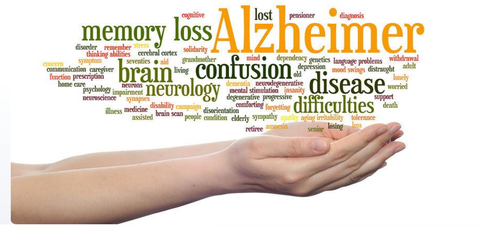
STATISTICAL ANALYSIS TO RECOGNIZE EARLY SIGNS OF ALZHEIMER'S

MAJOR SAMPSON(M20180743), FARID ULLAH(M20180216) & SHAWKAWTUL AZIZ(M20181035)

Department of DATA SCIENCE & ADVANCED ANALYTICS, NOVA IMS



Introduction

Alzheimer's disease (AD) is an irreversible, progressive brain disease that slowly destroys memory and thinking skills, and eventually even the ability to carry out the simplest tasks.

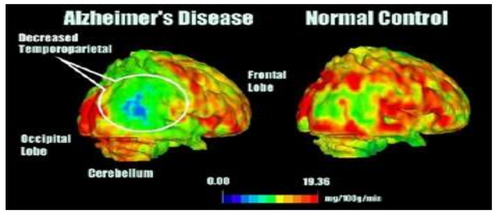
Alzheimer's disease is the most common cause of dementia among the elderly. AD's first symptoms usually manifest after the age of 60.

- Dementia involves the loss of cognitive functioning such as thinking, remembering, and reasoning, to such an extent that it interferes with a person's daily life and activities.

It is common known fact that aging is the primary factor that results in Alzheimer disease. However, aging is not the only factor that can cause the disease. There are other factors that we can examine to help developing a better understanding of Alzheimer disease in the work.

Statistics:

- AD currently affects 12million people worldwide (4.5 million in America).
- This number is likely to triple with the aging of the baby-boom generation by 2050.
- The prevalence rate for AD is about 7% for individuals aged 65 or more, and the risk doubles every 5 years after age 65.



Objective

The goal of the project is to observe how different features could potentially cause dementia and use the features of data to train the supervised learning model to predict early stages of Alzheimer disease.

From the dataset, the values of the Clinical Dementia Rating (CDR) will be used to determine whether the subjects have Alzheimer disease or not. The CDR feature has the value of 0.5, 1.0, 2.0. The value 0 means non-demented; 0.5 means very mild dementia; 1 means mild dementia; 2 means moderate dementia. In other words, if a subject who has the CDR value greater than zero, then the subject has dementia.

Though the level of dementia is not the major interest of the project, So the CDR values will be converted to binary values which are 0 and 1. The value 0 in CDR means the subject is not demented; the value 1 in CDR means the subject is demented.

The machine learning algorithm Lasso regression , k-nearest neighbors algorithm (k-NN), and Supported Vector Machine(SVM) will be implemented based on the CDR values. These algorithms can be used to provide the predictions of diagnosing dementia. We will choose the machine learning algorithm that gives me the higher Accuracy.

Data EXPLORATION

The data was obtained from Open Access Series of Imaging Studies: http://www.oasis-brains.org/ and hosted on https://www.kaggle.com/jboysen/mri-and-alzheimers .

It consists of a longitudinal collection of 150 subjects aged 60 to 96 years old. 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. The subjects are all right-handed and include both men and women.

72 of the subjects were characterized as non-demented throughout the study. 64 of the included subjects were characterized as demented at the time of their initial visits and remained so for subsequent scans, including 51 individuals with mild to moderate Alzheimer's disease. Another 14 subjects were characterized as nondemented at the time of their initial visit and were subsequently characterized as demented at a later visit.

Variable Name:	Variable Description:
ID	Identification
M/F	Gender (M if Male, F if Female)
Hand	Handedness
Age	Age in years
EDUC	Years of education
SES	Socioeconomic Status
MMSE	Mini Mental State Examination
CDR	Clinical Dementia Rating
eTIV	Estimated Total Intracranial Volume
nWBV	Normalize Whole Brain Volume
ASF	Atlas Scaling Factor
Delay	Delay

Summary/Visualisation:

After generating various descriptions of the data. We find there to be 371 observations and 15 columns.

