Predicting Alzheimer's Disease using Socioeconomic and MRI Imaging Data from Demented and Nondemented Adults

Applying Random Forest machine learning algorithm to classify Alzheimer's patients

Sejal Davla, PhD

Introduction

The **Open Access Series of Imaging Studies (OASIS)** is a project aimed at making neuroimaging data sets of the brain freely available to the scientific community. This freely avilable neuroimaging longitudinal data consists 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 on T1-weighted MRI scanner. 72 of the subjects were characterized as nondemented throughout the study. 64 of the subjects were characterized as demented at the time of their initial visits and remained so for subsequent scans. Another 14 subjects were characterized as nondemented at the time of their initial visit and were subsequently characterized as demented at a later visit.

The data contains the following attributes:

Column name	Information
Subject.ID	Unique ID of the patient
MRI.ID	Unique Id generated after conducting MRI on patient (This information combines subject ID with visit number, therefore **it was removed from the analysis**)
Group	Includes three subject categories. i) **Nondemented** (Normal), ii) **Demented** (Patients with mild to severe dementia), and iii) **Converted** (Previously Normal but developed dementia later)
Visit	Number of follow-up visit for each MRI scan
MR.Delay	The number of day between two medical visits
M.F	Gender (The column was renamed to gender for better readability)

Column name	Information
Hand	Handedness (All subjects are right-handed so **the column was removed from the analysis**)
Age	Age in years
EDUC	Years of education (This column was renamed as education)
SES	Socioeconomic status assessed by the Hollingshead Index of Social Position (**1 = highest status to 5 = lowest status**)
MMSE	Mini-Mental State Examination score (**0 = worst to 30 = best**)
CDR	[Clinical Dementia Rating](https://knightadrc.wustl.edu/professionals-clinicians/cdr-dementia-staging-instrument/#:~:text=The%20Clinical%20Dementia%20Rating%20(CDR,Affairs%2C% **(0 = no dementia, 0.5 = very mild AD, 1 = mild AD, 2 = moderate AD)**
eTIV	Estimated total intracranial volume (mm^3)
nWBV	Normalized whole-brain volume
ASF	Atlas scaling factor (unitless). Calculated by transforming native-space brain and skull to the atlas target

Loading and Understanding data

```
df <- read_csv(here("oasis_longitudinal.csv")) |>
  clean_names() |>
  select(-c(hand, mri_id)) |>
  rename(gender = m_f)

long <- data.table(df)
head(long)
```

```
   subject_id       group visit mr_delay gender age educ ses mmse cdr e_tiv
1:  OAS2_0001 Nondemented     1        0      M  87   14   2   27 0.0  1987
2:  OAS2_0001 Nondemented     2      457      M  88   14   2   30 0.0  2004
3:  OAS2_0002    Demented     1        0      M  75   12  NA   23 0.5  1678
4:  OAS2_0002    Demented     2      560      M  76   12  NA   28 0.5  1738
5:  OAS2_0002    Demented     3     1895      M  80   12  NA   22 0.5  1698
6:  OAS2_0004 Nondemented     1        0      F  88   18   3   28 0.0  1215
   n_wbv   asf
1: 0.696 0.883
```

```
2: 0.681 0.876
3: 0.736 1.046
4: 0.713 1.010
5: 0.701 1.034
6: 0.710 1.444
```

