Prediction model

S I

May 20, 2018

Download data

```
url_train <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-training.csv"
dat_train <- "pml-training.csv"
download.file(url=url_train, destfile=dat_train, method = "auto")
url_test <- "https://d396qusza40orc.cloudfront.net/predmachlearn/pml-testing.csv"
dat_test <- "pml-testing.csv"
download.file(url=url_test, destfile=dat_test, method = "auto")
```

Import and clean data.

We perform mainly the following steps: Import data and convert empty values to NA. Check number and percentage of NAs in test set. Remove columns with only NAs in test set. We are left with two datasets that have 60 variables, instead of 160. Check to see that colnames are the same in the two new datasets Remove id columns from the new datasets

```
df_train <- read.csv(dat_train, na.strings=c("NA",""), header=TRUE)
colnames_train <- colnames(df_train)
df_test <- read.csv(dat_test, na.strings=c("NA",""), header=TRUE)
colnames_test <- colnames(df_test)
colSums(!is.na(df_test))
```

```
##                        X                user_name     raw_timestamp_part_1
##                       20                       20                       20
##     raw_timestamp_part_2           cvtd_timestamp               new_window
##                       20                       20                       20
##               num_window                roll_belt               pitch_belt
##                       20                       20                       20
##                yaw_belt         total_accel_belt       kurtosis_roll_belt
##                       20                       20                        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                        0
##           max_picth_belt             max_yaw_belt            min_roll_belt
##                        0                        0                        0
##           min_pitch_belt             min_yaw_belt      amplitude_roll_belt
##                        0                        0                        0
##     amplitude_pitch_belt       amplitude_yaw_belt     var_total_accel_belt
##                        0                        0                        0
##            avg_roll_belt         stddev_roll_belt            var_roll_belt
##                        0                        0                        0
##           avg_pitch_belt        stddev_pitch_belt           var_pitch_belt
##                        0                        0                        0
##             avg_yaw_belt          stddev_yaw_belt             var_yaw_belt
##                        0                        0                        0
```

```
##             gyros_belt_x             gyros_belt_y             gyros_belt_z
##                       20                       20                       20
##             accel_belt_x             accel_belt_y             accel_belt_z
##                       20                       20                       20
##            magnet_belt_x            magnet_belt_y            magnet_belt_z
##                       20                       20                       20
##                roll_arm                 pitch_arm                  yaw_arm
##                       20                       20                       20
##          total_accel_arm            var_accel_arm            avg_roll_arm
##                       20                        0                        0
##          stddev_roll_arm             var_roll_arm            avg_pitch_arm
##                        0                        0                        0
##         stddev_pitch_arm            var_pitch_arm              avg_yaw_arm
##                        0                        0                        0
##           stddev_yaw_arm             var_yaw_arm             gyros_arm_x
##                        0                        0                       20
##              gyros_arm_y              gyros_arm_z              accel_arm_x
##                       20                       20                       20
##              accel_arm_y              accel_arm_z             magnet_arm_x
##                       20                       20                       20
##             magnet_arm_y             magnet_arm_z        kurtosis_roll_arm
##                       20                       20                        0
##        kurtosis_picth_arm         kurtosis_yaw_arm        skewness_roll_arm
##                        0                        0                        0
##        skewness_pitch_arm         skewness_yaw_arm             max_roll_arm
##                        0                        0                        0
##             max_picth_arm              max_yaw_arm             min_roll_arm
##                        0                        0                        0
##             min_pitch_arm              min_yaw_arm       amplitude_roll_arm
##                        0                        0                        0
##       amplitude_pitch_arm        amplitude_yaw_arm            roll_dumbbell
##                        0                        0                       20
##            