MA335: Modelling Experimental and Observational Data

(Final Report)

Name – Harshita Goswami

Reg No – 2213557

Date – 18-06-2023

Abstract:

This complete examination of a dataset with an Alzheimer's disease focus is presented in the final report. The goal is to analyze the dataset using sophisticated statistical techniques and machine learning algorithms to find trends, correlations, and forecasting models related to the onset and progression of Alzheimer's disease. Descriptive statistics, graphical displays, clustering algorithms, logistic regression modeling, and feature selection techniques were all used in the investigation. The results help explain the connections between several factors and the diagnosis and progression of Alzheimer's disease. To improve the precision and practicality of the predictive models created in this study, additional investigation and validation are required.

CONTENTS:

INTRODUCTION-

The dataset used for this analysis contains a wealth of information such as demographic variables, clinical ratings, and neuroimaging measurements. With this comprehensive dataset, we will explore the complexities of Alzheimer's disease and shed light on potential risk factors, disease markers, and predictors that may influence diagnosis and disease progression. The primary objective of this report is to analyze the dataset using advanced statistical methods and machine learning algorithms to gain insight into the relationships between the features contained and the diagnosis of Alzheimer's disease. Through applying descriptive statistics, graphical representations, clustering algorithms, logistic regression modelling, and feature selection methods, we aim to uncover meaningful patterns, associations, and predictive models that can contribute to understanding and treating Alzheimer's disease.

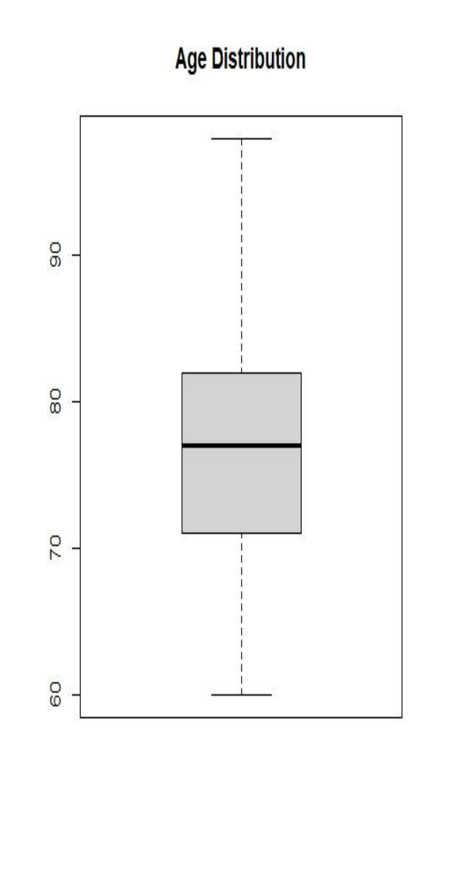
Descriptive Analysis: A Comprehensive analysis using descriptive statistics, including graphs and numerical representations-

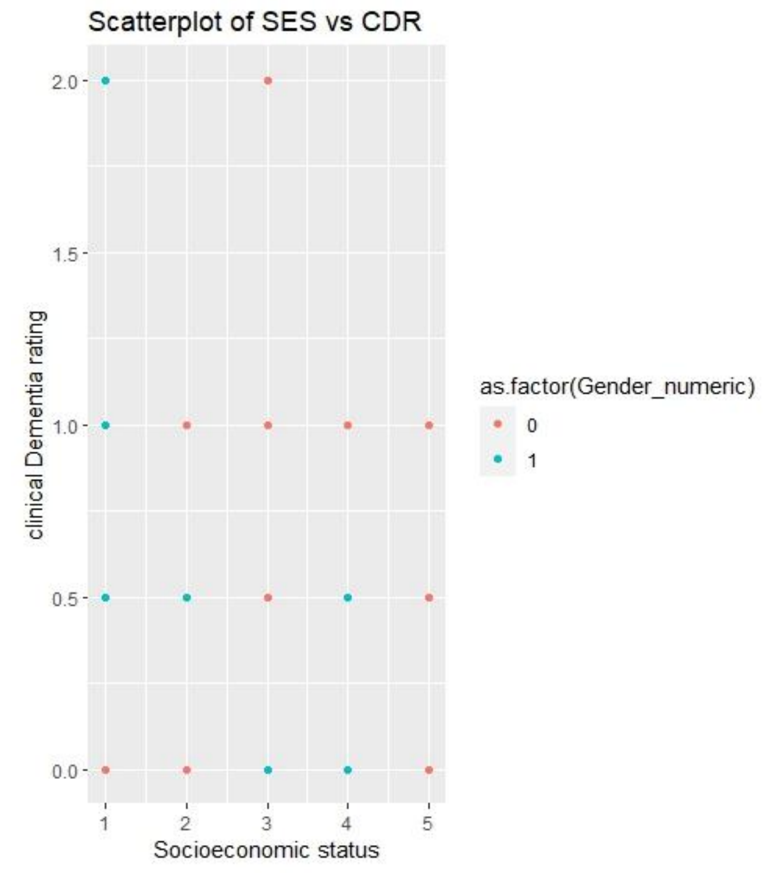
The Table below shows the mean, standard deviation, and skewness for each of the variables in our data set.

