Human Activity Recognition

The data set contains 561 features which were calculated from time series data from a waist sensor. The data represents 6 possible human activities, walking, walking up stairs, walking down stairs, sitting, standing, and lying (coded as numbers 1-6 respectively).

Loading packages:

In [49]:
```
library(dplyr)
library(ggplot2)
library(lattice)
library(stringr)
library(gridExtra)
library(caret)
library(rpart)
library(readr)
library(e1071)
options(repos='https://cran.cnr.berkeley.edu/')
install.packages('fastICA')
install.packages('klaR')
install.packages('kknn')
install.packages('gbm')

library(fastICA)
library(klaR)
library(kknn)
library(gbm)
```

...

Data Exploration

Loading Data:

In [26]:
```
X <- read_table('C:/Datasets/UCI HAR Dataset/train/X_train.txt', col_names=FALSE)
y <- read.csv('C:/Datasets/UCI HAR Dataset/train/y_train.txt', header = FALSE)
```

...

In [27]:
```
dim(as.matrix(X))
head(X,5)
```

7352 561

X1	X2	X3	X4	X5	X6	X7	X8	X
0.2885845	-0.02029417	-0.1329051	-0.9952786	-0.9831106	-0.9135264	-0.9951121	-0.9831846	-0.923527
0.2784188	-0.01641057	-0.1235202	-0.9982453	-0.9753002	-0.9603220	-0.9988072	-0.9749144	-0.957686
0.2796531	-0.01946716	-0.1134617	-0.9953796	-0.9671870	-0.9789440	-0.9965199	-0.9636684	-0.977468
0.2791739	-0.02620065	-0.1232826	-0.9960915	-0.9834027	-0.9906751	-0.9970995	-0.9827498	-0.989302
0.2766288	-0.01656965	-0.1153619	-0.9981386	-0.9808173	-0.9904816	-0.9983211	-0.9796719	-0.990441

In [28]:
```
head(y,5)
```

V1
5
5
5
5
5

The response vector is an integer vector, which will be converted to a factor.

In [29]:
```
y[,1] <- factor(y[,1])
summary(y)
```

```
V1
1:1226
2:1073
3: 986
4:1286
5:1374
6:1407
```

Check for duplicates and missing values.

In [30]:
```
sum(duplicated(X))
```

0

In [31]:
```
sum(is.na(X))
sum(is.na(y))
```

0

0

The X matrix and y vector will be combined into a data frame for further processing.

In [32]:
```
df <- as.data.frame(X)
df$y <- y
```

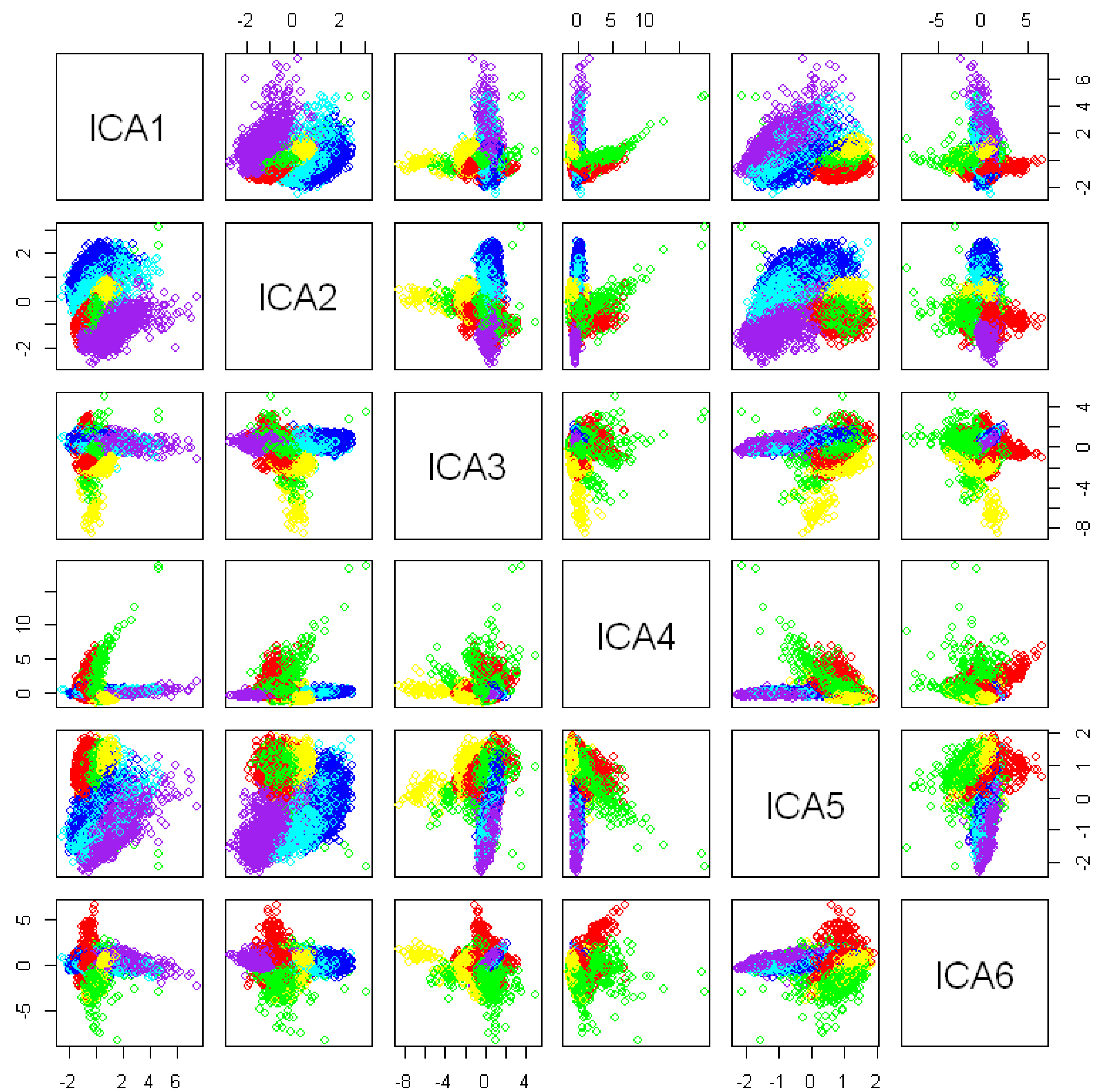
Since the features are not readily interpretable, further summarizing of the data will not provide much insight. Instead, the data will be visualized in pairwise plots.