pitch_dumbbell             yaw_dumbbell   kurtosis_roll_dumbbell
##                       20                       20                        0
##   kurtosis_picth_dumbbell    kurtosis_yaw_dumbbell   skewness_roll_dumbbell
##                        0                        0                        0
##   skewness_pitch_dumbbell    skewness_yaw_dumbbell        max_roll_dumbbell
##                        0                        0                        0
##        max_picth_dumbbell         max_yaw_dumbbell        min_roll_dumbbell
##                        0                        0                        0
##        min_pitch_dumbbell         min_yaw_dumbbell  amplitude_roll_dumbbell
##                        0                        0                        0
## amplitude_pitch_dumbbell   amplitude_yaw_dumbbell     total_accel_dumbbell
##                        0                        0                       20
##        var_accel_dumbbell         avg_roll_dumbbell     stddev_roll_dumbbell
##                        0                        0                        0
##         var_roll_dumbbell        avg_pitch_dumbbell    stddev_pitch_dumbbell
##                        0                        0                        0
##        var_pitch_dumbbell         avg_yaw_dumbbell      stddev_yaw_dumbbell
##                        0                        0                        0
##          var_yaw_dumbbell         gyros_dumbbell_x         gyros_dumbbell_y
##                        0                       20                       20
##          gyros_dumbbell_z         accel_dumbbell_x         accel_dumbbell_y
##                       20                       20                       20
```

```
##         accel_dumbbell_z        magnet_dumbbell_x        magnet_dumbbell_y
##                       20                       20                       20
##        magnet_dumbbell_z             roll_forearm            pitch_forearm
##                       20                       20                       20
##              yaw_forearm    kurtosis_roll_forearm   kurtosis_picth_forearm
##                       20                        0                        0
##     kurtosis_yaw_forearm    skewness_roll_forearm   skewness_pitch_forearm
##                        0                        0                        0
##     skewness_yaw_forearm         max_roll_forearm        max_picth_forearm
##                        0                        0                        0
##          max_yaw_forearm         min_roll_forearm        min_pitch_forearm
##                        0                        0                        0
##          min_yaw_forearm   amplitude_roll_forearm  amplitude_pitch_forearm
##                        0                        0                        0
##    amplitude_yaw_forearm      total_accel_forearm        var_accel_forearm
##                        0                       20                        0
##         avg_roll_forearm      stddev_roll_forearm         var_roll_forearm
##                        0                        0                        0
##        avg_pitch_forearm     stddev_pitch_forearm        var_pitch_forearm
##                        0                        0                        0
##          avg_yaw_forearm       stddev_yaw_forearm          var_yaw_forearm
##                        0                        0                        0
##          gyros_forearm_x         gyros_forearm_y          gyros_forearm_z
##                       20                       20                       20
##          accel_forearm_x         accel_forearm_y          accel_forearm_z
##                       20                       20                       20
##         magnet_forearm_x        magnet_forearm_y         magnet_forearm_z
##                       20                       20                       20
##               problem_id
##                       20
```

```
colMeans(is.na(df_test))*100
```

```
##                        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                      100
##      kurtosis_picth_belt        kurtosis_yaw_belt       skewness_roll_belt
##                      100                      100                      100
##     skewness_roll_belt.1        skewness_yaw_belt            max_roll_belt
##                      100                      100                      100
##           max_picth_belt             max_yaw_belt            min_roll_belt
##                      100                      100                      100
##           min_pitch_belt             min_yaw_belt       amplitude_roll_belt
##                      100                      100                      100
##     amplitude_pitch_belt       amplitude_yaw_belt      var_total_accel_belt
##                      100                      100                      100
##            avg_roll_belt         stddev_roll_belt            var_roll_belt
##                      100                      100                      100
##           avg_pitch_belt        stddev_pitch_belt           var_pitch_belt
##                      100                      100                      100
```

