HarvardX: PH125.9x

Data Science: Capstone

Sergio Grueso Sáez
16/10/2019

Contents:

INTRODUCTION

Project Presentation:

Dementia is a neurodegenerative disease and according to OMS it affects nearly 7.7 million people every year, having a high prevalence between elderly. Approximately 60% of people with dementia have Alzheimer.

There is no cure for Alzheimer but the devastating effects can be slowed down. By identifying people with higher risk it is possible to extend the years of their life by slowing the neuronal degeneration. That can be done by introducing them in new habits that involve healthier diet, exercise regularly and cognitive training.

Can Alzheimer be predicted? The aim of this project is to answer this question by creating a model that based on a collection of variables predicts an outcome of whether a person with different values on those variables is considered demented or non-demented. This project uses the database provided by the Open Access Series of Imaging Studies (OASIS). OASIS is made available by the Washington University Alzheimer's Disease Research Center, Dr. Randy Buckner at the Howard Hughes Medical Institute (HHMI) at Harvard University, the Neuroinformatics Research Group (NRG) at Washington University School of Medicine, and the Biomedical Informatics Research Network (BIRN).

Dataset:

The dataset consists on longitudinal collection of 150 subjects aged 60 to 96. Each subject was scanned on two or more visits, separated by at least one year for a total of 373 imaging sessions. For each subject, 3 or 4 individual T1-weighted MRI scans obtained in single scan sessions are included (for more information visit https://www.kaggle.com/ jboy sen/mri-and-alzheimers). First all the libraries and the longitudinal dataset that will be used are downloaded. All can be found in https://github.com/sgrueso/MRI.git.

```
if(!require(tidyverse)) install.packages("tidyverse")
if(!require(caret)) install.packages("caret")
if(!require(dslabs)) install.packages("dslabs")
```

```
if(!require(dbplyr)) install.packages("dbplyr")
if(!require(ggplot2)) install.packages("ggplot2")
if(!require(Rborist)) install.packages("Rborist")
if(!require(randomForest)) install.packages("randomForest")

longitudinal <- read.csv("../MRI/oasis_longitudinal.csv", header = TRUE,
stringsAsFactors = FALSE)
```

Here is a summary of the longitudinal dataset and a description of the variables that may have confusing names:

- EDUC: years of education
- SES: Socio-economic state
- ASF: computed scaling factor that transforms native-space brain and skull to the atlas target (i.e., the determinant of the transform matrix)
- eTIV: estimated total intracranial volume (in cm3). Is the automated estimate of total intracranial volume in native space derived from the ASF
- MMSE: Mini-Mental State Exam. Is a test used to identify dementia
- CDR: Also a test to identify dementia
- nWBV: Normalized whole-brain volume (%) Automated tissue segmentation based estimate of brain volume (gray-plus white-matter).

```
summary(longitudinal)
```

```
##   Subject.ID           MRI.ID             Group              Visit
##  Length:373         Length:373         Length:373         Min.   :1.000
##  Class :character   Class :character   Class :character   1st Qu.:1.000
##  Mode  :character   Mode  :character   Mode  :character   Median :2.000
##                                                           Mean   :1.882
##                                                           3rd Qu.:2.000
##                                                           Max.   :5.000
##
##     MR.Delay          M.F                Hand                Age
##  Min.   :   0.0   Length:373         Length:373         Min.   :60.00
##  1st Qu.:   0.0   Class :character   Class :character   1st Qu.:71.00
##  Median : 552.0   Mode  :character   Mode  :character   Median :77.00
##  Mean   : 595.1                                        Mean   :77.01
##  3rd Qu.: 873.0                                        3rd Qu.:82.00
##  Max.   :2639.0                                        Max.   :98.00
##
##       EDUC            SES             MMSE            CDR
##  Min.   : 6.0   Min.   :1.00   Min.   : 4.00   Min.   :0.0000
```

```
##  1st Qu.:12.0   1st Qu.:2.00   1st Qu.:27.00   1st Qu.:0.0000
##  Median :15.0   Median :2.00   Median :29.00   Median :0.0000
##  Mean   :14.6   Mean   :2.46   Mean   :27.34   Mean   :0.2909
##  3rd Qu.:16.0   3rd Qu.:3.00   3rd Qu.:30.00   3rd Qu.:0.5000
##  Max.   :23.0   Max.   :5.00   Max.   :30.00   Max.   :2.0000
##                 NA's   :19     NA's   :2
##       eTIV           nWBV            ASF
##  Min.   :1106   Min.   :0.6440   Min.   :0.876
##  1st Qu.:1357   1st Qu.:0.7000   1st Qu.:1.099
##  Median :1470   Median :0.7290   Median :1.194
##  Mean   :1488   Mean   :0.7296   Mean   :1.195
##  3rd Qu.:1597   3rd Qu.:0.7560   3rd Qu.:1.293
##  Max.   :2004   Max.   :0.8370   Max.   :1.587
##
```