Data Visualization

Since the data contains so many features, and the features in themselves already are not so easily interpretable, visualization will be performed by first using ICA to extract independent components.

In [47]:
```
ICA <- preProcess(X,method='ica',n.comp=6)
Xica <- predict(ICA, X)
```

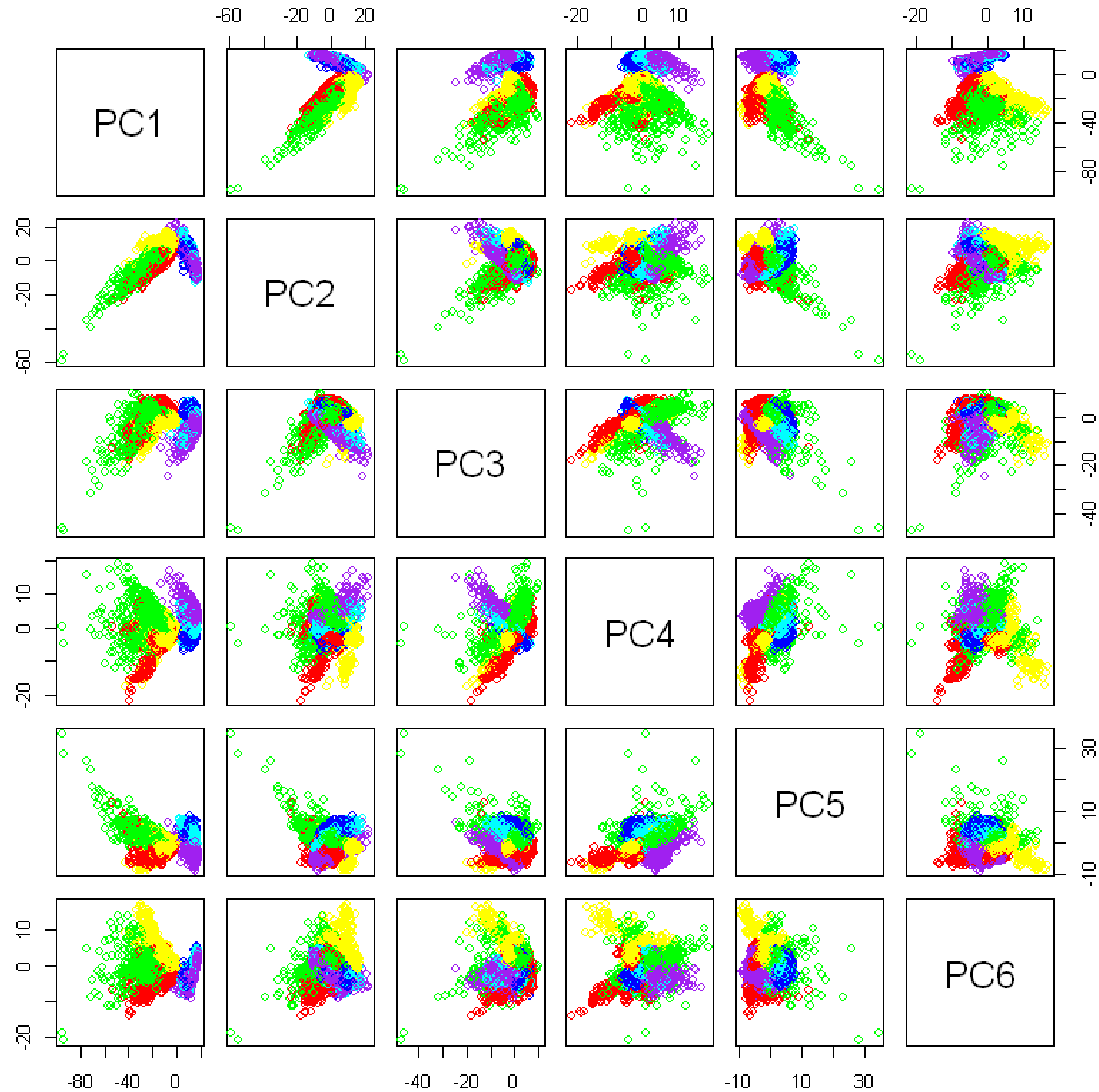
In [48]:
```
color <- character(7352)
colors <- c('red','yellow','green','cyan','blue','purple')
for(i in 1:6){
    color[y$V1 == paste(i)]<-colors[i]
}
pairs(Xica[,1:6],col=color)
```