```
##             avg_yaw_belt          stddev_yaw_belt             var_yaw_belt
##                      100                      100                      100
##            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                      100                      100
##          stddev_roll_arm            var_roll_arm           avg_pitch_arm
##                      100                      100                      100
##         stddev_pitch_arm           var_pitch_arm             avg_yaw_arm
##                      100                      100                      100
##           stddev_yaw_arm             var_yaw_arm             gyros_arm_x
##                      100                      100                        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                      100
##       kurtosis_picth_arm       kurtosis_yaw_arm       skewness_roll_arm
##                      100                      100                      100
##       skewness_pitch_arm       skewness_yaw_arm            max_roll_arm
##                      100                      100                      100
##            max_picth_arm            max_yaw_arm            min_roll_arm
##                      100                      100                      100
##            min_pitch_arm            min_yaw_arm      amplitude_roll_arm
##                      100                      100                      100
##      amplitude_pitch_arm      amplitude_yaw_arm           roll_dumbbell
##                      100                      100                        0
##           pitch_dumbbell             yaw_dumbbell   kurtosis_roll_dumbbell
##                        0                        0                      100
##  kurtosis_picth_dumbbell  kurtosis_yaw_dumbbell  skewness_roll_dumbbell
##                      100                      100                      100
##  skewness_pitch_dumbbell  skewness_yaw_dumbbell       max_roll_dumbbell
##                      100                      100                      100
##       max_picth_dumbbell       max_yaw_dumbbell       min_roll_dumbbell
##                      100                      100                      100
##       min_pitch_dumbbell       min_yaw_dumbbell  amplitude_roll_dumbbell
##                      100                      100                      100
## amplitude_pitch_dumbbell amplitude_yaw_dumbbell     total_accel_dumbbell
##                      100                      100                        0
##       var_accel_dumbbell       avg_roll_dumbbell     stddev_roll_dumbbell
##                      100                      100                      100
##        var_roll_dumbbell      avg_pitch_dumbbell    stddev_pitch_dumbbell
##                      100                      100                      100
##       var_pitch_dumbbell       avg_yaw_dumbbell     stddev_yaw_dumbbell
##                      100                      100                      100
##         var_yaw_dumbbell       gyros_dumbbell_x       gyros_dumbbell_y
##                      100                        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_picth_forearm
##                        0                      100                      100
##     kurtosis_yaw_forearm    skewness_roll_forearm   skewness_pitch_forearm
##                      100                      100                      100
##     skewness_yaw_forearm         max_roll_forearm        max_picth_forearm
##                      100                      100                      100
##          max_yaw_forearm         min_roll_forearm        min_pitch_forearm
##                      100                      100                      100
##          min_yaw_forearm   amplitude_roll_forearm  amplitude_pitch_forearm
##                      100                      100                      100
##    amplitude_yaw_forearm      total_accel_forearm        var_accel_forearm
##                      100                        0                      100
##         avg_roll_forearm      stddev_roll_forearm         var_roll_forearm
##                      100                      100                      100
##        avg_pitch_forearm     stddev_pitch_forearm        var_pitch_forearm
##                      100                      100                      100
##          avg_yaw_forearm       stddev_yaw_forearm          var_yaw_forearm
##                      100                      100                      100
##          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
##               problem_id
##                        0
```

```
df_testNoNA <- df_test[, colSums(is.na(df_test)) != nrow(df_test)]
df_trainSub <- df_train[, colSums(is.na(df_test)) != nrow(df_test)]
dim(df_testNoNA)
```

```
## [1] 20 60
```

```
dim(df_trainSub)
```

```
## [1] 19622    60
```

```
colnames_trainSub <- colnames(df_trainSub)
colnames_testNoNA <- colnames(df_testNoNA)
setdiff(colnames_testNoNA,colnames_trainSub)
```

```
## [1] "problem_id"
```

```
setdiff(colnames_trainSub,colnames_testNoNA)
```

```
## [1] "classe"
```

```
df_testTrim<- df_testNoNA[,c(-1, -60)]
df_trainTrim<- df_trainSub[,-1]
```

