Thema09DementiaPrediction

Ewoud

2023-09-06

## Introduction

The question this research is aiming to give an answer to is: *How accurate can a machine learning algorithm be, that predicts if a subject has dementia using different clinical parameters?*

Dataset: <https://www.kaggle.com/datasets/shashwatwork/dementia-prediction-dataset>

## Codebook

## # A tibble: 15 × 2  
## variable\_id item\_text   
## <chr> <chr>   
## 1 Subject ID "Id of subject"   
## 2 MRI ID "Id of MRI"   
## 3 Group "Converted / Demented/ Nondemented"   
## 4 Visit "Number of visit "   
## 5 MR Delay "Delay with MRI"   
## 6 M/F "Gender : Male / Female "   
## 7 Hand "Handedness"   
## 8 Age "Age of the subject at time of visit"  
## 9 EDUC "Years of education"   
## 10 SES "Socioeconomic status"   
## 11 MMSE "Mini-Mental State Examination score"  
## 12 CDR "Clinical Dementia Rating"   
## 13 eTIV "Estimated total intracranial volume"  
## 14 nWBV "Normalized whole-brain volume"   
## 15 ASF "Atlas scaling factor"

| Name | description | unit | value |
| --- | --- | --- | --- |
| Subject.ID | Id of the patient | ? | ? |
| MRI ID | Id of MRI | ? | ? |
| Group | Converted / Demented/ Nondemented | ? | ? |
| Visit | Number of visit | ? | ? |
| MR Delay | Delay with MRI | ? | ? |
| M/F | Gender : Male / Female | ? | ? |
| Hand | Handedness | ? | ? |
| Age | Age of the subject at time of visit | ? | ? |
| EDUC | Weight of shell | ? | ? |
| SES | Socioeconomic status | ? | ? |
| MMSE | Mini-Mental State Examination score | ? | ? |
| CDR | Clinical Dementia Rating | ? | ? |
| eTIV | Estimated total intracranial volume | ? | ? |
| nWBV | Normalized whole-brain volume | ? | ? |
| ASF | Atlas scaling factor | ? | ? |

### Description of some of the rows

MRI :

SES : Socioeconomic status as assessed by the Hollingshead Index of Social Position and classified into categories from 1 (highest status) to 5 (lowest status)

MMSE : Mini–Mental State Examination (MMSE) The Mini–Mental State Examination (MMSE) or Folstein test is a 30-point questionnaire that is used extensively in clinical and research settings to measure cognitive impairment. It is commonly used in medicine and allied health to screen for dementia. It is also used to estimate the severity and progression of cognitive impairment and to follow the course of cognitive changes in an individual over time; thus making it an effective way to document an individual’s response to treatment. The MMSE’s purpose has been not, on its own, to provide a diagnosis for any particular nosological entity.

Interpretations

Any score greater than or equal to 24 points (out of 30) indicates a normal cognition. Below this, scores can indicate severe (≤9 points), moderate (10–18 points) or mild (19–23 points) cognitive impairment. The raw score may also need to be corrected for educational attainment and age. That is, a maximal score of 30 points can never rule out dementia. Low to very low scores correlate closely with the presence of dementia, although other mental disorders can also lead to abnormal findings on MMSE testing. The presence of purely physical problems can also interfere with interpretation if not properly noted; for example, a patient may be physically unable to hear or read instructions properly, or may have a motor deficit that affects writing and drawing skills.

CDR : Clinical Dementia Rating (CDR) The CDR™ in one aspect is a 5-point scale used to characterize six domains of cognitive and functional performance applicable to Alzheimer disease and related dementias: Memory, Orientation, Judgment & Problem Solving, Community Affairs, Home & Hobbies, and Personal Care. The necessary information to make each rating is obtained through a semi-structured interview of the patient and a reliable informant or collateral source (e.g., family member) referred to as the CDR™ Assessment Protocol.