For comparison, a similar plot is generated using PCA.

In [45]:
```
PCA <- preProcess(X,method='pca')
Xpca <- predict(PCA, X)
```

```
In [46]: pairs(Xpca[,1:6],col=color)
```



Under ICA and PCA coordinates, the first three activity levels are sometimes separated from the last three levels. This makes sense because the first three involve walking (straight, up stairs, and down stairs) and the last three are stationary activities (sitting, standing, lying).

Both techniques separate the groups to a small degree, but there is still significant overlap. Modeling will be performed using the full set of predictors if possible.

Modeling

Linear Discriminant Analysis

In [52]:
```
LDAmodel <- train(X, y$V1, method = 'lda',trControl =trainControl(method='repeatedcv
LDAmodel
```

```
Warning message:
"Setting row names on a tibble is deprecated."Warning message in lda.default(x, g
rouping, ...):
"variables are collinear"Warning message:
"Setting row names on a tibble is deprecated."Warning message in lda.default(x, g
rouping, ...):
"variables are collinear"Warning message:
"Setting row names on a tibble is deprecated."Warning message in lda.default(x, g
rouping, ...):
"variables are collinear"Warning message:
"Setting row names on a tibble is deprecated."Warning message in lda.default(x, g
rouping, ...):
"variables are collinear"Warning message:
"Setting row names on a tibble is deprecated."Warning message in lda.default(x, g
rouping, ...):
"variables are collinear"Warning message:
"Setting row names on a tibble is deprecated."Warning message in lda.default(x, g
rouping, ...):
"variables are collinear"Warning message:
```

Logistic Regression

In [54]:
```
Logisticmodel <- train(X, y$V1, method = 'multinom', MaxNWts = 4000, trControl =trai
Logisticmodel
```

```
Warning message:
"Setting row names on a tibble is deprecated."

# weights:  3378 (2810 variable)
initial  value 10535.545679
iter  10 value 4657.862717
iter  20 value 1469.346111
iter  30 value 964.551730
iter  40 value 601.231132
iter  50 value 240.204960
iter  60 value 94.408349
iter  70 value 41.713469
iter  80 value 4.568333
iter  90 value 0.111541
iter 100 value 0.010883
final  value 0.010883
stopped after 100 iterations

Warning message:
"Setting row names on a tibble is deprecated."
```

Support Vector Machine

SVM and kNN need scaled features.