METHOD

Data Analysis

Analysis

The analysis is started by removing Hand variable. As said in the description every subject is right handed an therefore it is not relevant to keep this variable.

```
longitudinal$Hand <- NULL
```

M.F variable name is changed to “Sex” to be more descriptive

```
names(longitudinal)[names(longitudinal) == "M.F"] <- "Sex"
```

Quick easy search for null values

```
sum(is.na(longitudinal))
```

```
## [1] 21
```

19 observations have one missing value from SES variable and 2 observations have missing values from SES and MMSE column. Subjects with 2 missing values are dropped and any missing value is replaced by the median of that variable. Median is used instead of mean because SES is a variable with integer values from 1 to 5 that indicates socioeconomic status and we do not want decimals values.

```
ind <- which(is.na(longitudinal$MMSE))
longitudinal <- longitudinal[-ind, ]

ind <- which(is.na(longitudinal$SES))
longitudinal$SES[ind] <- median(longitudinal$SES[-ind])
```

Group variable is the variable we want to predict. “Demented” is changed to 1 and “non-demented” to 0. This is in order to facilitate predictions, if the algorithm predicts anything greater than 0.5 (>0.5) the prediction will be “1” meaning that the subject has or is in risk of dementia.

```
longitudinal$Group <- ifelse(longitudinal$Group=="Demented", 1, 0)
longitudinal$Group <- as.factor(longitudinal$Group)
```

Correlations between variables are displayed:

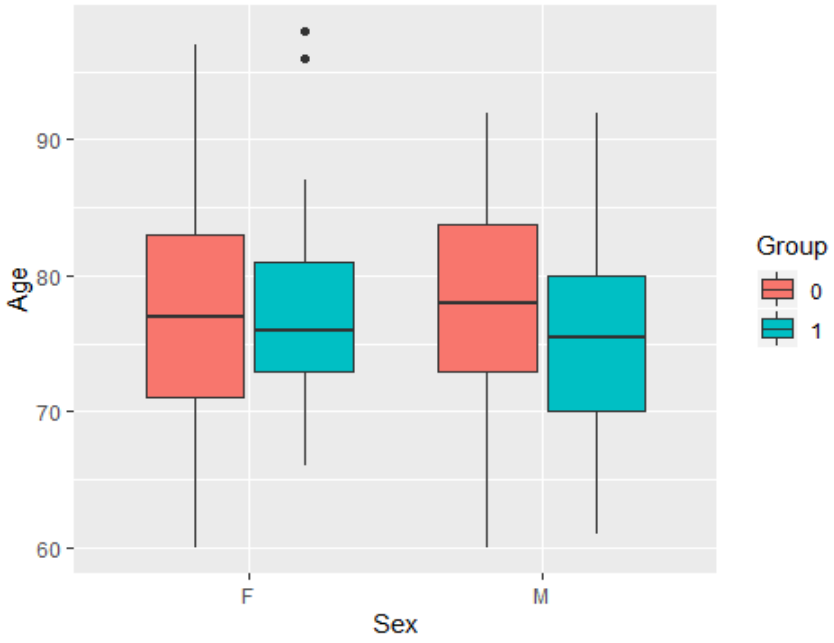
```
longitudinal %>% select(MR.Delay, Age, EDUC, MMSE, CDR, eTIV, nWBV, ASF)
%>% cor()
```

```
##             MR.Delay         Age        EDUC        MMSE         CDR
## MR.Delay  1.00000000  0.20550115  0.05354613  0.06584393 -0.06729902
## Age       0.20550115  1.00000000 -0.02859911  0.05561218 -0.02514668
## EDUC      0.05354613 -0.02859911  1.00000000  0.19488436 -0.14561447
## MMSE      0.06584393  0.05561218  0.19488436  1.00000000 -0.68651948
## CDR      -0.06729902 -0.02514668 -0.14561447 -0.68651948  1.00000000
## eTIV      0.12451322  0.04143405  0.25088721 -0.03208381  0.04245037
## nWBV     -0.10582013 -0.51825368 -0.01118340  0.34191241 -0.35047275
## ASF      -0.12997991 -0.03395663 -0.23448371  0.04005216 -0.05376694
##                eTIV       nWBV         ASF
## MR.Delay  0.12451322 -0.1058201 -0.12997991
## Age       0.04143405 -0.5182537 -0.03395663
## EDUC      0.25088721 -0.0111834 -0.23448371
## MMSE     -0.03208381  0.3419124  0.04005216
## CDR       0.04245037 -0.3504728 -0.05376694
## eTIV      1.00000000 -0.2099807 -0.98912256
## nWBV     -0.20998073  1.0000000  0.21391424
## ASF      -0.98912256  0.2139142  1.00000000
```