Variable	Mean	SD	Skewness
Age	76.7160883	7.8050714	0.1923390
EDUC	14.6151420	2.9268764	-0.0913250
SES	2.5457413	1.1230986	0.1541552
MMSE	27.3423181	3.8612273	-2.3051450
CDR	0.2728707	0.3821437	1.4856801
eTIV	1493.5772871	179.7190789	0.5087890
nWBV	0.7305962	0.0381020	0.1851246
ASF	1.1916057	0.1396627	0.0727897
Gender_numeric	0.4321767	0.4961618	0.2738243

From the above table it is interesting to note that the mean MMSE score of 27.3423181 suggests that, on average, the individuals in the dataset obtained a relatively high score on the Mini-Mental State Examination. This indicates a reasonably good level of cognitive function in the sample. However, the negative skewness value of -2.3051450 indicates that there may be a subset of individuals with lower MMSE scores, potentially indicating cognitive impairment or dementia. The estimated total intracranial volume, or mean eTIV of 1493.5772871, gives us information on the average brain size of the people in the dataset. It's crucial to note that the SD of 179.7190789 indicates that there may be some variation in brain sizes within the sample. And also The variable Gender_numeric mean value of 0.4321767 and skewness of 0.2738243 show that there is a little imbalance in the representation of gender in the dataset.

Graphical Analysis: for graphical analysis, we will see the boxplot of Age distribution, and a scatterplot of SES and CDR as factors of Gender.