Data Exploration and Cleaning

```r
data_summary <- describe(df)
data_summary
```

```
df 

 13  Variables      373  Observations
--------------------------------------------------------------------------------
subject_id 
       n  missing distinct 
     373        0      150 

lowest : OAS2_0001 OAS2_0002 OAS2_0004 OAS2_0005 OAS2_0007
highest: OAS2_0182 OAS2_0183 OAS2_0184 OAS2_0185 OAS2_0186
--------------------------------------------------------------------------------
group 
       n  missing distinct 
     373        0        3 
                                            
Value        Converted    Demented Nondemented
Frequency           37         146         190
Proportion       0.099       0.391       0.509
--------------------------------------------------------------------------------
visit 
       n  missing distinct     Info     Mean      Gmd 
     373        0        5    0.874    1.882   0.9552 
                                        
Value          1     2     3     4     5
Frequency    150   144    58    15     6
Proportion 0.402 0.386 0.155 0.040 0.016

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
mr_delay 
       n  missing distinct     Info     Mean      Gmd      .05      .10 
     373        0      201    0.935    595.1    682.6        0        0 
     .25      .50      .75      .90      .95
       0      552      873     1561     1828

lowest :    0  182  212  248  352, highest: 2386 2400 2508 2517 2639
--------------------------------------------------------------------------------
gender
       n  missing distinct
     373        0        2

Value          F     M
Frequency    213   160
Proportion 0.571 0.429
--------------------------------------------------------------------------------
age
       n  missing distinct     Info     Mean      Gmd      .05      .10
     373        0       39    0.998    77.01    8.703     65.0     67.2
     .25      .50      .75      .90      .95
    71.0     77.0     82.0     87.8     90.0

lowest : 60 61 62 63 64, highest: 94 95 96 97 98
--------------------------------------------------------------------------------
educ
       n  missing distinct     Info     Mean      Gmd      .05      .10
     373        0       12    0.962     14.6    3.183       11       12
     .25      .50      .75      .90      .95
      12       15       16       18       18

Value          6     8    11    12    13    14    15    16    17    18    20
Frequency      3     9    11   103    27    33    17    81     9    64    13
Proportion 0.008 0.024 0.029 0.276 0.072 0.088 0.046 0.217 0.024 0.172 0.035

Value         23
Frequency      3
Proportion 0.008

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
ses
       n  missing distinct     Info     Mean      Gmd
     354       19        5    0.938     2.46    1.266

Value          1     2     3     4     5
Frequency     88   103    82    74     7
Proportion 0.249 0.291 0.232 0.209 0.020

For the frequency table, variable is rounded to the nearest 0
```

```
--------------------------------------------------------------------------------
mmse
       n  missing distinct     Info     Mean      Gmd      .05      .10
     371        2       18    0.954    27.34    3.417       20       22
     .25      .50      .75      .90      .95
      27       29       30       30       30

Value          4     7    15    16    17    18    19    20    21    22    23
Frequency      1     1     2     3     5     2     3     7    11     7    11
Proportion 0.003 0.003 0.005 0.008 0.013 0.005 0.008 0.019 0.030 0.019 0.030

Value         24    25    26    27    28    29    30
Frequency      4    12    20    32    45    91   114
Proportion 0.011 0.032 0.054 0.086 0.121 0.245 0.307

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
cdr
       n  missing distinct     Info     Mean      Gmd
     373        0        4    0.794   0.2909   0.3683

Value        0.0   0.5   1.0   2.0
Frequency    206   123    41     3
Proportion 0.552 0.330 0.110 0.008

For the frequency table, variable is rounded to the nearest 0
--------------------------------------------------------------------------------
e_tiv
       n  missing distinct     Info     Mean      Gmd      .05      .10
     373        0      286        1     1488    197.7     1234     1289
     .25      .50      .75      .90      .95
    1357     1470     1597     1731     1817

lowest : 1106 1123 1143 1151 1154, highest: 1928 1931 1957 1987 2004
--------------------------------------------------------------------------------
n_wbv
       n  missing distinct     Info     Mean      Gmd      .05      .10
     373        0      136        1   0.7296  0.04232   0.6746   0.6822
     .25      .50      .75      .90      .95
  0.7000   0.7290   0.7560   0.7796   0.7940

lowest : 0.644 0.646 0.652 0.657 0.66 , highest: 0.817 0.819 0.822 0.827 0.837
--------------------------------------------------------------------------------
asf
       n  missing distinct     Info     Mean      Gmd      .05      .10
     373        0      265        1    1.195   0.1563   0.9656   1.0134
     .25      .50      .75      .90      .95 
  1.0990   1.1940   1.2930   1.3618   1.4222 

lowest : 0.876 0.883 0.897 0.909 0.91 , highest: 1.521 1.525 1.535 1.563 1.587
--------------------------------------------------------------------------------
```

Handling missing values in the data

```
# Get a summary of missing (NA) values in the data
colSums(is.na(df))
```

```
subject_id      group      visit   mr_delay     gender        age       educ 
         0          0          0          0          0          0          0 
       ses       mmse        cdr      e_tiv      n_wbv        asf 
        19          2          0          0          0          0 
```

```
NA_rows <- df[!complete.cases(df), ]

unique(NA_rows$subject_id)
```

```
[1] "OAS2_0002" "OAS2_0007" "OAS2_0063" "OAS2_0099" "OAS2_0114" "OAS2_0160"
[7] "OAS2_0181" "OAS2_0182"
```

Out of 150 subject data, 8 subject data has NA values in the ses (socioeconomic status), mmse (mini mental examination score) columns. Because there are enough data points in the analysis, rows with missing ses and mmse values were removed from the analysis **instead of imputing mean or median values**. This strengthens data modeling without diminishing statistical power.

```
df_new <- df[complete.cases(df), ]
colSums(is.na(df_new)) # there are no NA values
```

```
subject_id      group      visit   mr_delay     gender        age       educ 
         0          0          0          0          0          0          0 
       ses       mmse        cdr      e_tiv      n_wbv        asf 
         0          0          0          0          0          0 
```

