portion

Apply your machine learning algorithm to the 20 test cases available in the test data above and submit your predictions in appropriate format to the Course Project Prediction Quiz for automated grading.

# Reproducibility

Due to security concerns with the exchange of R code, your code will not be run during the evaluation by your classmates. Please be sure that if they download the repo, they will be able to view the compiled HTML version of your analysis.

library(caret)

## Warning: package 'caret' was built under R version 3.3.2

## Loading required package: lattice

## Warning: package 'lattice' was built under R version 3.3.1

## Loading required package: ggplot2

## Warning: package 'ggplot2' was built under R version 3.3.2

library(rpart)  
library(rpart.plot)

## Warning: package 'rpart.plot' was built under R version 3.3.2

library(RColorBrewer)

## Warning: package 'RColorBrewer' was built under R version 3.3.1

library(rattle)

## Warning: package 'rattle' was built under R version 3.3.2

## Rattle: A free graphical interface for data mining with R.  
## Version 4.1.0 Copyright (c) 2006-2015 Togaware Pty Ltd.  
## Type 'rattle()' to shake, rattle, and roll your data.

library(randomForest)

## Warning: package 'randomForest' was built under R version 3.3.2

## randomForest 4.6-12

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:ggplot2':  
##   
## margin

Finally, load the same seed with the following line of code:

set.seed(12345)  
##The training data set can be found on the following URL:  
  
trainUrl <- "http://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv"  
##The testing data set can be found on the following URL:  
  
testUrl <- "http://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv"  
##Both files are downloaded manually  
  
training <- read.csv("pml-training.csv", na.strings=c("NA","#DIV/0!",""))  
testing <- read.csv("pml-testing.csv", na.strings=c("NA","#DIV/0!",""))  
#check data  
head(training)