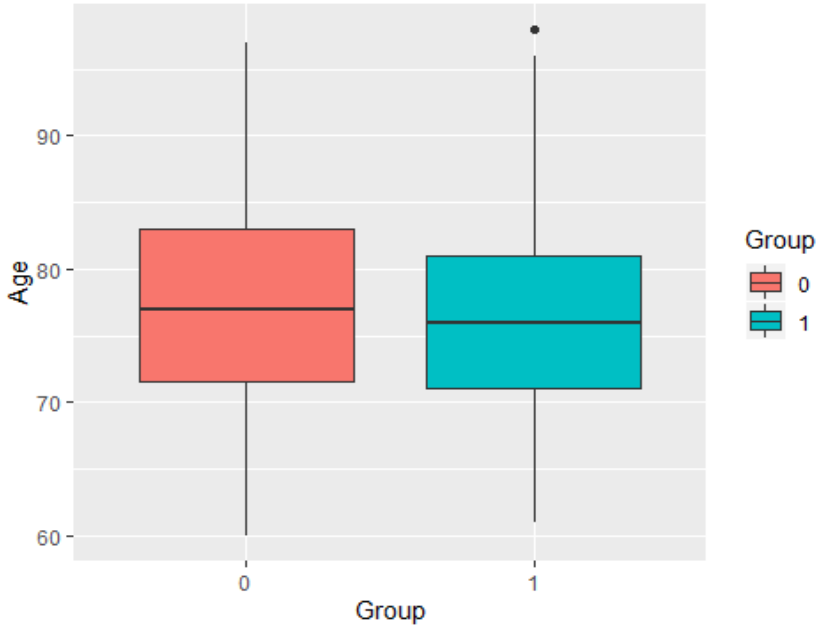
Data visualization:

Data visualization analysis can be started by observing Age and Sex distribution separated by Group variable:

```
longitudinal %>% ggplot(aes(Sex, Age, fill= Group)) + geom_boxplot()
```