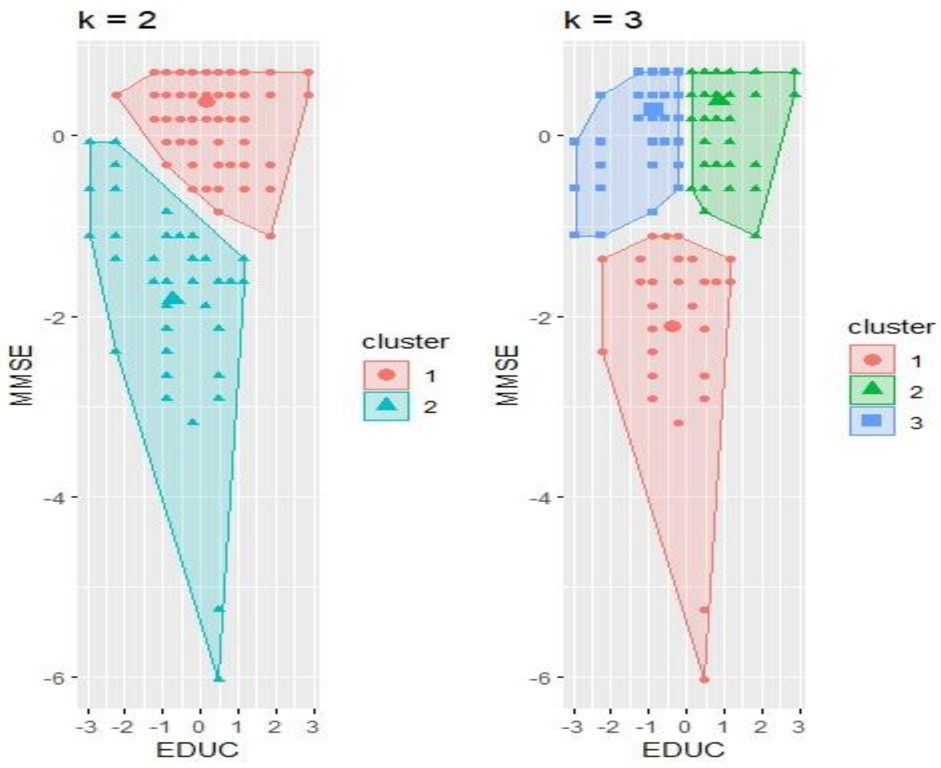




From the above Boxplot, we observe that The box ranges from approximately 70 to 83 years old, indicating that the middle 50% of the age group falls within this range. The median (50th percentile) lies roughly in the middle of the box, indicating that the distribution is roughly symmetrical. Scatterplot- in the plot below we have taken 2 variables to compare which are Socioeconomic status and Clinical Dementia Rating as a factor of gender to check the relation between these two variables on Dementia. Here we can see from the graph below that- it is interesting to note that females with a middle socioeconomic status (CDR 2.0) have moderate dementia in contrast to males with low socioeconomic status. There are more females of different socioeconomic statuses, suffering from mild dementia (CDR 1.0)

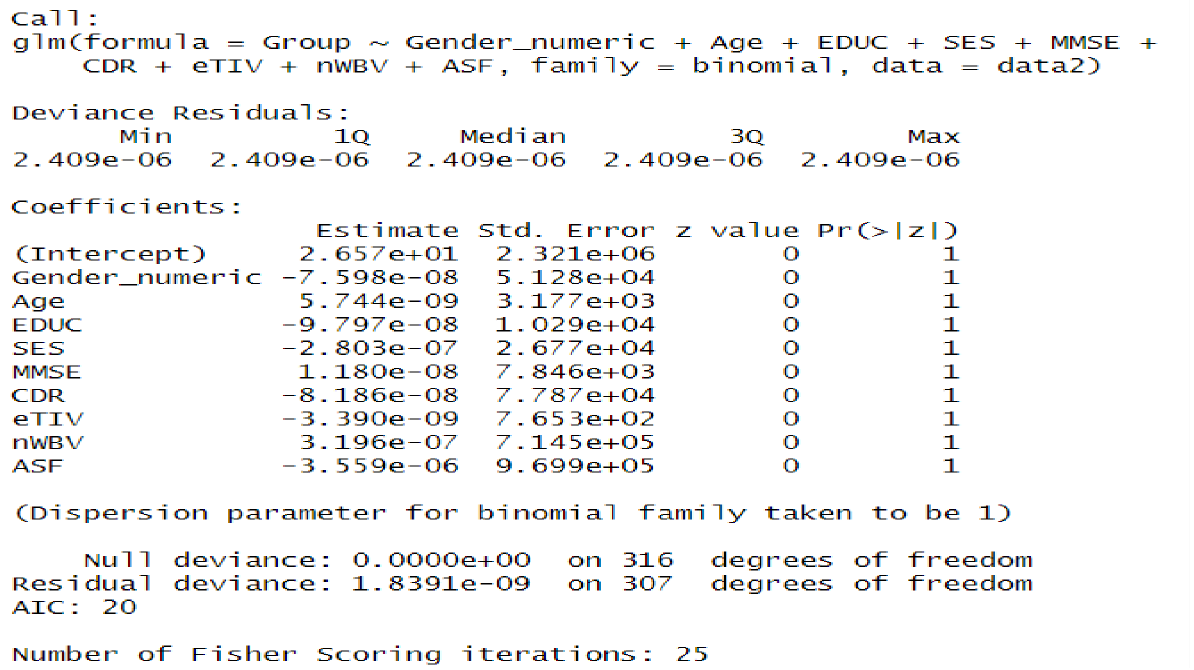
Clustering Algorithm: We will Implement clustering algorithms to explore possible groupings or subgroups in a dataset. By applying unsupervised learning techniques such as k-means clustering and hierarchical clustering, we hope to identify various patterns and clusters that may exist among variables. **K-means clustering:** On the data, we used K-means clustering with two clusters. The clusters are 262 and 55 in size, respectively, meaning that one cluster has 262 data points and the other 55. The average values of the variables inside each cluster are represented by the cluster means. "EDUC" and "MMSE" are the two variables in this situation. The average value of "EDUC" and "MMSE" for Cluster 1 is 0.1536599 and 0.3809656, respectively. The average value of "EDUC" and "MMSE" for Cluster 2 is -0.7319799 and -1.8147816 respectively. These cluster means to shed light on each cluster's properties. Compared to Cluster 2, Cluster 1, "EDUC" and "MMSE" had greater values. This shows that the data points in Cluster 1 generally had superior Mini-Mental State Examination (MMSE) test scores and higher average levels of schooling. Contrarily, Cluster 2 exhibits lower values for "EDUC" and "MMSE" than Cluster 1 does. This suggests that, on average, the data points in Cluster 2 have lower levels of education and perform worse on the MMSE test.