## 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  
## 3 3 carlitos 1323084231 820366 05/12/2011 11:23  
## 4 4 carlitos 1323084232 120339 05/12/2011 11:23  
## 5 5 carlitos 1323084232 196328 05/12/2011 11:23  
## 6 6 carlitos 1323084232 304277 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  
## 3 no 11 1.42 8.07 -94.4 3  
## 4 no 12 1.48 8.05 -94.4 3  
## 5 no 12 1.48 8.07 -94.4 3  
## 6 no 12 1.45 8.06 -94.4 3  
## kurtosis\_roll\_belt kurtosis\_picth\_belt kurtosis\_yaw\_belt  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## skewness\_roll\_belt skewness\_roll\_belt.1 skewness\_yaw\_belt max\_roll\_belt  
## 1 NA NA NA NA  
## 2 NA NA NA NA  
## 3 NA NA NA NA  
## 4 NA NA NA NA  
## 5 NA NA NA NA  
## 6 NA NA NA NA  
## max\_picth\_belt max\_yaw\_belt min\_roll\_belt min\_pitch\_belt min\_yaw\_belt  
## 1 NA NA NA NA NA  
## 2 NA NA NA NA NA  
## 3 NA NA NA NA NA  
## 4 NA NA NA NA NA  
## 5 NA NA NA NA NA  
## 6 NA NA NA NA NA  
## amplitude\_roll\_belt amplitude\_pitch\_belt amplitude\_yaw\_belt  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## var\_total\_accel\_belt avg\_roll\_belt stddev\_roll\_belt var\_roll\_belt  
## 1 NA NA NA NA  
## 2 NA NA NA NA  
## 3 NA NA NA NA  
## 4 NA NA NA NA  
## 5 NA NA NA NA  
## 6 NA NA NA NA  
## avg\_pitch\_belt stddev\_pitch\_belt var\_pitch\_belt avg\_yaw\_belt  
## 1 NA NA NA NA  
## 2 NA NA NA NA  
## 3 NA NA NA NA  
## 4 NA NA NA NA  
## 5 NA NA NA NA  
## 6 NA NA NA NA  
## stddev\_yaw\_belt var\_yaw\_belt gyros\_belt\_x gyros\_belt\_y gyros\_belt\_z  
## 1 NA NA 0.00 0.00 -0.02  
## 2 NA NA 0.02 0.00 -0.02  
## 3 NA NA 0.00 0.00 -0.02  
## 4 NA NA 0.02 0.00 -0.03  
## 5 NA NA 0.02 0.02 -0.02  
## 6 NA NA 0.02 0.00 -0.02  
## accel\_belt\_x accel\_belt\_y accel\_belt\_z magnet\_belt\_x magnet\_belt\_y  
## 1 -21 4 22 -3 599  
## 2 -22 4 22 -7 608  
## 3 -20 5 23 -2 600  
## 4 -22 3 21 -6 604  
## 5 -21 2 24 -6 600  
## 6 -21 4 21 0 603  
## magnet\_belt\_z roll\_arm pitch\_arm yaw\_arm total\_accel\_arm var\_accel\_arm  
## 1 -313 -128 22.5 -161 34 NA  
## 2 -311 -128 22.5 -161 34 NA  
## 3 -305 -128 22.5 -161 34 NA  
## 4 -310 -128 22.1 -161 34 NA  
## 5 -302 -128 22.1 -161 34 NA  
## 6 -312 -128 22.0 -161 34 NA  
## avg\_roll\_arm stddev\_roll\_arm var\_roll\_arm avg\_pitch\_arm stddev\_pitch\_arm  
## 1 NA NA NA NA NA  
## 2 NA NA NA NA NA  
## 3 NA NA NA NA NA  
## 4 NA NA NA NA NA  
## 5 NA NA NA NA NA  
## 6 NA NA NA NA NA  
## var\_pitch\_arm avg\_yaw\_arm stddev\_yaw\_arm var\_yaw\_arm gyros\_arm\_x  
## 1 NA NA NA NA 0.00  
## 2 NA NA NA NA 0.02  
## 3 NA NA NA NA 0.02  
## 4 NA NA NA NA 0.02  
## 5 NA NA NA NA 0.00  
## 6 NA NA NA NA 0.02  
## gyros\_arm\_y gyros\_arm\_z accel\_arm\_x accel\_arm\_y accel\_arm\_z magnet\_arm\_x  
## 1 0.00 -0.02 -288 109 -123 -368  
## 2 -0.02 -0.02 -290 110 -125 -369  
## 3 -0.02 -0.02 -289 110 -126 -368  
## 4 -0.03 0.02 -289 111 -123 -372  
## 5 -0.03 0.00 -289 111 -123 -374  
## 6 -0.03 0.00 -289 111 -122 -369  
## magnet\_arm\_y magnet\_arm\_z kurtosis\_roll\_arm kurtosis\_picth\_arm  
## 1 337 516 NA NA  
## 2 337 513 NA NA  
## 3 344 513 NA NA  
## 4 344 512 NA NA  
## 5 337 506 NA NA  
## 6 342 513 NA NA  
## kurtosis\_yaw\_arm skewness\_roll\_arm skewness\_pitch\_arm skewness\_yaw\_arm  
## 1 NA NA NA NA  
## 2 NA NA NA NA  
## 3 NA NA NA NA  
## 4 NA NA NA NA  
## 5 NA NA NA NA  
## 6 NA NA NA NA  
## max\_roll\_arm max\_picth\_arm max\_yaw\_arm min\_roll\_arm min\_pitch\_arm  
## 1 NA NA NA NA NA  
## 2 NA NA NA NA NA  
## 3 NA NA NA NA NA  
## 4 NA NA NA NA NA  
## 5 NA NA NA NA NA  
## 6 NA NA NA NA NA  
## min\_yaw\_arm amplitude\_roll\_arm amplitude\_pitch\_arm amplitude\_yaw\_arm  
## 1 NA NA NA NA  
## 2 NA NA NA NA  
## 3 NA NA NA NA  
## 4 NA NA NA NA  
## 5 NA NA NA NA  
## 6 NA NA NA NA  
## roll\_dumbbell pitch\_dumbbell yaw\_dumbbell