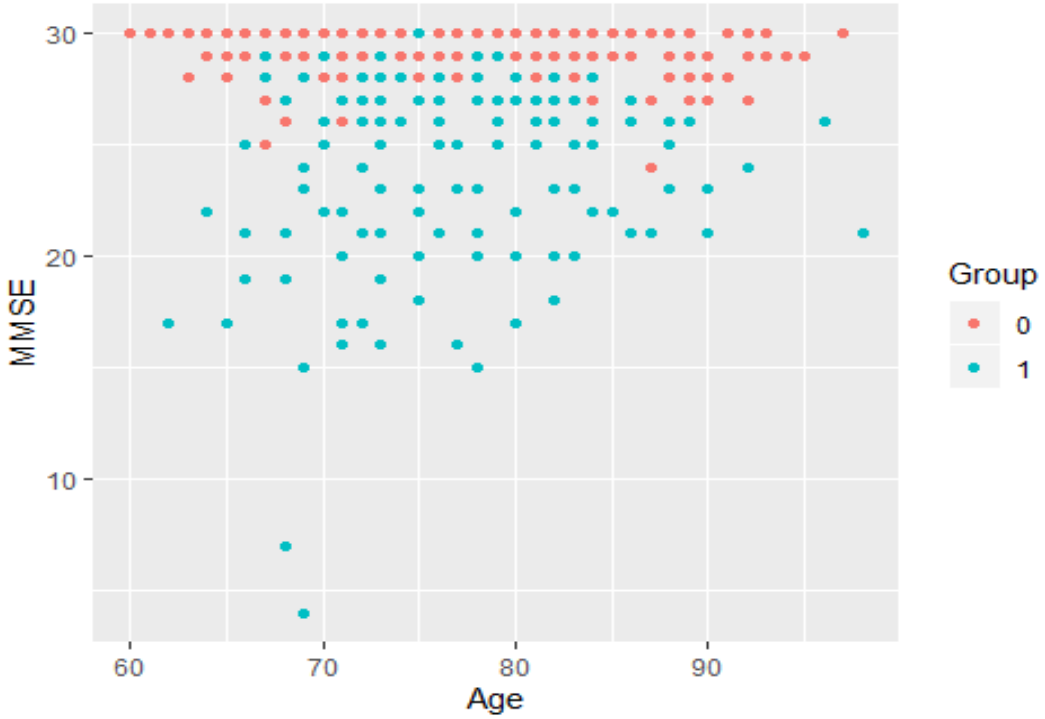



```
longitudinal %>% ggplot(aes(Group, Age, fill = Group)) + geom_boxplot()
```



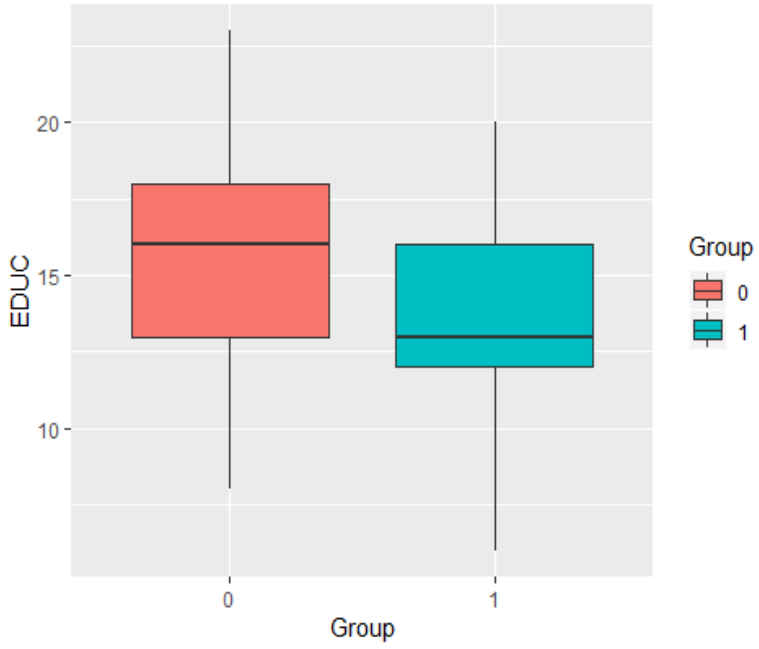
Results on the MMSE test plotted against Age with the colour depending on the Group:

```
longitudinal %>% ggplot(aes(Age, MMSE, col = Group)) + geom_point()
```