Data processing and model

For this stage we perform a series of steps as follows: Split the data 65% for training and 35% for testing. Then use the training set (df_trainTrim) as the source for the new training and testing sets and leave the test set (df_testTrim) untouched.

We use caret package to perform principle component analysis, use decision tree method and then Random Fores (we draw also some relevant plots)

```
library(lattice)
library(ggplot2)
library(caret)
```

```
## Warning: package 'caret' was built under R version 3.4.4
```

```
set.seed(54321)

TrainSub <- createDataPartition(y=df_trainTrim$classe, p=0.65, list=FALSE)
myTraining <- df_trainTrim[TrainSub, ]
myTesting <- df_trainTrim[-TrainSub, ]
dim(myTraining)
```

```
## [1] 12757    59
```

```
dim(myTesting)
```

```
## [1] 6865   59
```

```
nsv<- nearZeroVar(df_trainTrim, saveMetrics = TRUE)
nsv
```

```
##                      freqRatio percentUnique zeroVar   nzv
## user_name             1.100679    0.03057792   FALSE FALSE
## raw_timestamp_part_1  1.000000    4.26562022   FALSE FALSE
## raw_timestamp_part_2  1.000000   85.53154622   FALSE FALSE
## cvtd_timestamp        1.000668    0.10192641   FALSE FALSE
## new_window           47.330049    0.01019264   FALSE  TRUE
## num_window            1.000000    4.37264295   FALSE FALSE
## roll_belt             1.101904    6.77810621   FALSE FALSE
## pitch_belt            1.036082    9.37722964   FALSE FALSE
## yaw_belt              1.058480    9.97349913   FALSE FALSE
## total_accel_belt      1.063160    0.14779329   FALSE FALSE
## gyros_belt_x          1.058651    0.71348486   FALSE FALSE
## gyros_belt_y          1.144000    0.35164611   FALSE FALSE
## gyros_belt_z          1.066214    0.86127816   FALSE FALSE
## accel_belt_x          1.055412    0.83579655   FALSE FALSE
## accel_belt_y          1.113725    0.72877383   FALSE FALSE
## accel_belt_z          1.078767    1.52379982   FALSE FALSE
## magnet_belt_x         1.090141    1.66649679   FALSE FALSE
## magnet_belt_y         1.099688    1.51870350   FALSE FALSE
## magnet_belt_z         1.006369    2.32901845   FALSE FALSE
## roll_arm             52.338462   13.52563449   FALSE FALSE
## pitch_arm            87.256410   15.73234125   FALSE FALSE
## yaw_arm              33.029126   14.65701763   FALSE FALSE
## total_accel_arm       1.024526    0.33635715   FALSE FALSE
## gyros_arm_x           1.015504    3.27693405   FALSE FALSE
## gyros_arm_y           1.454369    1.91621649   FALSE FALSE
```