kurtosis\_roll\_dumbbell  
## 1 13.05217 -70.49400 -84.87394 NA  
## 2 13.13074 -70.63751 -84.71065 NA  
## 3 12.85075 -70.27812 -85.14078 NA  
## 4 13.43120 -70.39379 -84.87363 NA  
## 5 13.37872 -70.42856 -84.85306 NA  
## 6 13.38246 -70.81759 -84.46500 NA  
## kurtosis\_picth\_dumbbell kurtosis\_yaw\_dumbbell skewness\_roll\_dumbbell  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## skewness\_pitch\_dumbbell skewness\_yaw\_dumbbell max\_roll\_dumbbell  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## max\_picth\_dumbbell max\_yaw\_dumbbell min\_roll\_dumbbell min\_pitch\_dumbbell  
## 1 NA NA NA NA  
## 2 NA NA NA NA  
## 3 NA NA NA NA  
## 4 NA NA NA NA  
## 5 NA NA NA NA  
## 6 NA NA NA NA  
## min\_yaw\_dumbbell amplitude\_roll\_dumbbell amplitude\_pitch\_dumbbell  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## amplitude\_yaw\_dumbbell total\_accel\_dumbbell var\_accel\_dumbbell  
## 1 NA 37 NA  
## 2 NA 37 NA  
## 3 NA 37 NA  
## 4 NA 37 NA  
## 5 NA 37 NA  
## 6 NA 37 NA  
## avg\_roll\_dumbbell stddev\_roll\_dumbbell var\_roll\_dumbbell  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## avg\_pitch\_dumbbell stddev\_pitch\_dumbbell var\_pitch\_dumbbell  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## avg\_yaw\_dumbbell stddev\_yaw\_dumbbell var\_yaw\_dumbbell gyros\_dumbbell\_x  
## 1 NA NA NA 0  
## 2 NA NA NA 0  
## 3 NA NA NA 0  
## 4 NA NA NA 0  
## 5 NA NA NA 0  
## 6 NA NA NA 0  
## gyros\_dumbbell\_y gyros\_dumbbell\_z accel\_dumbbell\_x accel\_dumbbell\_y  
## 1 -0.02 0.00 -234 47  
## 2 -0.02 0.00 -233 47  
## 3 -0.02 0.00 -232 46  
## 4 -0.02 -0.02 -232 48  
## 5 -0.02 0.00 -233 48  
## 6 -0.02 0.00 -234 48  
## accel\_dumbbell\_z magnet\_dumbbell\_x magnet\_dumbbell\_y magnet\_dumbbell\_z  
## 1 -271 -559 293 -65  
## 2 -269 -555 296 -64  
## 3 -270 -561 298 -63  
## 4 -269 -552 303 -60  
## 5 -270 -554 292 -68  
## 6 -269 -558 294 -66  
## roll\_forearm pitch\_forearm yaw\_forearm kurtosis\_roll\_forearm  
## 1 28.4 -63.9 -153 NA  
## 2 28.3 -63.9 -153 NA  
## 3 28.3 -63.9 -152 NA  
## 4 28.1 -63.9 -152 NA  
## 5 28.0 -63.9 -152 NA  
## 6 27.9 -63.9 -152 NA  
## kurtosis\_picth\_forearm kurtosis\_yaw\_forearm skewness\_roll\_forearm  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## skewness\_pitch\_forearm skewness\_yaw\_forearm max\_roll\_forearm  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## max\_picth\_forearm max\_yaw\_forearm min\_roll\_forearm min\_pitch\_forearm  
## 1 NA NA NA NA  
## 2 NA NA NA NA  
## 3 NA NA NA NA  
## 4 NA NA NA NA  
## 5 NA NA NA NA  
## 6 NA NA NA NA  
## min\_yaw\_forearm amplitude\_roll\_forearm amplitude\_pitch\_forearm  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## amplitude\_yaw\_forearm total\_accel\_forearm var\_accel\_forearm  
## 1 NA 36 NA  
## 2 NA 36 NA  
## 3 NA 36 NA  
## 4 NA 36 NA  
## 5 NA 36 NA  
## 6 NA 36 NA  
## avg\_roll\_forearm stddev\_roll\_forearm var\_roll\_forearm avg\_pitch\_forearm  
## 1 NA NA NA NA  
## 2 NA NA NA NA  
## 3 NA NA NA NA  
## 4 NA NA NA NA  
## 5 NA NA NA NA  
## 6 NA NA NA NA  
## stddev\_pitch\_forearm var\_pitch\_forearm avg\_yaw\_forearm  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## stddev\_yaw\_forearm var\_yaw\_forearm gyros\_forearm\_x gyros\_forearm\_y  
## 1 NA NA 0.03 0.00  
## 2 NA NA 0.02 0.00  
## 3 NA NA 0.03 -0.02  
## 4 NA NA 0.02 -0.02  
## 5 NA NA 0.02 0.00  
## 6 NA NA 0.02 -0.02  
## gyros\_forearm\_z accel\_forearm\_x accel\_forearm\_y accel\_forearm\_z  
## 1 -0.02 192 203 -215  
## 2 -0.02 192 203 -216  
## 3 0.00 196 204 -213  
## 4 0.00 189 206 -214  
## 5 -0.02 189 206 -214  
## 6 -0.03 193 203 -215  
## magnet\_forearm\_x magnet\_forearm\_y magnet\_forearm\_z classe  
## 1 -17 654 476 A  
## 2 -18 661 473 A  
## 3 -18 658 469 A  
## 4 -16 658 469 A  
## 5 -17 655 473 A  
## 6 -9 660 478 A