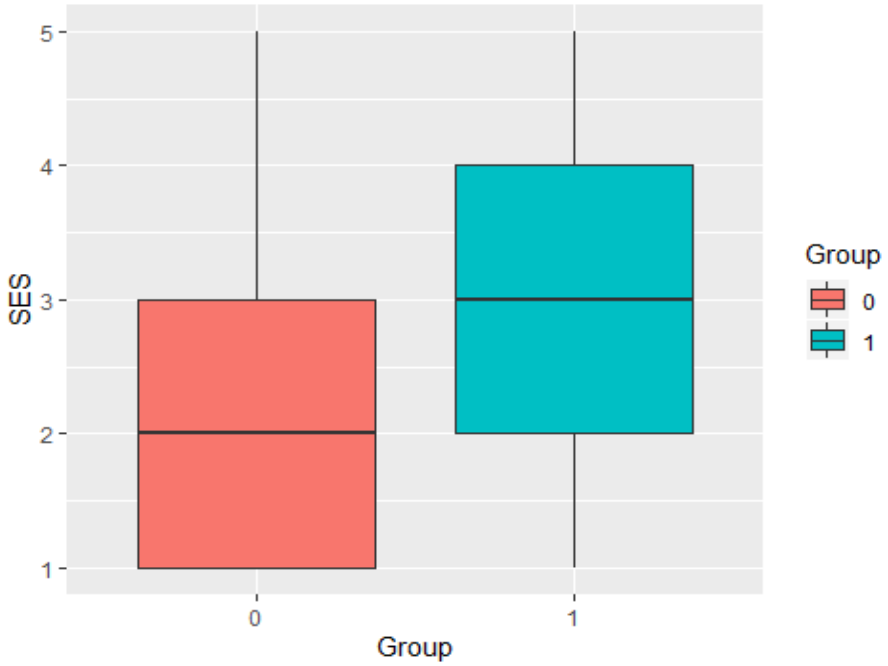
Comparison between Education and Group is made:

```
longitudinal %>% ggplot(aes(Group, EDUC, fill = Group)) + geom_boxplot()
```



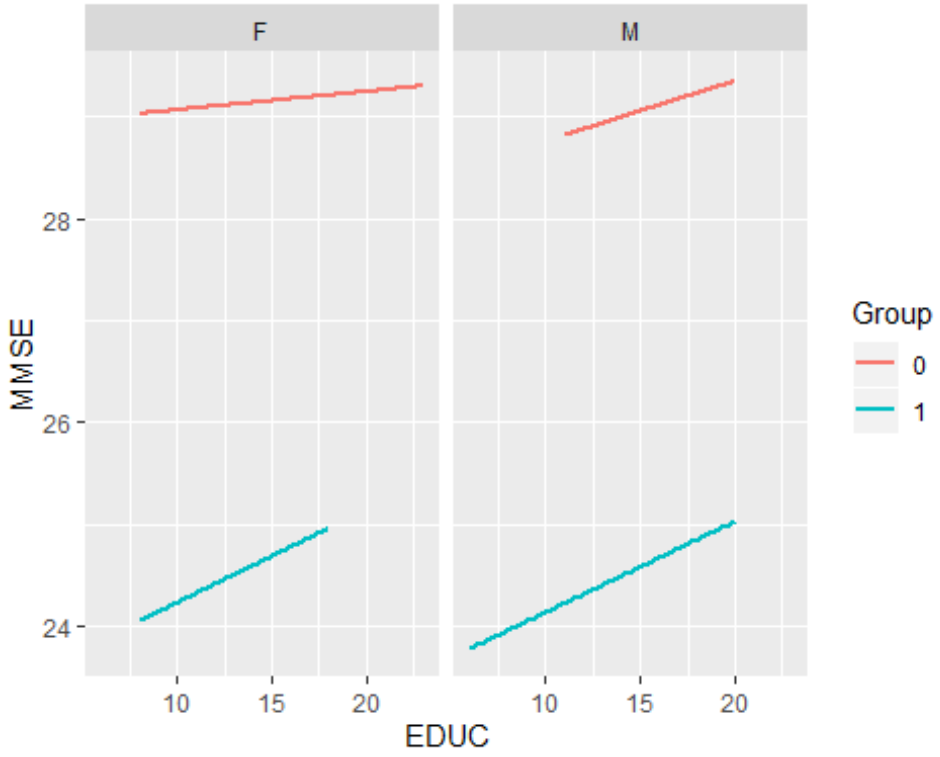
Comparison between SES and Group is also made:

```
longitudinal %>% ggplot(aes(Group, SES, fill = Group)) + geom_boxplot()
```