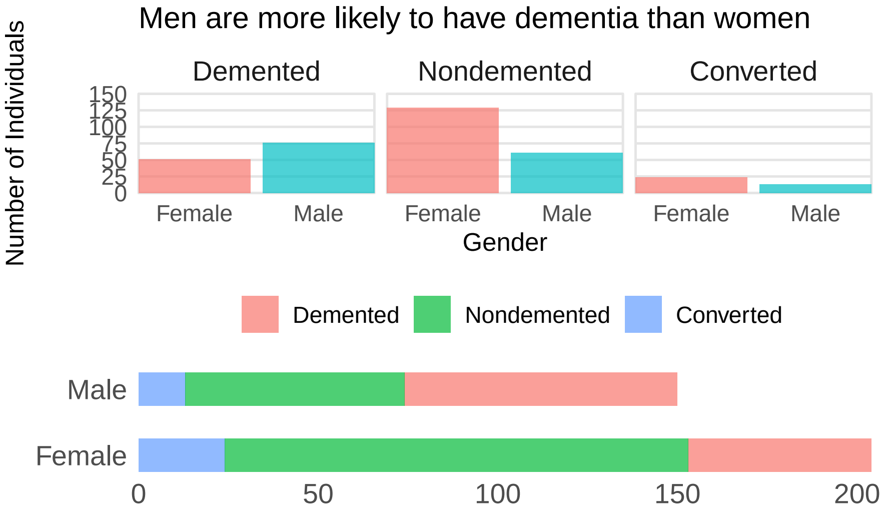
```
df_new$gender <- as.factor(df_new$gender)
df_new$group <- as.factor(df_new$group)
df_new$visit <- as.factor(df_new$visit)
df_new$ses <- as.factor(df_new$ses)
```

```
df_new$cdr <- as.factor(df_new$cdr)
```

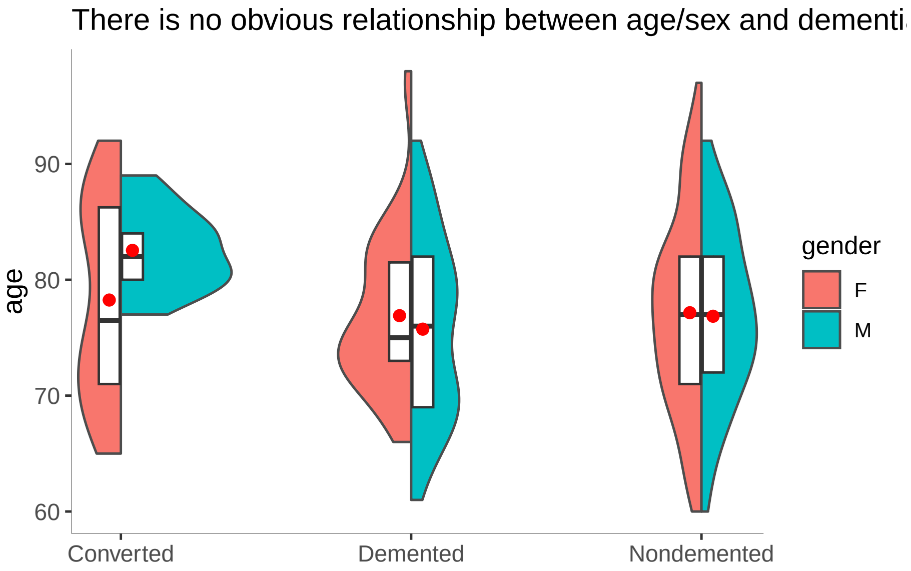
Perform Univariate and Bivariate Exploratory Data Analysis

There are two objectives for performing exploratory data analysis. First is to explore data distribution and understand if specific variables are under- or over-represented in the dataset. Second objective is to determine relationship between variables that will help make assumptions in the modeling step.