head(testing)

## X user\_name raw\_timestamp\_part\_1 raw\_timestamp\_part\_2 cvtd\_timestamp  
## 1 1 pedro 1323095002 868349 05/12/2011 14:23  
## 2 2 jeremy 1322673067 778725 30/11/2011 17:11  
## 3 3 jeremy 1322673075 342967 30/11/2011 17:11  
## 4 4 adelmo 1322832789 560311 02/12/2011 13:33  
## 5 5 eurico 1322489635 814776 28/11/2011 14:13  
## 6 6 jeremy 1322673149 510661 30/11/2011 17:12  
## new\_window num\_window roll\_belt pitch\_belt yaw\_belt total\_accel\_belt  
## 1 no 74 123.00 27.00 -4.75 20  
## 2 no 431 1.02 4.87 -88.90 4  
## 3 no 439 0.87 1.82 -88.50 5  
## 4 no 194 125.00 -41.60 162.00 17  
## 5 no 235 1.35 3.33 -88.60 3  
## 6 no 504 -5.92 1.59 -87.70 4  
## kurtosis\_roll\_belt kurtosis\_picth\_belt kurtosis\_yaw\_belt  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## skewness\_roll\_belt skewness\_roll\_belt.1 skewness\_yaw\_belt max\_roll\_belt  
## 1 NA NA NA NA  
## 2 NA NA NA NA  
## 3 NA NA NA NA  
## 4 NA NA NA NA  
## 5 NA NA NA NA  
## 6 NA NA NA NA  
## max\_picth\_belt max\_yaw\_belt min\_roll\_belt min\_pitch\_belt min\_yaw\_belt  
## 1 NA NA NA NA NA  
## 2 NA NA NA NA NA  
## 3 NA NA NA NA NA  
## 4 NA NA NA NA NA  
## 5 NA NA NA NA NA  
## 6 NA NA NA NA NA  
## amplitude\_roll\_belt amplitude\_pitch\_belt amplitude\_yaw\_belt  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## var\_total\_accel\_belt avg\_roll\_belt stddev\_roll\_belt var\_roll\_belt  
## 1 NA NA NA NA  
## 2 NA NA NA NA  
## 3 NA NA NA NA  
## 4 NA NA NA NA  
## 5 NA NA NA NA  
## 6 NA NA NA NA  
## avg\_pitch\_belt stddev\_pitch\_belt var\_pitch\_belt avg\_yaw\_belt  
## 1 NA NA NA NA  
## 2 NA NA NA NA  
## 3 NA NA NA NA  
## 4 NA NA NA NA  
## 5 NA NA NA NA  
## 6 NA NA NA NA  
## stddev\_yaw\_belt var\_yaw\_belt gyros\_belt\_x gyros\_belt\_y gyros\_belt\_z  
## 1 NA NA -0.50 -0.02 -0.46  
## 2 NA NA -0.06 -0.02 -0.07  
## 3 NA NA 0.05 0.02 0.03  
## 4 NA NA 0.11 0.11 -0.16  
## 5 NA NA 0.03 0.02 0.00  
## 6 NA NA 0.10 0.05 -0.13  
## accel\_belt\_x accel\_belt\_y accel\_belt\_z magnet\_belt\_x magnet\_belt\_y  
## 1 -38 69 -179 -13 581  
## 2 -13 11 39 43 636  
## 3 1 -1 49 29 631  
## 4 46 45 -156 169 608  
## 5 -8 4 27 33 566  
## 6 -11 -16 38 31 638  
## magnet\_belt\_z roll\_arm pitch\_arm yaw\_arm total\_accel\_arm var\_accel\_arm  
## 1 -382 40.7 -27.80 178 10 NA  
## 2 -309 0.0 0.00 0 38 NA  
## 3 -312 0.0 0.00 0 44 NA  
## 4 -304 -109.0 55.00 -142 25 NA  
## 5 -418 76.1 2.76 102 29 NA  
## 6 -291 0.0 0.00 0 14 NA  
## avg\_roll\_arm stddev\_roll\_arm var\_roll\_arm avg\_pitch\_arm stddev\_pitch\_arm  
## 1 NA NA NA NA NA  
## 2 NA NA NA NA NA  
## 3 NA NA NA NA NA  
## 4 NA NA NA NA NA  
## 