visualization of the clustering results:



here, two separate clustering analyses were performed: k-means with 2 clusters and k-means with 3 clusters. It is interesting to note that 2-cluster clustering shows an uneven distribution of data points between the two clusters whereas In 3 Clustering- three clusters have sizes 44, 145, and 128, respectively. Compared to the 2-cluster clustering, this indicates a more balanced distribution of data points across the clusters. Overall, the 2-cluster clustering shows a clear separation between the two clusters, while the 3-cluster clustering introduces an additional cluster that captures further variations in the data.

Logistic regression: we will fit a logistic regression model to predict the variable "Group" (i.e., Alzheimer’s diagnosis) using the remaining variables in the dataset. Using this model, we will evaluate the predictive ability of the included variables and investigate their link with the diagnosis of Alzheimer's disease. Logistic regression is a widely used statistical technique for binary classification tasks. Here, we have a summary of the model: -

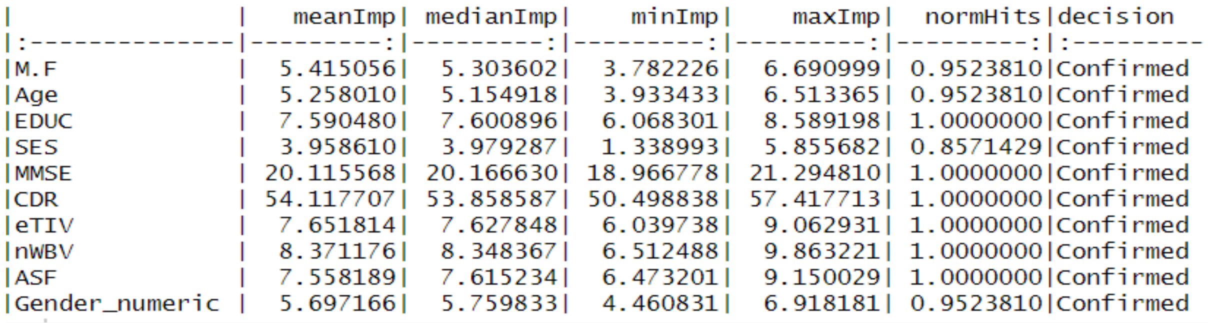


```
Call:
glm(formula = Group ~ Gender_numeric + Age + EDUC + SES + MMSE +
    CDR + eTIV + nWBV + ASF, family = binomial, data = data2)

Deviance Residuals:
      Min         1Q     Median         3Q        Max
2.409e-06  2.409e-06  2.409e-06  2.409e-06  2.409e-06

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     2.657e+01  2.321e+06       0        1
Gender_numeric -7.598e-08  5.128e+04       0        1
Age             5.744e-09  3.177e+03       0        1
EDUC           -9.797e-08  1.029e+04       0        1
SES            -2.803e-07  2.677e+04       0        1
MMSE            1.180e-08  7.846e+03       0        1
CDR            -8.186e-08  7.787e+04       0        1
eTIV           -3.390e-09  7.653e+02       0        1
nWBV            3.196e-07  7.145e+05       0        1
ASF            -3.559e-06  9.699e+05       0        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 0.0000e+00  on 316  degrees of freedom
Residual deviance: 1.8391e-09  on 307  degrees of freedom
AIC: 20

Number of Fisher Scoring iterations: 25
```

Here from the summary of the model we can observe that The residual deviance, which measures the goodness-of-fit of the model, is extremely close to zero (1.8391e-09), This indicates that the model explains a large proportion of the variability in the data and fits the observed data extremely well. The AIC value of 20 suggests that the model fits the data well with low complexity. It is important to note that the lack of a statistically significant predictor raises concerns about the model's ability to accurately predict the group variable. Further research may be needed to determine the effectiveness and usefulness of this model for predicting group categories.