As previous plots have shown, Education and MMSE have a negative correlation with Dementia. Therefore a final plot of MMSE against education is suggested with the expectation to find a positive correlation between these 2 variables, and also, independently of the value in the variables Sex and Group.

```
longitudinal %>% ggplot(aes(EDUC, MMSE, col = Group)) +
                 geom_smooth(method = "lm", se = FALSE) +
                 facet_wrap(~Sex)
```



In summary, patients with Alzheimer (dementia) compared with non-demented subjects; tend to be less educated, in a lower social economic status and with lower punctuations in MMSE.

Age is normally the principal risk factor but in this study the mean age of the group with dementia is actually lower. That's probably because the expectancy of life in people with dementia is much lower. Therefore we do not see a lot of people arriving to mid or late 90s when suffering Alzheimer's disease.

Modelling approach

Train and test set partitions are made:

```
set.seed(1, sample.kind="Rounding")

test_index <- createDataPartition(y = longitudinal$Group, times = 1, p =
0.2, list = FALSE)
train_set <- longitudinal[-test_index,]
test_set  <- longitudinal[test_index,]
```

Naive Baseline Model:

The obvious way of predicting dementia seems to be MMSE. According to MMSE test manual, a value lower or equal than 27 is considered low enough to predict dementia. The model has a 0.84 accuracy with only one parameter. But can I improve the accuracy if the model includes more parameters like Age, Education or Social Economic Status?

```
y_hat <- ifelse(test_set$MMSE <= 27, 1, 0)

mean(y_hat == test_set$Group)

## [1] 0.84
```

GLM

MMSE is a test and includes scales that have into account age and level of education of the subject. That is represented in our study in the variables Age and EDUC so the next model must include both variables. A General Regression Model is used to train our model.

```
set.seed(1, sample.kind="Rounding")

model1 <- train(Group ~ Age + EDUC + MMSE,
                method = "glm",
                data = train_set)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

confusionMatrix(predict(model1, test_set), test_set$Group)$overall["Accur
acy"]

##  Accuracy
## 0.8666667
```

The model is improved by adding both variables. Can more variables improve more the accuracy? The variables selected are the ones that do not have very high correlation between each other and that are relevant to consider if a subject has dementia (for example Subject.ID or MRI.ID are not selected). Indeed, more variables improve the model.

```r
set.seed(1, sample.kind="Rounding")

model2 <- train(Group ~ Age + EDUC + MMSE + SES + eTIV + nWBV,
                method = "glm",
                data = train_set)

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

confusionMatrix(predict(model2, test_set), test_set$Group)$overall["Accur
acy"]

## Accuracy
##     0.88
```

Random Forest

Random forest seems more useful with 6 predictors and therefore is used to train the model. We get a better accuracy than just MMSE model and glm and variable importance is shown.