5 NA NA NA NA NA  
## 6 NA NA NA NA NA  
## var\_pitch\_arm avg\_yaw\_arm stddev\_yaw\_arm var\_yaw\_arm gyros\_arm\_x  
## 1 NA NA NA NA -1.65  
## 2 NA NA NA NA -1.17  
## 3 NA NA NA NA 2.10  
## 4 NA NA NA NA 0.22  
## 5 NA NA NA NA -1.96  
## 6 NA NA NA NA 0.02  
## gyros\_arm\_y gyros\_arm\_z accel\_arm\_x accel\_arm\_y accel\_arm\_z magnet\_arm\_x  
## 1 0.48 -0.18 16 38 93 -326  
## 2 0.85 -0.43 -290 215 -90 -325  
## 3 -1.36 1.13 -341 245 -87 -264  
## 4 -0.51 0.92 -238 -57 6 -173  
## 5 0.79 -0.54 -197 200 -30 -170  
## 6 0.05 -0.07 -26 130 -19 396  
## magnet\_arm\_y magnet\_arm\_z kurtosis\_roll\_arm kurtosis\_picth\_arm  
## 1 385 481 NA NA  
## 2 447 434 NA NA  
## 3 474 413 NA NA  
## 4 257 633 NA NA  
## 5 275 617 NA NA  
## 6 176 516 NA NA  
## kurtosis\_yaw\_arm skewness\_roll\_arm skewness\_pitch\_arm skewness\_yaw\_arm  
## 1 NA NA NA NA  
## 2 NA NA NA NA  
## 3 NA NA NA NA  
## 4 NA NA NA NA  
## 5 NA NA NA NA  
## 6 NA NA NA NA  
## max\_roll\_arm max\_picth\_arm max\_yaw\_arm min\_roll\_arm min\_pitch\_arm  
## 1 NA NA NA NA NA  
## 2 NA NA NA NA NA  
## 3 NA NA NA NA NA  
## 4 NA NA NA NA NA  
## 5 NA NA NA NA NA  
## 6 NA NA NA NA NA  
## min\_yaw\_arm amplitude\_roll\_arm amplitude\_pitch\_arm amplitude\_yaw\_arm  
## 1 NA NA NA NA  
## 2 NA NA NA NA  
## 3 NA NA NA NA  
## 4 NA NA NA NA  
## 5 NA NA NA NA  
## 6 NA NA NA NA  
## roll\_dumbbell pitch\_dumbbell yaw\_dumbbell kurtosis\_roll\_dumbbell  
## 1 -17.73748 24.96085 126.23596 NA  
## 2 54.47761 -53.69758 -75.51480 NA  
## 3 57.07031 -51.37303 -75.20287 NA  
## 4 43.10927 -30.04885 -103.32003 NA  
## 5 -101.38396 -53.43952 -14.19542 NA  
## 6 62.18750 -50.55595 -71.12063 NA  
## kurtosis\_picth\_dumbbell kurtosis\_yaw\_dumbbell skewness\_roll\_dumbbell  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## skewness\_pitch\_dumbbell skewness\_yaw\_dumbbell max\_roll\_dumbbell  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## max\_picth\_dumbbell max\_yaw\_dumbbell min\_roll\_dumbbell min\_pitch\_dumbbell  
## 1 NA NA NA NA  
## 2 NA NA NA NA  
## 3 NA NA NA NA  
## 4 NA NA NA NA  
## 5 NA NA NA NA  
## 6 NA NA NA NA  
## min\_yaw\_dumbbell amplitude\_roll\_dumbbell amplitude\_pitch\_dumbbell  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## amplitude\_yaw\_dumbbell total\_accel\_dumbbell var\_accel\_dumbbell  
## 1 NA 9 NA  
## 2 NA 31 NA  
## 3 NA 29 NA  
## 4 NA 18 NA  
## 5 NA 4 NA  
## 6 NA 29 NA  
## avg\_roll\_dumbbell stddev\_roll\_dumbbell var\_roll\_dumbbell  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## avg\_pitch\_dumbbell stddev\_pitch\_dumbbell var\_pitch\_dumbbell  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## avg\_yaw\_dumbbell