***Table 1 & 2**: we can conclude that the average age of the subjects are 77 years old, the estimated total intracranial volume is 1488, Clinical Dementia Rating is 0.29 and the Normalized whole brain volume is 0.79.*

```
> summary(Data)
   Subject.ID          MRI.ID             Group              Visit           MR.Delay
 Length:373         Length:373         Length:373         Min.   :1.000   Min.   :   0.0
 Class :character   Class :character   Class :character   1st Qu.:1.000   1st Qu.:   0.0
 Mode  :character   Mode  :character   Mode  :character   Median :2.000   Median : 552.0
                                                          Mean   :1.882   Mean   : 595.1
                                                          3rd Qu.:2.000   3rd Qu.: 873.0
                                                          Max.   :5.000   Max.   :2639.0

     M.F                Hand                Age             EDUC            SES
 Length:373         Length:373         Min.   :60.00   Min.   : 6.0   Min.   :1.00
 Class :character   Class :character   1st Qu.:71.00   1st Qu.:12.0   1st Qu.:2.00
 Mode  :character   Mode  :character   Median :77.00   Median :15.0   Median :2.00
                                       Mean   :77.01   Mean   :14.6   Mean   :2.46
                                       3rd Qu.:82.00   3rd Qu.:16.0   3rd Qu.:3.00
                                       Max.   :98.00   Max.   :23.0   Max.   :5.00
                                                                      NA's   :19
     MMSE            CDR              eTIV            nWBV            ASF
 Min.   : 4.00   Min.   :0.0000   Min.   :1106   Min.   :0.6440   Min.   :0.876
 1st Qu.:27.00   1st Qu.:0.0000   1st Qu.:1357   1st Qu.:0.7000   1st Qu.:1.099
 Median :29.00   Median :0.0000   Median :1470   Median :0.7290   Median :1.194
 Mean   :27.34   Mean   :0.2909   Mean   :1488   Mean   :0.7296   Mean   :1.195
 3rd Qu.:30.00   3rd Qu.:0.5000   3rd Qu.:1597   3rd Qu.:0.7560   3rd Qu.:1.293
 Max.   :30.00   Max.   :2.0000   Max.   :2004   Max.   :0.8370   Max.   :1.587
 NA's   :2
```

***Figure** 3: From the correlation graph of the relationship of the each feature, we can notice that eTIV and ASF have certain degree of correlation which is not surprising since the value of total intracranial volume is estimated by the atlas scaling factor. The Age and the nWBV values are also showing some degree of correlation. It clearly shows that when the age increases, the normalized whole brain volume(nWBV) decreases accordingly.*

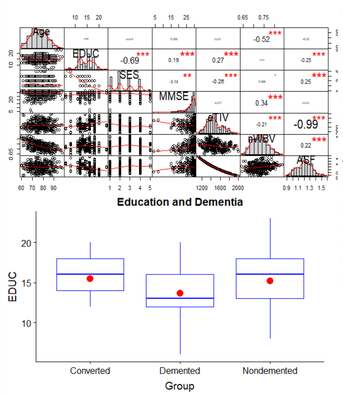


Figure 4. Educ vs Demantia by Group Graph:

it is indicated that the more education one has, the less likely dementia occurs. There appears to be a difference in median years of education between Demented and Nondemented. The nondemented group has a higher average years of education (red dot) and higher median years of education than Demented.

Methodology

Converting the CDR values to binary values so that the machine learning algorithm can "learn" to classify the result. If CDR > 0, then convert the values to 1. Otherwise, set it to 0.

```
#Factor Variables, by creating a dummy variable
# converted variables to factor and collapse CDR into demented and non demented
Data$Group<- as.factor(Data$Group)
Data$M.F<- as.factor(Data$M.F)
Data$CDR<- as.factor(ifelse(Data$CDR==.5, 1, Data$CDR))
```

The Subject.ID, MRI.ID, MR.Delay, and Hand features are useless in our prediction since, and we dropped them. Also, SES & MMSE of the subjects have missing information on the data. If we delete all the subjects who have missing data, the data set will lose its quality. We imputed the missing values with median to fill in the empty columns.

```
#Imputation of the Missing Values
sort(apply(Data, 2, function(x)sum(is.na(x))), decreasing = TRUE)

#Inputing median of the feature MMSE & SES
Data$MMSE<- ifelse(is.na(Data$MMSE), median(Data$MMSE, na.rm = TRUE), Data$MMSE)
```

Scale all numeric predictors. Some models benefit from having variables on the same scale. We could use the preProcess function in caret package, however since there are so few variables, we just do the center and scaling manually.

```
#center and scale
#Next I center and scale all numeric predictors.
#Some models benefit from having variables on the same scale
numeric_variables$visit<- scale(numeric_variables$visit)
numeric_variables$Age<- scale(numeric_variables$Age)
numeric_variables$EDUC<- scale(numeric_variables$EDUC)
numeric_variables$MMSE<- scale(numeric_variables$MMSE)
numeric_variables$nWBV<- scale(numeric_variables$nWBV)
numeric_variables$ASF<- scale(numeric_variables$ASF)
```

Multicolinearity : It is important to check if any predictors are correlated with each other to avoid. If there are, predictive performance may suffer and numerical instability is introduced. The caret package has a function findCorrelation which does a great job of dealing with this issue.

```
#Numeric Variables, Multicolinearity
numeric_variables<- Data[, sapply(Data, is.numeric)]
correlations<- cor(numeric_variables)
highcorr<- findCorrelation(correlations, cutoff = .75)
numeric_variables<- numeric_variables[,-highcorr]
```

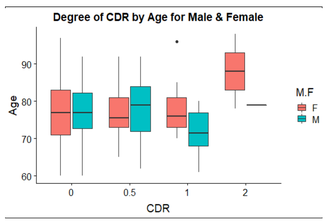


Fig 5. CDR by Age & Gender:
The frequency of women is much higher than men which further back the claim that women are more prone to AD than men: https://alz.org/blog/alz/february_2016/why_does_alzheimer_s_disease_affect_more_women_tha