In [55]:
```r
scaling <- preProcess(X, method='scale')
scaledX <- predict(scaling, X)
```

In [56]:
```r
SVMLinmodel <- train(scaledX, y$V1, method = 'svmLinear', trControl =trainControl(me
SVMLinmodel
```

```
Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."

Support Vector Machines with Linear Kernel
```

In [57]:
```r
SVMRadmodel <- train(scaledX, y$V1, method = 'svmRadial', trControl =trainControl(me
SVMRadmodel
```

```
Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
```

k Nearest Neighbors

In [60]:
```
kNNmodel <- train(scaledX, y$V1, method = 'knn', trControl =trainControl(method='rep
kNNmodel
```

```
Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
```

Naive Bayes

In [61]:
```
NBmodel <- train(X, y$V1, method = 'nb',trControl =trainControl(method='repeatedcv',
NBmodel
```

```
Warning message:
"Setting row names on a tibble is deprecated."Warning message in FUN(X[[i]],
...):
"Numerical 0 probability for all classes with observation 5"Warning message in FU
N(X[[i]], ...):
"Numerical 0 probability for all classes with observation 8"Warning message in FU
N(X[[i]], ...):
"Numerical 0 probability for all classes with observation 9"Warning message in FU
N(X[[i]], ...):
"Numerical 0 probability for all classes with observation 99"Warning message in F
UN(X[[i]], ...):
"Numerical 0 probability for all classes with observation 127"Warning message in
FUN(X[[i]], ...):
"Numerical 0 probability for all classes with observation 208"Warning message in
FUN(X[[i]], ...):
"Numerical 0 probability for all classes with observation 209"Warning message in
FUN(X[[i]], ...):
"Numerical 0 probability for all classes with observation 214"Warning message in
FUN(X[[i]], ...):
```

Decision Tree

In [62]:
```
Treemodel <- train(X, y$V1, method = 'rpart',trControl =trainControl(method='repeate
Treemodel
```

```
Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."

CART
```

In [64]:
```
Grid = expand.grid(cp=c(0.0001, 0.001, 0.01, 0.1))
Treemodel <- train(X, y$V1, method = 'rpart',trControl =trainControl(method='repeate
Treemodel
```

```
Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."
```

In [66]:
```
Grid = expand.grid(cp=c(0.000001, 0.00001, 0.0001))
Treemodel <- train(X, y$V1, method = 'rpart',trControl =trainControl(method='repeate
Treemodel
```

```
Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."

CART
```

Random Forest

In [71]:
```
RFmodel <- train(X, y$V1, method = 'rf',trControl =trainControl(method='repeatedcv',
RFmodel
```

```
Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
"Setting row names on a tibble is deprecated."Warning message:
```

All the models that involve hyperplanes of separation perform better in cross validation than models that do not assume linear decision boundaries. This is probably due to the large number of features, which causes overfitting in those models. The models with hyplerplanes of separation have high accuracy in cross validation, which indicate minimal overfitting even with the large number of features. LDA, linear SVM, and logistic regression have similar cross validation accuracy, but LDA is faster. Therefore LDA will be chosen as the final model.

In [74]:
```
Xtest <- read_table('C:/Datasets/UCI HAR Dataset/test/X_test.txt', col_names=FALSE)
ytest <- read.csv('C:/Datasets/UCI HAR Dataset/test/y_test.txt', header = FALSE)
ytest[,1] <- factor(ytest[,1])
dftest <- as.data.frame(Xtest)
dftest$y <- ytest
```

...

In [78]:
```
ypred <- factor(predict(LDAmodel,Xtest))
confusionMatrix(ypred,ytest[,1])
```

```
Confusion Matrix and Statistics

          Reference
Prediction   1   2   3   4   5   6
         1 490  11   1   0   0   0
         2   6 460  14   1   0   0
         3   0   0 405   0   0   0
         4   0   0   0 434  22   0
         5   0   0   0  56 510   0
         6   0   0   0   0   0 537

Overall Statistics

               Accuracy : 0.9623
                 95% CI : (0.9548, 0.9689)
    No Information Rate : 0.1822
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.9547
 Mcnemar's Test P-Value : NA

Statistics by Class:

                     Class: 1 Class: 2 Class: 3 Class: 4 Class: 5 Class: 6
Sensitivity            0.9879   0.9766   0.9643   0.8839   0.9586   1.0000
Specificity            0.9951   0.9915   1.0000   0.9910   0.9768   1.0000
Pos Pred Value         0.9761   0.9563   1.0000   0.9518   0.9011   1.0000
Neg Pred Value         0.9975   0.9955   0.9941   0.9771   0.9908   1.0000
Prevalence             0.1683   0.1598   0.1425   0.1666   0.1805   0.1822
Detection Rate         0.1663   0.1561   0.1374   0.1473   0.1731   0.1822
Detection Prevalence   0.1703   0.1632   0.1374   0.1547   0.1921   0.1822
Balanced Accuracy      0.9915   0.9841   0.9821   0.9375   0.9677   1.0000
```

The LDA model is fairly accurate. Most of the confusion is between walking/walking upstairs/wallking downstairs and between standing and sitting. This is understandable because walking is similar to walking on stairs and because the orientation of the waist is similar when standing and sitting.