stddev\_yaw\_dumbbell var\_yaw\_dumbbell gyros\_dumbbell\_x  
## 1 NA NA NA 0.64  
## 2 NA NA NA 0.34  
## 3 NA NA NA 0.39  
## 4 NA NA NA 0.10  
## 5 NA NA NA 0.29  
## 6 NA NA NA -0.59  
## gyros\_dumbbell\_y gyros\_dumbbell\_z accel\_dumbbell\_x accel\_dumbbell\_y  
## 1 0.06 -0.61 21 -15  
## 2 0.05 -0.71 -153 155  
## 3 0.14 -0.34 -141 155  
## 4 -0.02 0.05 -51 72  
## 5 -0.47 -0.46 -18 -30  
## 6 0.80 1.10 -138 166  
## accel\_dumbbell\_z magnet\_dumbbell\_x magnet\_dumbbell\_y magnet\_dumbbell\_z  
## 1 81 523 -528 -56  
## 2 -205 -502 388 -36  
## 3 -196 -506 349 41  
## 4 -148 -576 238 53  
## 5 -5 -424 252 312  
## 6 -186 -543 262 96  
## roll\_forearm pitch\_forearm yaw\_forearm kurtosis\_roll\_forearm  
## 1 141 49.30 156.0 NA  
## 2 109 -17.60 106.0 NA  
## 3 131 -32.60 93.0 NA  
## 4 0 0.00 0.0 NA  
## 5 -176 -2.16 -47.9 NA  
## 6 150 1.46 89.7 NA  
## kurtosis\_picth\_forearm kurtosis\_yaw\_forearm skewness\_roll\_forearm  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## skewness\_pitch\_forearm skewness\_yaw\_forearm max\_roll\_forearm  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## max\_picth\_forearm max\_yaw\_forearm min\_roll\_forearm min\_pitch\_forearm  
## 1 NA NA NA NA  
## 2 NA NA NA NA  
## 3 NA NA NA NA  
## 4 NA NA NA NA  
## 5 NA NA NA NA  
## 6 NA NA NA NA  
## min\_yaw\_forearm amplitude\_roll\_forearm amplitude\_pitch\_forearm  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## amplitude\_yaw\_forearm total\_accel\_forearm var\_accel\_forearm  
## 1 NA 33 NA  
## 2 NA 39 NA  
## 3 NA 34 NA  
## 4 NA 43 NA  
## 5 NA 24 NA  
## 6 NA 43 NA  
## avg\_roll\_forearm stddev\_roll\_forearm var\_roll\_forearm avg\_pitch\_forearm  
## 1 NA NA NA NA  
## 2 NA NA NA NA  
## 3 NA NA NA NA  
## 4 NA NA NA NA  
## 5 NA NA NA NA  
## 6 NA NA NA NA  
## stddev\_pitch\_forearm var\_pitch\_forearm avg\_yaw\_forearm  
## 1 NA NA NA  
## 2 NA NA NA  
## 3 NA NA NA  
## 4 NA NA NA  
## 5 NA NA NA  
## 6 NA NA NA  
## stddev\_yaw\_forearm var\_yaw\_forearm gyros\_forearm\_x gyros\_forearm\_y  
## 1 NA NA 0.74 -3.34  
## 2 NA NA 1.12 -2.78  
## 3 NA NA 0.18 -0.79  
## 4 NA NA 1.38 0.69  
## 5 NA NA -0.75 3.10  
## 6 NA NA -0.88 4.26  
## gyros\_forearm\_z accel\_forearm\_x accel\_forearm\_y accel\_forearm\_z  
## 1 -0.59 -110 267 -149  
## 2 -0.18 212 297 -118  
## 3 0.28 154 271 -129  
## 4 1.80 -92 406 -39  
## 5 0.80 131 -93 172  
## 6 1.35 230 322 -144  
## magnet\_forearm\_x magnet\_forearm\_y magnet\_forearm\_z problem\_id  
## 1 -714 419 617 1  
## 2 -237 791 873 2  
## 3 -51 698 783 3  
## 4 -233 783 521 4  
## 5 375 -787 91 5  
## 6 -300 800 884 6

