# PML Final Project Assignement

Azeddine ELHASSOUNY

5 mars 2016

## Introduction

Instructions

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.

Review criteriamoins 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.

Prediction Assignment Writeupmoins 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.

## Geting and loading the data

```
training <- read.csv("pml-training.csv", na.strings=c("NA","#DIV/0!",""))
testing <- read.csv("pml-testing.csv", na.strings=c("NA","#DIV/0!",""))
```

## Partioning the training set into two

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

## Cleaning the data

### Remove nearZero variance variables

```
nzv <- nearZeroVar(myTraining, saveMetrics=TRUE)
myTraining <- myTraining[,nzv$nzv==FALSE]

nzv<- nearZeroVar(myTesting,saveMetrics=TRUE)
myTesting <- myTesting[,nzv$nzv==FALSE]
```

### Remove variables with more of NA and 7 firsts columns

```
feature_set <- colnames(myTraining[colSums(is.na(myTraining)) == 0])[-(1:7)]
model_data <- myTraining[feature_set]
```

## Transform the myTesting and testing data sets to have same names columns and class type

```
clean1 <- colnames(model_data)
clean2 <- colnames(model_data[, -52])  # remove the classe column
myTesting <- myTesting[clean1]         # allow only variables in myTesting
that are also in myTraining
testing <- testing[clean2]             # allow only variables in testing that
are also in myTraining
```

## Building model combined 5 random forest, using 150 trees

```
registerDoParallel()
x <-model_data[-ncol(model_data)]
y <- model_data$classe

rf <- foreach(ntree=rep(150, 6), .combine=randomForest::combine,
.packages='randomForest') %dopar% {
  randomForest(x, y, ntree=ntree)
}
```

## Prediction and confusionmatrix

```
predictions2 <- predict(rf, newdata=myTesting)
confusionMatrix(predictions2,myTesting$classe)
```

## Submited data

```
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)
  }
}


x <- testing
x <- x[feature_set[feature_set!='classe']]
answers <- predict(rf, newdata=x)

answers

pml_write_files(answers)
```