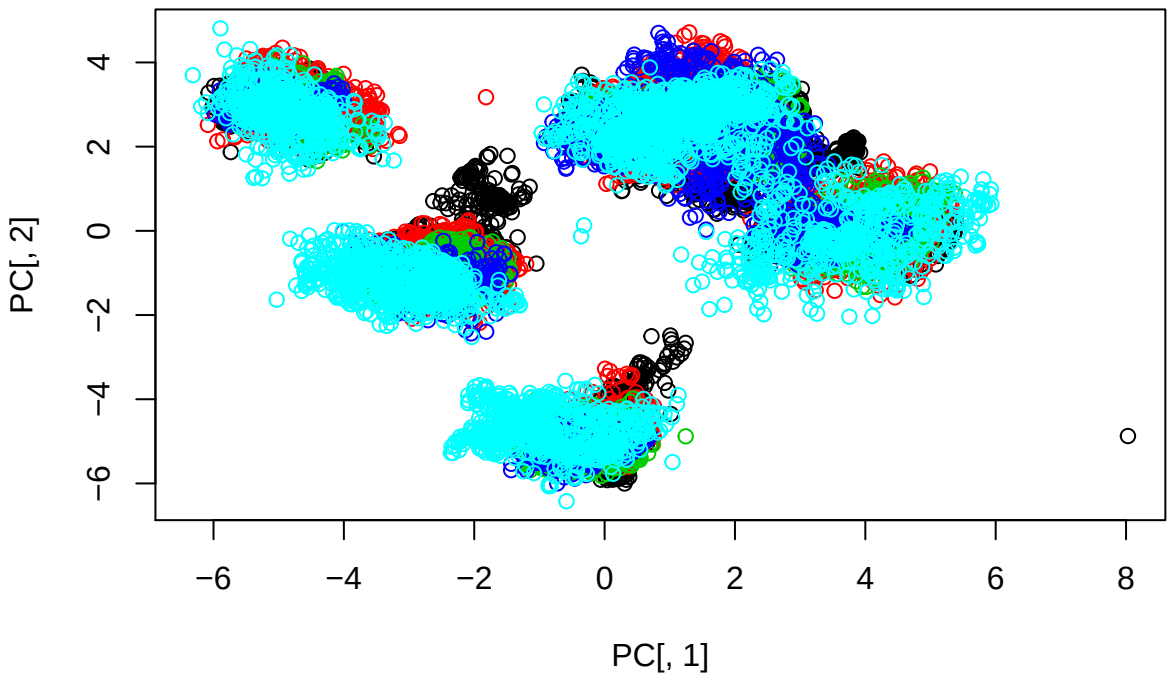
```
## gyros_arm_z           1.110687    1.26388747   FALSE FALSE
## accel_arm_x           1.017341    3.95984099   FALSE FALSE
## accel_arm_y           1.140187    2.73672409   FALSE FALSE
## accel_arm_z           1.128000    4.03628580   FALSE FALSE
## magnet_arm_x          1.000000    6.82397309   FALSE FALSE
## magnet_arm_y          1.056818    4.44399144   FALSE FALSE
## magnet_arm_z          1.036364    6.44684538   FALSE FALSE
## roll_dumbbell         1.022388   84.20650290   FALSE FALSE
## pitch_dumbbell        2.277372   81.74498012   FALSE FALSE
## yaw_dumbbell          1.132231   83.48282540   FALSE FALSE
## total_accel_dumbbell  1.072634    0.21914178   FALSE FALSE
## gyros_dumbbell_x      1.003268    1.22821323   FALSE FALSE
## gyros_dumbbell_y      1.264957    1.41677709   FALSE FALSE
## gyros_dumbbell_z      1.060100    1.04984201   FALSE FALSE
## accel_dumbbell_x      1.018018    2.16593619   FALSE FALSE
## accel_dumbbell_y      1.053061    2.37488533   FALSE FALSE
## accel_dumbbell_z      1.133333    2.08949139   FALSE FALSE
## magnet_dumbbell_x     1.098266    5.74864948   FALSE FALSE
## magnet_dumbbell_y     1.197740    4.30129447   FALSE FALSE
## magnet_dumbbell_z     1.020833    3.44511263   FALSE FALSE
## roll_forearm         11.589286   11.08959331   FALSE FALSE
## pitch_forearm        65.983051   14.85577413   FALSE FALSE
## yaw_forearm          15.322835   10.14677403   FALSE FALSE
## total_accel_forearm   1.128928    0.35674243   FALSE FALSE
## gyros_forearm_x       1.059273    1.51870350   FALSE FALSE
## gyros_forearm_y       1.036554    3.77637346   FALSE FALSE
## gyros_forearm_z       1.122917    1.56457038   FALSE FALSE
## accel_forearm_x       1.126437    4.04647844   FALSE FALSE
## accel_forearm_y       1.059406    5.11160942   FALSE FALSE
## accel_forearm_z       1.006250    2.95586586   FALSE FALSE
## magnet_forearm_x      1.012346    7.76679238   FALSE FALSE
## magnet_forearm_y      1.246914    9.54031189   FALSE FALSE
## magnet_forearm_z      1.000000    8.57710733   FALSE FALSE
## classe                1.469581    0.02548160   FALSE FALSE
```

```
M <- abs(cor(df_trainTrim[,c(-1,-4,-5,-59)]))
diag(M) <- 0
which(M>0.8, arr.ind = TRUE)
```

```