## Partioning the training set into two

Partioning the training set into two Partioning Training data set into two data sets, 60% for myTraining, 40% for myTesting:

inTrain <- createDataPartition(y=training$classe, p=0.6, list=FALSE)  
myTraining <- training[inTrain, ]; myTesting <- training[-inTrain, ]  
dim(myTraining); dim(myTesting)

## [1] 11776 160

## [1] 7846 160

## Cleaning the data

The following transformations were used to clean the data:

Transformation 1: Cleaning NearZeroVariance Variables Run this code to view possible NZV Variables:

myDataNZV <- nearZeroVar(myTraining, saveMetrics=TRUE)  
##Run this code to create another subset without NZV variables:  
myNZVvars <- names(myTraining) %in% c("new\_window", "kurtosis\_roll\_belt", "kurtosis\_picth\_belt",  
"kurtosis\_yaw\_belt", "skewness\_roll\_belt", "skewness\_roll\_belt.1", "skewness\_yaw\_belt",  
"max\_yaw\_belt", "min\_yaw\_belt", "amplitude\_yaw\_belt", "avg\_roll\_arm", "stddev\_roll\_arm",  
"var\_roll\_arm", "avg\_pitch\_arm", "stddev\_pitch\_arm", "var\_pitch\_arm", "avg\_yaw\_arm",  
"stddev\_yaw\_arm", "var\_yaw\_arm", "kurtosis\_roll\_arm", "kurtosis\_picth\_arm",  
"kurtosis\_yaw\_arm", "skewness\_roll\_arm", "skewness\_pitch\_arm", "skewness\_yaw\_arm",  
"max\_roll\_arm", "min\_roll\_arm", "min\_pitch\_arm", "amplitude\_roll\_arm", "amplitude\_pitch\_arm",  
"kurtosis\_roll\_dumbbell", "kurtosis\_picth\_dumbbell", "kurtosis\_yaw\_dumbbell", "skewness\_roll\_dumbbell",  
"skewness\_pitch\_dumbbell", "skewness\_yaw\_dumbbell", "max\_yaw\_dumbbell", "min\_yaw\_dumbbell",  
"amplitude\_yaw\_dumbbell", "kurtosis\_roll\_forearm", "kurtosis\_picth\_forearm", "kurtosis\_yaw\_forearm",  
"skewness\_roll\_forearm", "skewness\_pitch\_forearm", "skewness\_yaw\_forearm", "max\_roll\_forearm",  
"max\_yaw\_forearm", "min\_roll\_forearm", "min\_yaw\_forearm", "amplitude\_roll\_forearm",  
"amplitude\_yaw\_forearm", "avg\_roll\_forearm", "stddev\_roll\_forearm", "var\_roll\_forearm",  
"avg\_pitch\_forearm", "stddev\_pitch\_forearm", "var\_pitch\_forearm", "avg\_yaw\_forearm",  
"stddev\_yaw\_forearm", "var\_yaw\_forearm")  
myTraining <- myTraining[!myNZVvars]  
#To check the new N?? of observations  
dim(myTraining)

## [1] 11776 100

##Transformation 2: Killing first column of Dataset - ID Removing first ID variable so that it does not interfer with ML Algorithms:  
myTraining <- myTraining[c(-1)]  
  
##Transformation 3: Cleaning Variables with too many NAs. For Variables that have more than a 60% threshold of NA's I'm going to leave them out:  
trainingV3 <- myTraining #creating another subset to iterate in loop  
for(i in 1:length(myTraining)) { #for every column in the training dataset  
 if( sum( is.na( myTraining[, i] ) ) /nrow(myTraining) >= .6 ) { #if n?? NAs > 60% of total observations  
 for(j in 1:length(trainingV3)) {  
 if( length( grep(names(myTraining[i]), names(trainingV3)[j]) ) ==1) { #if the columns are the same:  
 trainingV3 <- trainingV3[ , -j] #Remove that column  
 }   
 }   
 }  
}  
#To check the new N?? of observations  
dim(trainingV3)

## [1] 11776 58

#Seting back to our set:  
myTraining <- trainingV3  
rm(trainingV3)

## Now let us do the exact same 3 transformations but for our myTesting and testing data sets.

clean1 <- colnames(myTraining)  
clean2 <- colnames(myTraining[, -58]) #already with classe column removed  
myTesting <- myTesting[clean1]  
testing <- testing[clean2]  
  
#To check the new N?? of observations  
dim(myTesting)

## [1] 7846 58

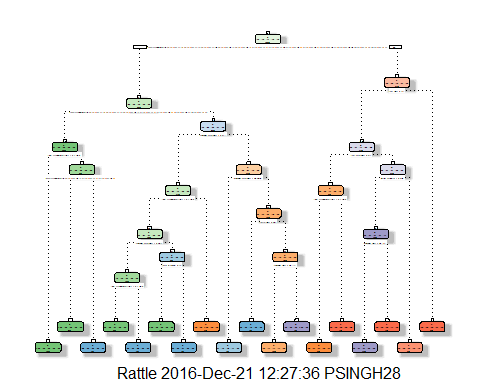
#To check the new N?? of observations  
dim(testing)

## [1] 20 57

#Note: The last column - problem\_id - which is not equal to training sets, was also "automagically" removed  
#No need for this code:  
#testing <- testing[-length(testing)]  
  