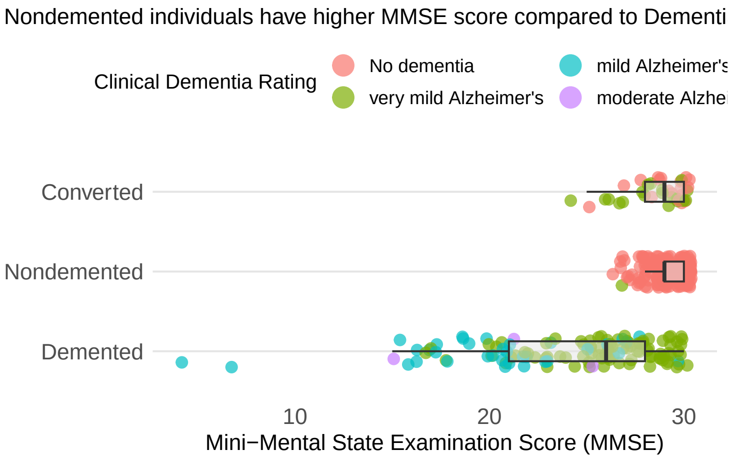
```
P1 <- df_new |>
    mutate(group = fct_relevel(group, c("Demented", "Nondemented", "Converted"))) |>
  ggplot(aes(x = gender, fill = gender)) +
  geom_bar(alpha = 0.7, width = 0.9) +
  facet_wrap(~group) +
  scale_y_continuous(limits = c(0, 150),
                     breaks = seq(0, 150, 25)) +
  scale_x_discrete(labels = c("Female", "Male")) +
  coord_cartesian(expand = FALSE,
                  clip = "off") +
  labs(x = "Gender",
       y = "Number of Individuals",
       title = "Men are more likely to have dementia than women") +
  theme_classic() +
  theme(strip.background = element_blank(),
        strip.text = element_text(size = 12),
        axis.text = element_text(size = 10),
        axis.line = element_blank(),
        axis.ticks = element_blank(),
        panel.grid.major.y = element_line(color = "grey90", size = 0.5),
        panel.background = element_rect(fill = NA, color = "grey90"),
        legend.position = "none")

P2 <- df_new |>
  group_by(gender, group) |>
  summarise(count = n()) |>
  mutate(prop = count / sum(count)) |>
  mutate(group = fct_relevel(group, c("Demented", "Nondemented", "Converted"))) |>
  ungroup() |>
  ggplot(aes(x = count, y = gender, fill = group)) +
  geom_col(width = 0.5, alpha = 0.7) +
  coord_cartesian(expand = FALSE) +
  scale_y_discrete(labels = c("Female", "Male")) +
  labs(x = "",
       y = "",
       fill = "") +
  theme_classic() +
  theme(axis.line = element_blank(),
        axis.ticks = element_blank(),
        legend.position = "top",
        axis.text = element_text(size = 12),
        legend.text = element_text(size = 10))

P1/P2
```



```
df_new |>
  select(group, gender, age) |>
  ggcraviola(craviola.width = 0.1,
             lines.col = "grey30",
             bins.quantiles = seq(0.25, 0.75, 0.25)) +
  scale_fill_manual(values = c("M" = "#00BFC4",
                                "F" = "#F8766D")) +
  labs(title = "There is no obvious relationship between age/sex and dementia diagnosis")
  theme_classic() +
  theme(axis.title.x = element_blank(),
        axis.line = element_line(size = 0.1, color = "grey30"),
        axis.text = element_text(size = 10),
        axis.title = element_text(size = 12))
```

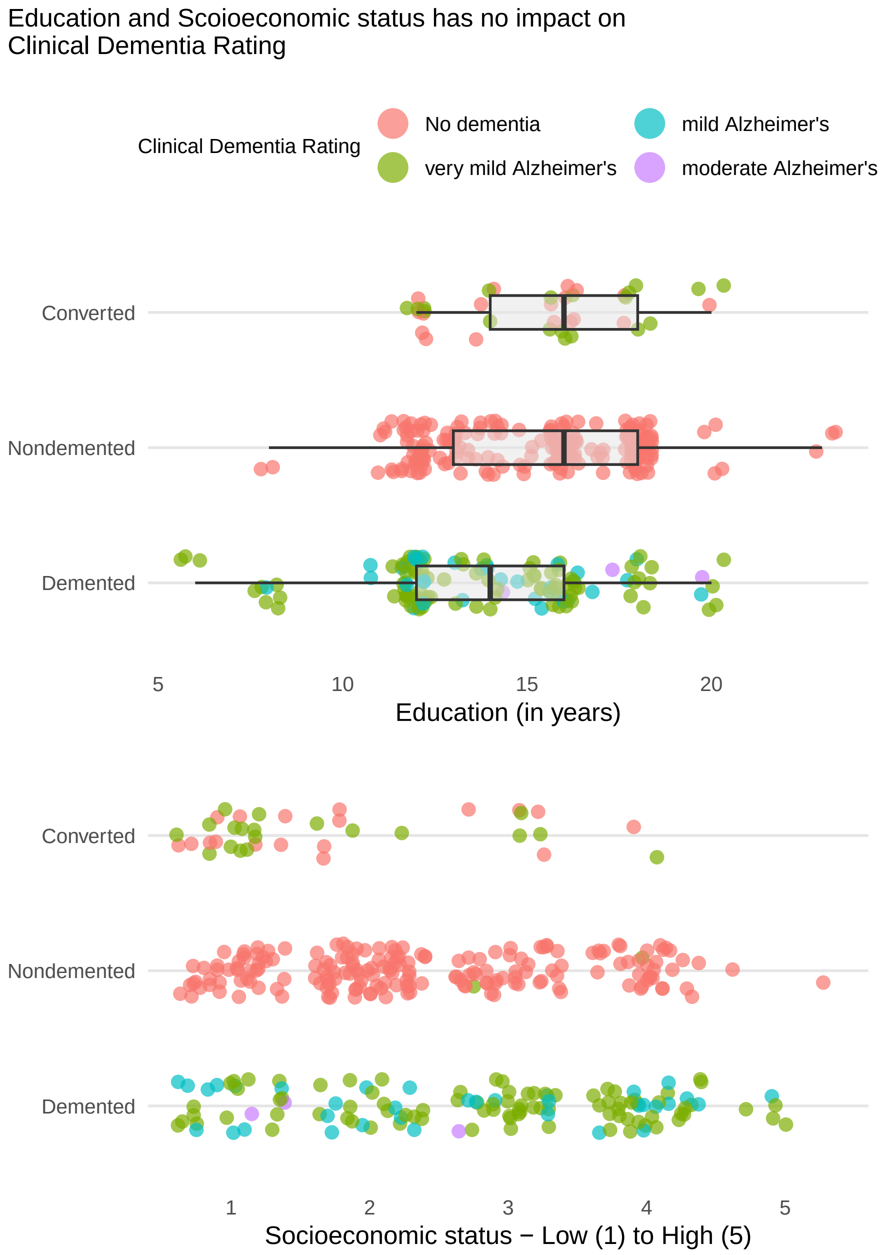


```
df_new |>
  select(group, mmse, cdr) |>
  mutate(group = fct_relevel(group, c("Demented", "Nondemented", "Converted"))) |>
  ggplot(aes(x = group, y = mmse)) +
  geom_point(aes(color = cdr), alpha = 0.7, size = 2.5, position = position_jitter(width
  geom_boxplot(fill = "grey90", width = 0.25, outlier.shape = NA, alpha = 0.5) +
  scale_color_discrete(labels = c("No dementia", "very mild Alzheimer's", "mild Alzheimer
  coord_flip() +
  labs(x = "",
       y = "Mini-Mental State Examination Score (MMSE)",
       title = "Nondemented individuals have higher MMSE score compared to Dementia patie
       color = "Clinical Dementia Rating") +
  theme_classic() +
  theme(plot.title = element_text(size = 12),
        plot.title.position = "plot",
        axis.title.y = element_blank(),
        axis.ticks = element_blank(),
        axis.line = element_blank(),
        axis.text = element_text(size = 12),
        axis.title.x = element_text(size = 12),
        panel.grid.major.y = element_line(color = "grey90", linewidth = 0.5),
        legend.position = "top",
        legend.text = element_text(size = 10)) +
  guides(color = guide_legend(ncol = 2,override.aes = list(size = 5)))
```