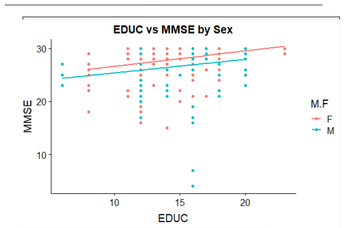


Fig 6. Educ vs MMSE:
The scatter plot with linear regression lines for Male and Female show a positive correlation between EDUC and MMSE. As suspected, it tends to be that the more years of education one has, the higher the MMSE score

Initial Implementation:

```
ROC curve variable importance

       Importance
MMSE      100.0000
nWBV       52.4295
M.F.F      34.8586
M.F.M      34.8586
EDUC       32.6727
SES        12.3204
Age         2.5732
visit       0.7786
ASF         0.0000
```

Fig 7 Splitting:
We first shuffle and split the data into training and testing set (CDR) and dropping some Group variables of no importance

```
#drop group, Dropping all 3 group variables.
train1<- as.data.frame(cbind(numeric_variables, factor_variables))

train1<-train1[,c(-8,-9,-10)]
```

Fig 8. Variable Importance:
Prior to the feature importance table above, we set the seed to make results reproducible. We use the glmnet method to fit a logistic regression model with the L1 penalty. This penalty can shrink some coefficients to 0. This conducts feature selection and helps with overfitting. Based on the coefficients, there weren't any variables dropped from the model. According to variable importance of the MMSE is by far the most important variable. Second most important is nWBV. Then, sex and EDUC are tied for third.

Implementation

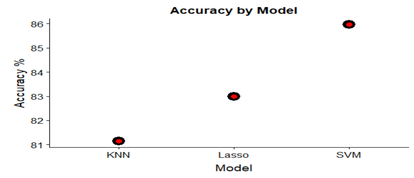
The process of implementing the three machine learning algorithms Support Vector Machine, KNN and Lasso Regression are relatively less complicated than the data processing. The only modification that we made was the change of parameters in each algorithm for Lasso: λ=.000517, SVM; Cost=8, KNN: k=5.

1. Logistic Regression with L1Penalty(Lasso): 83% and It seems that Age, EDUC, SES, and ASF have the largest coefficients.

2. KNN: The optimal tune is with k=5. This means that when predicting a new point, the five "closest" points determine what the new one will be.
With an accuracy of 81%, there is a drop in performance compared to logistic regression and support vector machine

```
#model Assessment
models<- list("Lasso" = lasso.mod, "SVM" = SVM.mod, "KNN" = knn.mod)
mod.accuracy<- sapply(models, function(x) max(x$results$Accuracy))
model<-names(which.max(mod.accuracy))
data.frame(model, mod.accuracy)%>%
  mutate(mod.accuracy=mod.accuracy*100)%>%
  ggplot(aes(x=model, y=mod.accuracy))+
  theme(plot.title = element_text(hjust = .5))+
  ylab("Accuracy %")+
  xlab("Model")+
  ggtitle("Accuracy by Model")
```

3. Support Vector Machine: Next we fit a support vector machine with accuracy of 86%. Notably better than the logistic regression model previously fitted. According to variable importance of the support vector machine, MMSE is by far the most important variable. Second most important is nWBV. Then, sex and EDUC are tied for third. Interesting to compare with the coefficients of the logistic regression at **Figure 10** below.



Conclusions & Improvements

The SVM model has the prediction accuracy of 86% which is very high. However, the time complexity is still a problem if the data set gets large. We could also see that Clinical Dementia Rating highly depends of result of Mini-Mental State Examination, while Age, Educational Level and Social-Economic Status have not great influence. Although it is important to remember that Dementia and Alzheimer's disease is complex mental issue, so we can not fully rely on Statistical & ML algorithms to make a diagnosis. But what we can do is consider that subject with specific characteristics is more likely to be diagnosed with Dementia based on information from other subjects with the same characteristics.

Furthermore, Random Forest Classifier algorithm could be useful to solve our problem. It uses averaging to improve the predictive accuracy and control overfitting. If we have a larger data set and want to avoid overfitting result, I think Random Forest Classifier would be a good choice.

References:

- http://www.oasis-brains.org/
- https://www.kaggle.com/jboysen/mri-and-alzheimers
- A Machine Learning Model to Predict the Onset of Alzheimer Disease using Potential Cerebrospinal Fluid (CSF) Biomarkers,Syed Asif Hassan et.al,IJACSA 2017.
- Brain Volume Decline in Aging,Anthony F. Fotenos et.al, ARCH NEUROL/VOL 65 (NO. 1), JAN 2008
- https://www.kaggle.com/ruslankl/dementia-prediction-w-tree-based-models/report
- Alzheimer's Disease, Hong-An Nguyen,Medicinal Chemistry,April 28, 2009.