##                   row col
## yaw_belt            6   4
## total_accel_belt    7   4
## accel_belt_y       12   4
## accel_belt_z       13   4
## accel_belt_x       11   5
## magnet_belt_x      14   5
## roll_belt           4   6
## roll_belt           4   7
## accel_belt_y       12   7
## accel_belt_z       13   7
## pitch_belt          5  11
## magnet_belt_x      14  11
## roll_belt           4  12
## total_accel_belt    7  12
## accel_belt_z       13  12
```

```
## roll_belt          4  13
## total_accel_belt   7  13
## accel_belt_y      12  13
## pitch_belt         5  14
## accel_belt_x      11  14
## gyros_arm_y       22  21
## gyros_arm_x       21  22
## magnet_arm_x      27  24
## accel_arm_x       24  27
## magnet_arm_z      29  28
## magnet_arm_y      28  29
## accel_dumbbell_x  37  31
## accel_dumbbell_z  39  32
## gyros_dumbbell_z  36  34
## gyros_forearm_z   49  34
## gyros_dumbbell_x  34  36
## gyros_forearm_z   49  36
## pitch_dumbbell    31  37
## yaw_dumbbell      32  39
## gyros_forearm_z   49  48
## gyros_dumbbell_x  34  49
## gyros_dumbbell_z  36  49
## gyros_forearm_y   48  49
```

```
preProc <- preProcess(df_trainTrim[,c(-1,-4,-5,-59)], method = "pca", pcaComp = 2)
PC<-predict(preProc,df_trainTrim[,c(-1,-4,-5,-59)])
plot(PC[,1],PC[,2], col=df_trainTrim$classe)
```




```
#Plot(PC[,1],PC[,2], col=df_trainTrim$user_name)
```

```
modDT <- rpart(classe ~ ., data=myTraining, method="class")
predDT <- predict(modDT, myTesting, type = "class")
#fancyRpartPlot(modDT)

cfmDT<-confusionMatrix(predDT, myTesting$classe)
cfmDT
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 1880   58   10    3    0
##          B   56 1108   91   57    0
##          C   17  153 1077  177   55
##          D    0    9   10  707   63
##          E    0    0    9  181 1144
##
## Overall Statistics
##
##                Accuracy : 0.8618
##                  95% CI : (0.8534, 0.8698)
##     No Information Rate : 0.2845
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.825
```

```
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            0.9626   0.8343   0.8997   0.6284   0.9065
## Specificity            0.9855   0.9632   0.9291   0.9857   0.9661
## Pos Pred Value         0.9636   0.8445   0.7282   0.8961   0.8576
## Neg Pred Value         0.9851   0.9604   0.9777   0.9312   0.9787
## Prevalence             0.2845   0.1934   0.1744   0.1639   0.1838
## Detection Rate         0.2739   0.1614   0.1569   0.1030   0.1666
## Detection Prevalence   0.2842   0.1911   0.2154   0.1149   0.1943
## Balanced Accuracy      0.9741   0.8987   0.9144   0.8071   0.9363
```