Feature selection method. Here I have used Boruta feature selection, which will identify the significant attributes, plot the results, calculate attribute statistics, and display them in a table.

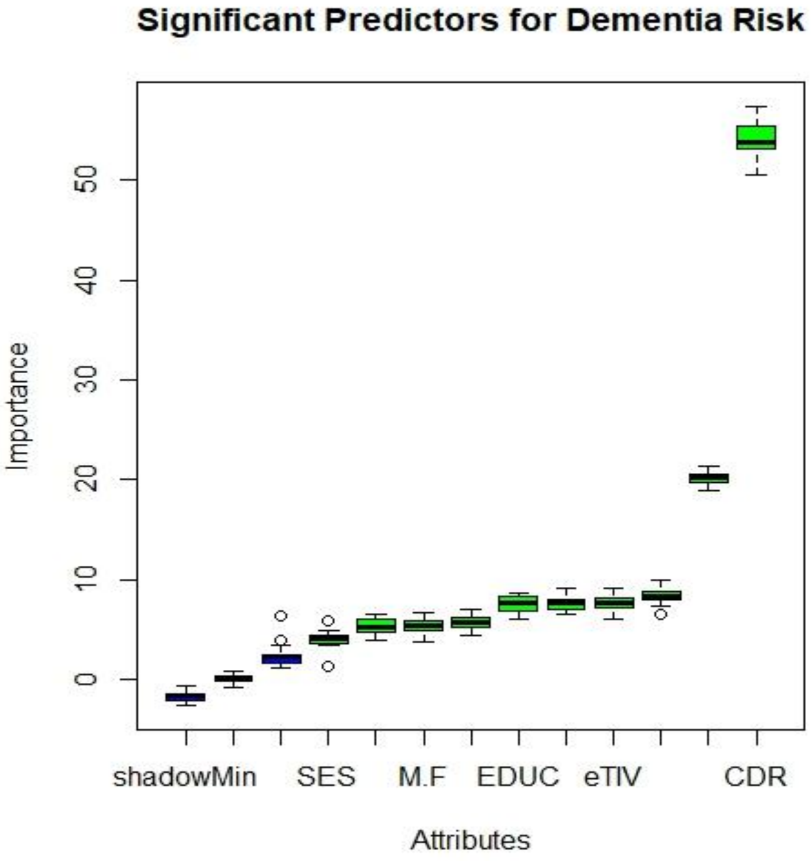
Table of Attribute statistics based on the Boruta selection results-



```
|               |   meanImp| medianImp|    minImp|    maxImp|  normHits|decision
|:--------------|---------:|---------:|---------:|---------:|---------:|:---------
|M.F            |  5.415056|  5.303602|  3.782226|  6.690999| 0.9523810|Confirmed
|Age            |  5.258010|  5.154918|  3.933433|  6.513365| 0.9523810|Confirmed
|EDUC           |  7.590480|  7.600896|  6.068301|  8.589198| 1.0000000|Confirmed
|SES            |  3.958610|  3.979287|  1.338993|  5.855682| 0.8571429|Confirmed
|MMSE           | 20.115568| 20.166630| 18.966778| 21.294810| 1.0000000|Confirmed
|CDR            | 54.117707| 53.858587| 50.498838| 57.417713| 1.0000000|Confirmed
|eTIV           |  7.651814|  7.627848|  6.039738|  9.062931| 1.0000000|Confirmed
|nWBV           |  8.371176|  8.348367|  6.512488|  9.863221| 1.0000000|Confirmed
|ASF            |  7.558189|  7.615234|  6.473201|  9.150029| 1.0000000|Confirmed
|Gender_numeric |  5.697166|  5.759833|  4.460831|  6.918181| 0.9523810|Confirmed
```

Here, we can see that the top features are- CDR (mean importance: 54.117707, median importance: 53.858587), eTIV (mean importance: 7.651814, median importance: 7.627848), nWBV (mean importance: 8.371176, median importance: 8.348367),ASF (mean importance: 7.558189, median importance: 7.615234),EDUC (mean importance: 7.590480, median importance: 7.600896) and The normHits values range from 0.8571429 to 1.0000000. This indicates that all of the features in the output are important and that the model is likely to be accurate.