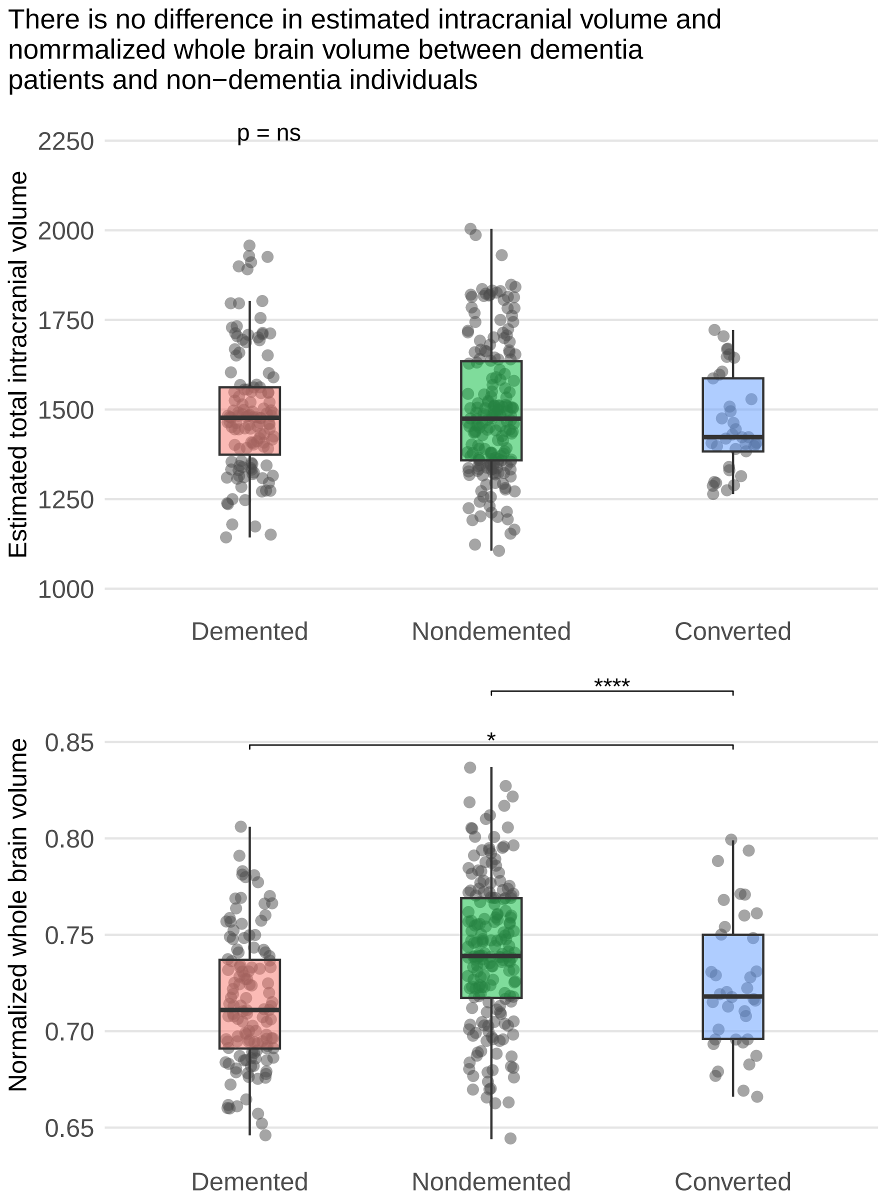
```
P3 <- df_new |>
  select(group, educ, cdr) |>
  mutate(group = fct_relevel(group, c("Demented", "Nondemented", "Converted"))) |>
  ggplot(aes(x = group, y = educ)) +
  geom_point(aes(color = cdr), alpha = 0.7, size = 2, position = position_jitter(width =
  geom_boxplot(fill = "grey90", width = 0.25, outlier.shape = NA, alpha = 0.5) +
  scale_color_discrete(labels = c("No dementia", "very mild Alzheimer's", "mild Alzheimer
  coord_flip() +
  labs(x = "",
       y = "Education (in years)",
       color = "Clinical Dementia Rating") +
  theme_classic() +
  theme(axis.title.y = element_blank(),
        axis.ticks = element_blank(),
        axis.line = element_blank(),
        axis.text = element_text(size = 8),
        axis.title.x = element_text(size = 10),
        panel.grid.major.y = element_line(color = "grey90", linewidth = 0.5),
        legend.position = "top",
        legend.text = element_text(size = 8),
        legend.title = element_text(size = 8)) +
  guides(color = guide_legend(ncol = 2,override.aes = list(size = 5)))
```

```
P4 <- df_new |>
  select(group, ses, cdr) |>
  mutate(group = fct_relevel(group, c("Demented", "Nondemented", "Converted"))) |>
  ggplot(aes(x = group, y = ses)) +
  geom_point(aes(color = cdr), alpha = 0.7, size = 2, position = position_jitter(width =
  scale_color_discrete(labels = c("No dementia", "very mild Alzheimer's", "mild Alzheimer
  coord_flip() +
  labs(x = "",
       y = "Socioeconomic status - Low (1) to High (5)",
       color = "Clinical Dementia Rating") +
  theme_classic() +
  theme(axis.title.y = element_blank(),
        axis.ticks = element_blank(),
        axis.line = element_blank(),
        axis.text = element_text(size = 8),
        axis.title.x = element_text(size = 10),
        panel.grid.major.y = element_line(color = "grey90", linewidth = 0.5),
        legend.position = "top",
        legend.text = element_text(size = 8),
        legend.title = element_text(size = 8)) +
  guides(color = guide_legend(ncol = 2,override.aes = list(size = 5)))

P3 / P4 +
  plot_annotation(title = "Education and Scoioeconomic status has no impact on \nClinical
                  theme = theme(legend.position = "top",
                                plot.title = element_text(size = 10))) +
  plot_layout(guides = "collect")
```



```
# shapiro.test(df_new$e_tiv) # data not normally distributed
# shapiro.test(df_new$n_wbv) # data normally distributed
```