##In order to ensure proper functioning of Decision Trees and especially RandomForest Algorithm with the Test data set (data set provided), we need to coerce the data into the same type.  
for (i in 1:length(testing) ) {  
 for(j in 1:length(myTraining)) {  
 if( length( grep(names(myTraining[i]), names(testing)[j]) ) ==1) {  
 class(testing[j]) <- class(myTraining[i])  
 }   
 }   
}  
#And to make sure Coertion really worked, simple smart ass technique:  
testing <- rbind(myTraining[2, -58] , testing) #note row 2 does not mean anything, this will be removed right.. now:  
testing <- testing[-1,]

## Using ML algorithms for prediction: Decision Tree

modFitA1 <- rpart(classe ~ ., data=myTraining, method="class")  
##Note: to view the decision tree with fancy run this command:  
  
fancyRpartPlot(modFitA1)



#Predicting:  
predictionsA1 <- predict(modFitA1, myTesting, type = "class")  
##(Moment of truth) Using confusion Matrix to test results:  
  
confusionMatrix(predictionsA1, myTesting$classe)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction A B C D E  
## A 2150 60 7 1 0  
## B 61 1260 69 64 0  
## C 21 188 1269 143 4  
## D 0 10 14 857 78  
## E 0 0 9 221 1360  
##   
## Overall Statistics  
##   
## Accuracy : 0.8789   
## 95% CI : (0.8715, 0.8861)  
## No Information Rate : 0.2845   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.8468   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: A Class: B Class: C Class: D Class: E  
## Sensitivity 0.9633 0.8300 0.9276 0.6664 0.9431  
## Specificity 0.9879 0.9693 0.9450 0.9845 0.9641  
## Pos Pred Value 0.9693 0.8666 0.7809 0.8936 0.8553  
## Neg Pred Value 0.9854 0.9596 0.9841 0.9377 0.9869  
## Prevalence 0.2845 0.1935 0.1744 0.1639 0.1838  
## Detection Rate 0.2740 0.1606 0.1617 0.1092 0.1733  
## Detection Prevalence 0.2827 0.1853 0.2071 0.1222 0.2027  
## Balanced Accuracy 0.9756 0.8997 0.9363 0.8254 0.9536

##Using ML algorithms for prediction: Random Forests  
modFitB1 <- randomForest(classe ~. , data=myTraining)  
#Predicting in-sample error:  
  
predictionsB1 <- predict(modFitB1, myTesting, type = "class")  
##(Moment of truth) Using confusion Matrix to test results:  
  
confusionMatrix(predictionsB1, myTesting$classe)

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction A B C D E  
## A 2231 2 0 0 0  
## B 1 1516 2 0 0  
## C 0 0 1366 3 0  
## D 0 0 0 1282 2  
## E 0 0 0 1 1440  
##   
## Overall Statistics  
##   
## Accuracy : 0.9986   
## 95% CI : (0.9975, 0.9993)  
## No Information Rate : 0.2845   
## P-Value [Acc > NIR] : < 2.2e-16   
##   
## Kappa : 0.9982   
## Mcnemar's Test P-Value : NA   
##   
## Statistics by Class:  
##   
## Class: A Class: B Class: C Class: D Class: E  
## Sensitivity 0.9996 0.9987 0.9985 0.9969 0.9986  
## Specificity 0.9996 0.9995 0.9995 0.9997 0.9998  
## Pos Pred Value 0.9991 0.9980 0.9978 0.9984 0.9993  
## Neg Pred Value 0.9998 0.9997 0.9997 0.9994 0.9997  
## Prevalence 0.2845 0.1935 0.1744 0.1639 0.1838  
## Detection Rate 0.2843 0.1932 0.1741 0.1634 0.1835  
## Detection Prevalence 0.2846 0.1936 0.1745 0.1637 0.1837  
## Balanced Accuracy 0.9996 0.9991 0.9990 0.9983 0.9992

Random Forests yielded better Results, as expected!

Generating Files to submit as answers for the Assignment: Finally, using the provided Test Set out-of-sample error.

For Random Forests we use the following formula, which yielded a much better prediction in in-sample:

predictionsB2 <- predict(modFitB1, testing, type = "class")  
#Function to generate files with predictions to submit for assignment  
  
pml\_write\_files = function(x){  
 n = length(x)  
 for(i in 1:n){  
 filename = paste0("problem\_id\_",i,".txt")  
 write.table(x[i],file=filename,quote=FALSE,row.names=FALSE,col.names=FALSE)  
 }  
}  
  
pml\_write\_files(predictionsB2)