The plot of the significant predictors for dementia risk using the Boruta feature selection method-



Here, in the plot, we can see that the graph shows the increasing order of importance for the variables, with CDR being the most crucial variable in relation to the outcome, followed by ETIV, EDUC, M.F, SES, and ShadowMin. The "ShadowMin" in the Boruta graph represents the minimum importance score of the shadow attribute.

CONCLUSION:

In conclusion, I can say that. The CDR feature can be used to track the progression of the disease and to predict whether a patient is at risk of developing more serious complications. Also, The BORUTA output serves as a starting point for feature selection by identifying potentially important features. It can guide you in selecting the most relevant features for model training, but further analysis and validation are necessary to assess the overall model fit and performance.

Appendix:

```
#Loading the required packages
library(knitr)
library(moments)
library(ggplot2)
library(dplyr)
library(cluster)
library(Boruta)
library(factoextra)
library(gridExtra)

# loading the data set
Data.set <- read.csv("project data.csv", header = TRUE)
# Converting M/F into numeric values
Data.set$Gender_numeric<- ifelse(Data.set$M.F == "M", 1, 0)
# Removing rows of Converted and missing values
data2 <- Data.set %>%
  filter(Group != "Converted")%>%
  na.omit()

# (1) Analyzing the Data set

# Calculate mean for each numeric variable
mean_values <- sapply(data2, function(x) ifelse(is.numeric(x), mean(x, na.rm = TRUE), NA))
# Calculate standard deviation for each numeric variable
sd_values <- sapply(data2, function(x) ifelse(is.numeric(x), sd(x, na.rm = TRUE), NA))
# Calculate skewness for each numeric variable
skewness_values <- sapply(data2, function(x) ifelse(is.numeric(x), skewness(x, na.rm = TRUE), NA))

# Create a data frame with the results
```

```
summary_df <- data.frame(
  Variable = names(data2),
  Mean = mean_values,
  SD = sd_values,
  Skewness = skewness_values
)
# Display the summary statistics in table format
kable(summary_df, format = "markdown")



# Box-plot
boxplot(data2$Age, main = "Age Distribution")

# Scatter-plot
ggplot(data2, aes(x = SES, y = CDR, color = as.factor(Gender_numeric))) +
  geom_point() +
  labs(title = "Scatterplot of SES vs CDR", x = "Socioeconomic status",y="clinical Dementia rating")

# (2) Selecting relevant variables for clustering

# Select the variables for clustering
clustering_data <- data2[, c("EDUC", "MMSE")]

# Standardize the variables
scaled_data <- scale(clustering_data)

# Perform k-means clustering for k = 2, 3, and 4
kmeans2 <- kmeans(scaled_data, centers = 2, nstart = 20)
kmeans3 <- kmeans(scaled_data, centers = 3, nstart = 20)

# Print the results of k-means clustering
print(kmeans2)
str(kmeans2)
```

```
print(kmeans3)

# Visualizing  the clustering results with k=2 and k=3
f1 <- fviz_cluster(kmeans2, geom = "point", data = clustering_data) + ggtitle("k = 2")
f2 <- fviz_cluster(kmeans3, geom = "point", data = clustering_data) + ggtitle("k = 3")

# Arrange the plots using gridExtra
grid.arrange(f1, f2, nrow = 1)

# (3) Fit logistic regression model
# Encode 'Group' variable as 0 and 1
data2$Group <- ifelse(data2$Group == "Demented", 0, 1)

# Fit logistic regression model
model <- glm(Group ~ Gender_numeric + Age + EDUC + SES + MMSE + CDR +
eTIV + nWBV + ASF, data = data2, family = binomial)

# Summary of the model
summary(model)

# (4)Perform feature selection using Boruta
# Creating the Boruta object
boruta1 <- Boruta(Group ~ ., data = data2, doTrace = 1)
# Printed significant attributes
signif <- names(boruta1$finalDecision)[boruta1$finalDecision %in% c("Confirmed",
"Tentative")]
print(signif)
# Plot results with x-axis label and increased size
plot(boruta1, xlab = "Attributes", main = "Significant Predictors for Dementia Risk",
cex.main = 1.2)
# Printed attribute statistics
attStats(boruta1)
tableB <- attStats(boruta1)
kable(tableB)
```