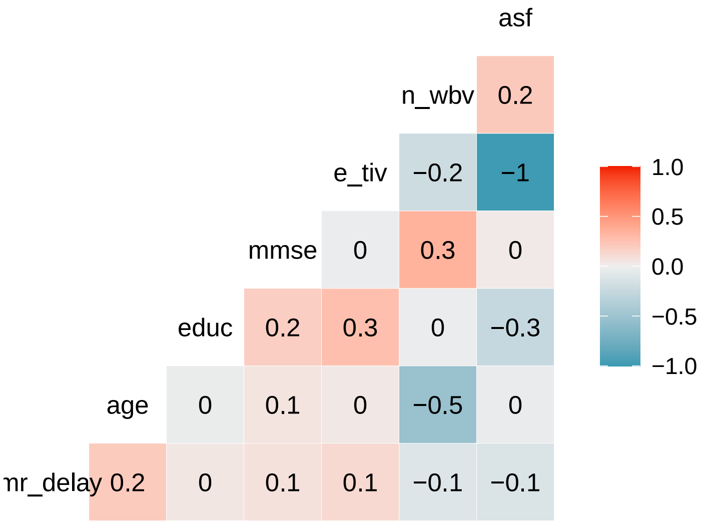
```r
stat.test <- df_new |>
  dunn_test(e_tiv ~ group)

stat.test1 <- df_new |>
  tukey_hsd(n_wbv ~ group) |>
  add_xy_position(x = "group", dodge = 0.8)

P5 <- df_new |>
  select(group, e_tiv) |>
  mutate(group = fct_relevel(group, c("Demented", "Nondemented", "Converted"))) |>
  ggplot(aes(x = group, y = e_tiv, fill = group)) +
  geom_point(color = "grey30", alpha = 0.5, size = 2, position = position_jitter(width =
  geom_boxplot(width = 0.25, outlier.shape = NA, alpha = 0.5) +
  scale_y_continuous(limits = c(1000,2250),
                     breaks = seq(1000, 2250,250)) +
  labs(y = "Estimated total intracranial volume") +
  theme_classic() +
  theme(axis.title.x = element_blank(),
        axis.ticks = element_blank(),
        axis.line = element_blank(),
        axis.text = element_text(size = 12),
        axis.title.y = element_text(size = 12),
        panel.grid.major.y = element_line(color = "grey90", linewidth = 0.5),
        legend.position = "none") +
  stat_kruskal_test(group.by = "x.var", label = "p = {p.signif}")

P6 <- df_new |>
  select(group, n_wbv) |>
  mutate(group = fct_relevel(group, c("Demented", "Nondemented", "Converted"))) |>
  ggplot(aes(x = group, y = n_wbv)) +
  geom_point(color = "grey30", alpha = 0.5, size = 2, position = position_jitter(width =
  geom_boxplot(aes(fill = group), width = 0.25, outlier.shape = NA, alpha = 0.5) +
  labs(y = "Normalized whole brain volume") +
  theme_classic() +
  theme(axis.title.x = element_blank(),
        axis.ticks = element_blank(),
        axis.line = element_blank(),
        axis.text = element_text(size = 12),
        axis.title.y = element_text(size = 12),
        panel.grid.major.y = element_line(color = "grey90", linewidth = 0.5),
        legend.position = "none") +
  stat_pvalue_manual(stat.test1, label = "p.adj.signif", hide.ns = TRUE, tip.length = 0.0

P5 / P6 +
    plot_annotation(title = "There is no difference in estimated intracranial volume and
```