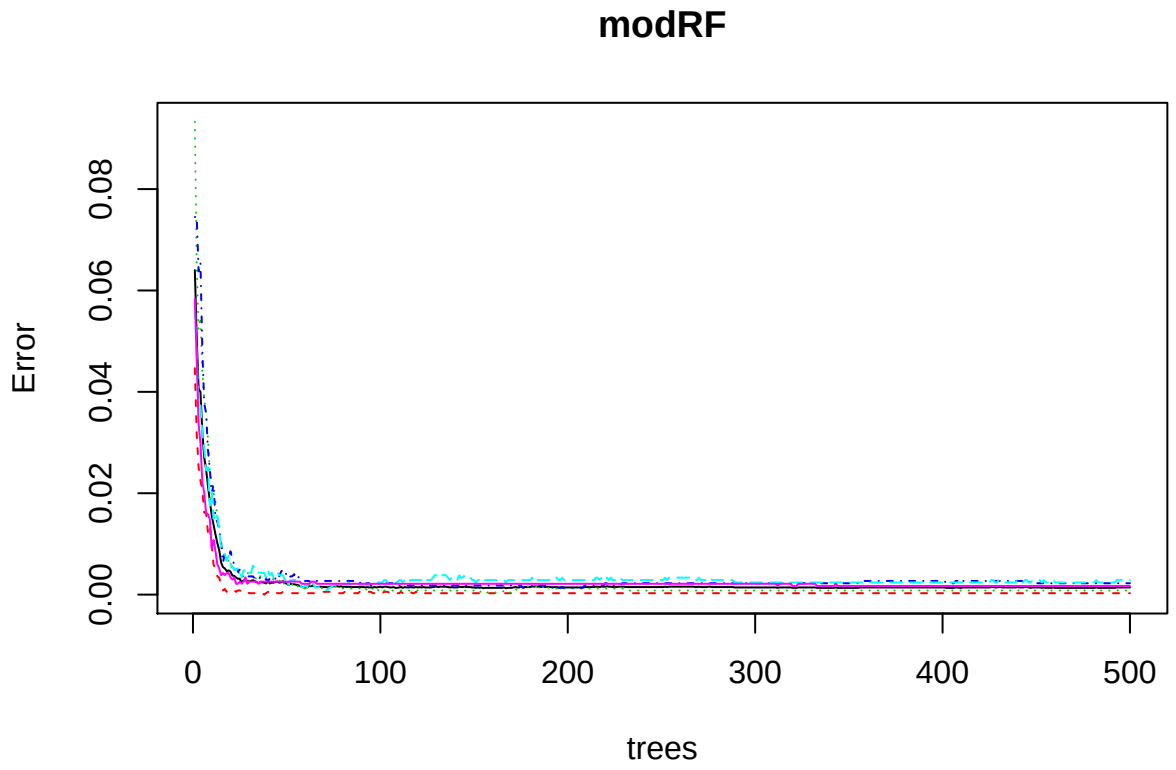
```
(accuracy_dt <- cfmDT$overall[1])
```

```
##  Accuracy
## 0.8617626
```

```
modRF <- randomForest(classe ~ ., data=myTraining)
predRF <- predict(modRF, myTesting, type = "class")
cfmRF<-confusionMatrix(predRF, myTesting$classe)
cfmRF
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    A    B    C    D    E
##          A 1952    0    0    0    0
##          B    1 1328    2    0    0
##          C    0    0 1194    3    0
##          D    0    0    1 1121    0
##          E    0    0    0    1 1262
##
## Overall Statistics
##
##                Accuracy : 0.9988
##                  95% CI : (0.9977, 0.9995)
##     No Information Rate : 0.2845
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.9985
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: A Class: B Class: C Class: D Class: E
## Sensitivity            0.9995   1.0000   0.9975   0.9964   1.0000
## Specificity            1.0000   0.9995   0.9995   0.9998   0.9998
## Pos Pred Value         1.0000   0.9977   0.9975   0.9991   0.9992
## Neg Pred Value         0.9998   1.0000   0.9995   0.9993   1.0000
## Prevalence             0.2845   0.1934   0.1744   0.1639   0.1838
## Detection Rate         0.2843   0.1934   0.1739   0.1633   0.1838
## Detection Prevalence   0.2843   0.1939   0.1744   0.1634   0.1840
## Balanced Accuracy      0.9997   0.9997   0.9985   0.9981   0.9999
```

```
plot(modRF)
```




```
(accuracy_rf <- cfmRF$overall[1])
```

```
##  Accuracy
## 0.9988347
```

```
predFinalDT <- predict(modDT, df_testTrim, type="class")
predFinalDT
```

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
##  B  A  C  A  A  E  D  C  A  A  B  C  B  A  E  E  A  B  B  B
## Levels: A B C D E
```

```
# We correct the incosistencies

testing <- rbind(myTraining[21, -59] , df_testTrim)
testing <- testing[-1,]
```

Quiz answer generation

```
library(compare)
```

```
library(utils)
predFinalRF <- predict(modRF, testing, type="class")
predFinalRF
```

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20
##  B  A  B  A  A  E  D  B  A  A  B  C  B  A  E  E  A  B  B  B
## Levels: A B C D E
```