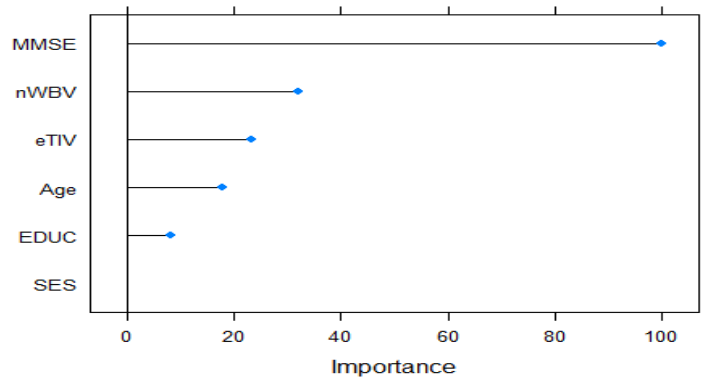
```r
set.seed(1, sample.kind="Rounding")

model3 <- train(Group ~ Age + EDUC + MMSE + SES + eTIV + nWBV,
                method = "rf",
                data = train_set)

confusionMatrix(predict(model3, test_set), test_set$Group)$overall["Accur
acy"]

##  Accuracy
## 0.8933333

plot(varImp(model3))
```

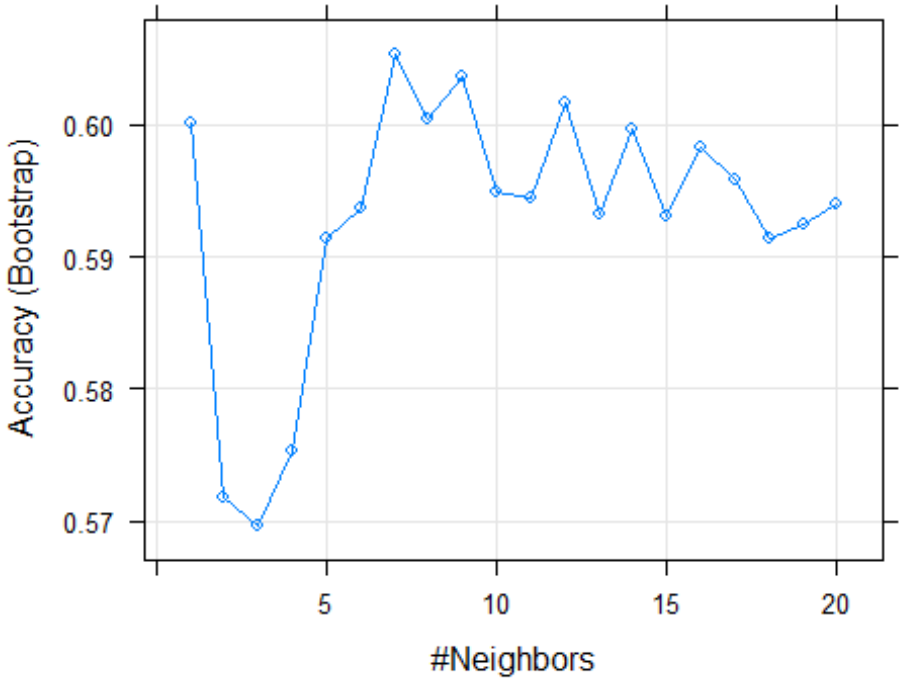


KNN

Another algorithm is KNN but the accuracy is only 0.627 with k = 7 as best “k”.

```
set.seed(1, sample.kind="Rounding")

model4 <- train(Group ~ MR.Delay + Age + EDUC + MMSE + SES + eTIV + nWBV,
                method = "knn",
                tuneGrid = data.frame(k = seq(1, 20, 1)),
                data = train_set)
plot(model4)
```




```
confusionMatrix(predict(model4, test_set), test_set$Group)$overall["Accur
acy"]

##  Accuracy
## 0.6266667
```

RESULTS

MODEL NAME	ACCURACY
KNN	0.627
MMSE	0.840
GLM (MMSE + Age + EDUC)	0.867
GLM (6 Variables)	0.880
Random Forest	0.893

CONCLUSIONS

With Random Forest nearly 9/10 subjects with Alzheimer are spotted. MMSE seems to be the best variable to predict Alzheimer but the accuracy of the MMSE test is improved by adding to the model MR.Delay, SES, eTIV and nWBV variables.

Alzheimer and any type of dementia are complex neurodegenerative diseases that involve more than one factor. Algorithms can provide valuable information, especially for detecting high risk patients, but diagnoses must be based on clinical expert opinions.