Second, we measure correlation between variables. The correlation matrix suggests weak or no correlation between numericals variables.

*There is a strong negative correlation between estimated total intracranial volume (eTIV) and Atlas Scaling Factor (ASF). The ***ASF*** is a one-parameter scaling factor that allows for comparison of the estimated total intracranial volume (***eTIV***) based on differences in human brain volume, therefore, the correlation is expected and not meaningful for the analysis. **I will drop ASF from the modeling to avoid multicolinearity.**

```
ggcorr(df_new,
       label = TRUE,
       legend.size = 10)
```



Conclusion I

Based on the exploratory data analysis, we derive the following conclusions.

1. Men are more likely to have dementia.
2. There is no obvious relationship between age/sex and dementia diagnosis.
3. Non-demented individuals have higher MMSE score compared to Dementia patients.
4. Education and Scoioeconomic status has no impact on Clinical Dementia Rating.
5. There is no difference in estimated intracranial volume and normalized whole brain volume between dementia patients and non-dementia individuals.

6. There is no correlation between MMSE score and estimated intracranial volume/normalized whole brain volume

Random Forest Classification Model

Based on the given data, can we predict dementia and Alzheimer's disease? This is a classification problem. We will employ decision tree which is a supervised learning algorithm to predict Alzheimer's disease based on socioeconomic factors.

```
model_data <- df_new |>
  select(-asf)

# partition data
set.seed(500)
ind <- sample(2, nrow(model_data), replace = T, prob = c(0.8, 0.2))
train <- df_new[ind == 1, ]
test <- df_new[ind == 2, ]
```

```
rf <- randomForest(group ~.,
                   data=train,
                   proximity=TRUE,
                   importance=TRUE,
                   predicted = TRUE)
print(rf)
```

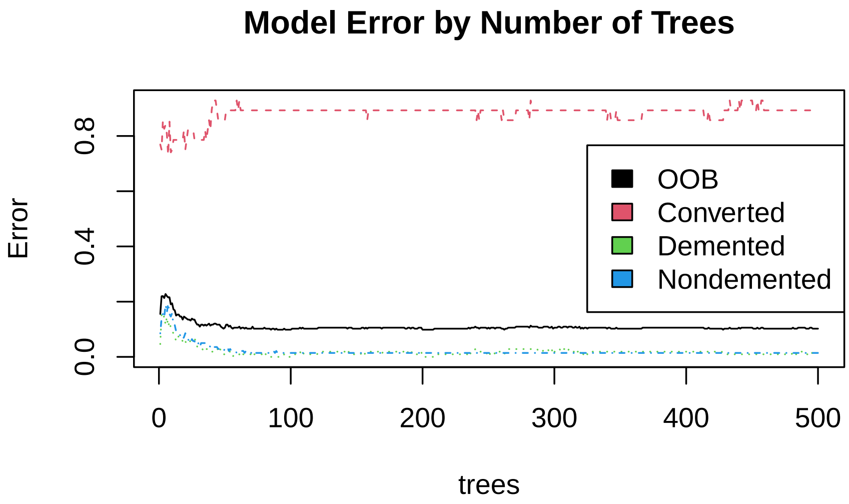
```

Call:
 randomForest(formula = group ~ ., data = train, proximity = TRUE,      importance = TRUE, p
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 10.22%
Confusion matrix:
            Converted Demented Nondemented class.error
Converted           3       11          14 0.892857143
Demented            1      105           0 0.009433962
Nondemented         0        2         138 0.014285714
```

Model cross-validation on test data

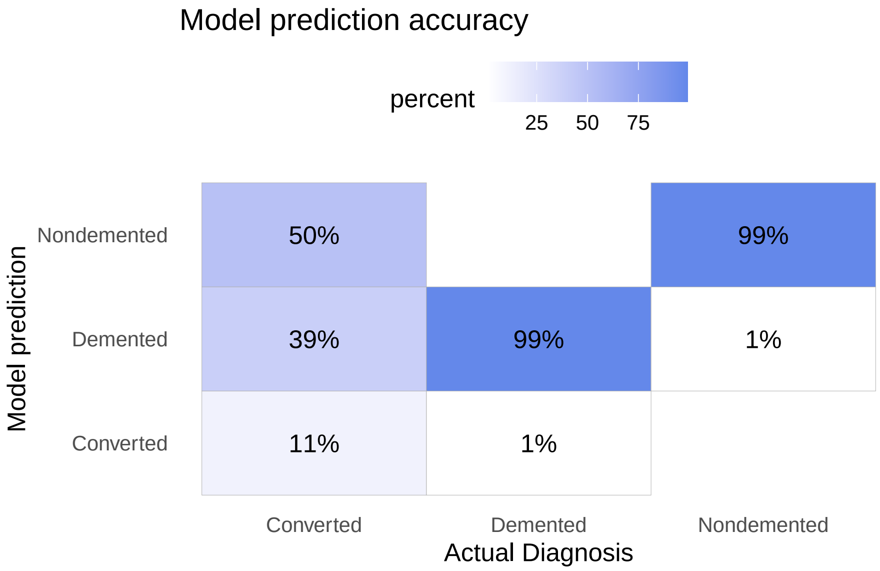
```r
test_pred <- predict(rf,
                    newdata = test)

accuracy <- mean(test_pred == test$group)*100
cat('Accuracy on testing data: ', round(accuracy, 2), '%',  sep='')
```

```
Accuracy on testing data: 92.5%
```

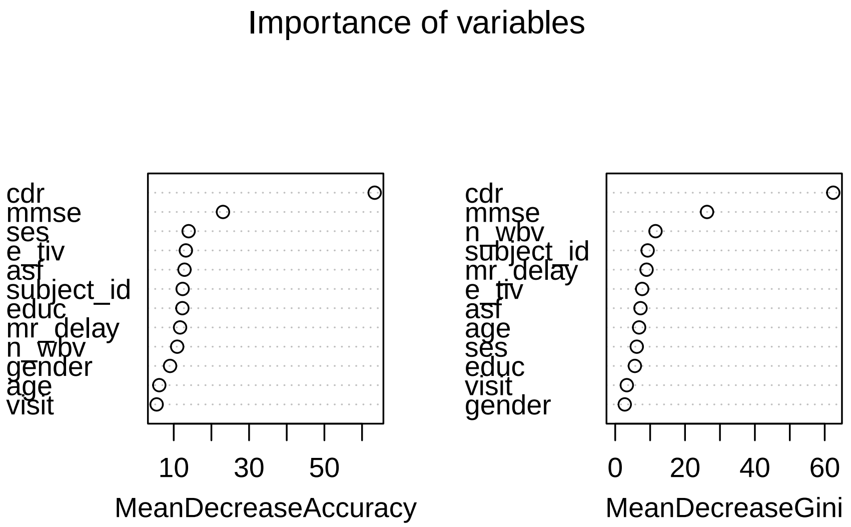
```r
plot(rf, main = "Model Error by Number of Trees")
legend(x = "right",
       legend = colnames(rf$err.rate),
       fill = 1:ncol(rf$err.rate))
```



```r
pred <- as.data.frame(predict(rf))

a <- train |>
  cbind(pred) |>
  group_by(group, predict(rf)) |>
  summarise(n = n()) |>
  mutate(freq = n/sum(n)) |>
  ungroup() |>
  rename("predict" = "predict(rf)") |>
  mutate(percent = round(freq, digits = 2) * 100)

a |>
  ggplot(aes(x = group, y = predict)) +
  geom_tile(aes(fill = percent), color = "grey70") +
  geom_text(aes(label = paste0(percent,"%"))) +
  scale_fill_gradient(low = "white", high = "#6488ea") +
  labs(x = "Actual Diagnosis",
       y = "Model prediction",
       title = "Model prediction accuracy") +
  theme_classic() +
  theme(axis.line = element_blank(),
        axis.ticks = element_blank(),
        legend.position = "top")
```



```
varImpPlot(rf, main = "Importance of variables")
```



Conclusion II

The Random Forest model shows great accuracy in predicting Alzheimer's disease diagnosis based on socioeconomic and brain imaging data. Among all the variables, clinical dementia rating (CDR) and mini-mental state examination score show greater reliability in accurately predicting dementia. While the prediction accuracy is 98% in classifying demented and nondemented individuals, the model performance reflects well on the data for converted patients. It is difficult to diagnose dementia in individuals when their CDR and MMSE scores do not suggest any cognitive decline. While Alzheimer's is a complex disease, features such as CDR and MMSE can be valuable in timely diagnosis.