The CDR™ Scoring Table provides descriptive anchors that guide the clinician in making appropriate ratings based on interview data and clinical judgment. In addition to ratings for each domain, an overall CDR™ score may be calculated through the use of an CDR™ Scoring Algorithm. This score is useful for characterizing and tracking a patient’s level of impairment/dementia:

0 = Normal 0.5 = Very Mild Dementia 1 = Mild Dementia 2 = Moderate Dementia 3 = Severe Dementia

eTIV: Estimated total intracranial volume (eTIV) The ICV measure, sometimes referred to as total intracranial volume (TIV), refers to the estimated volume of the cranial cavity as outlined by the supratentorial dura matter or cerebral contour when dura is not clearly detectable. ICV is often used in studies involved with analysis of the cerebral structure under different imaging modalities, such as Magnetic Resonance (MR), MR and Diffusion Tensor Imaging (DTI), MR and Single-photon Emission Computed Tomography (SPECT), Ultrasound and Computed Tomography (CT). ICV consistency during aging makes it a reliable tool for correction of head size variation across subjects in studies that rely on morphological features of the brain. ICV, along with age and gender are reported as covariates to adjust for regression analyses in investigating progressive neurodegenerative brain disorders, such as Alzheimer’s disease, aging and cognitive impairment. ICV has also been utilized as an independent voxel based morphometric feature to evaluate age-related changes in the structure of premorbid brai, determine characterizing atrophy patterns in subjects with mild cognitive impairment (MCI) and Alzheimer’s disease (AD), delineate structural abnormalities in the white matter (WM) in schizophrenia, epilepsy, and gauge cognitive efficacy.

nWBV : Normalized whole-brain volume, expressed as a percent of all voxels in the atlas-masked image that are labeled as gray or white matter by the automated tissue segmentation process

ASF: Atlas scaling factor (unitless). Computed scaling factor that transforms native-space brain and skull to the atlas target (i.e., the determinant of the transform matrix)

## Cleaning

First thing to do is to clean the dataset of zero values or outliers that can obstruct this research

### Missing values

Lets filter the rows out of the dataset with 0 or na.

Data1\_filtered <- Data1 %>% drop\_na()

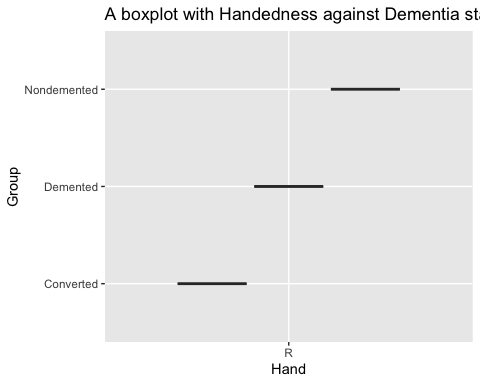
19 object are filtered

### finding outliers

## Subject ID MRI ID Group Visit   
## Length:354 Length:354 Length:354 Length:354   
## Class :character Class :character Class :character Class :character   
## Mode :character Mode :character Mode :character Mode :character   
##   
##   
##   
## MR Delay M/F Hand Age   
## Min. : 0.0 Length:354 Length:354 Min. :60.00   
## 1st Qu.: 0.0 Class :character Class :character 1st Qu.:71.00   
## Median : 559.5 Mode :character Mode :character Median :77.00   
## Mean : 601.4 Mean :77.03   
## 3rd Qu.: 882.5 3rd Qu.:82.00   
## Max. :2639.0 Max. :98.00   
## EDUC SES MMSE CDR eTIV   
## Min. : 6.00 Min. :1.00 Min. : 4.00 Min. :0.0000 Min. :1106   
## 1st Qu.:12.00 1st Qu.:2.00 1st Qu.:27.00 1st Qu.:0.0000 1st Qu.:1358   
## Median :15.00 Median :2.00 Median :29.00 Median :0.0000 Median :1470   
## Mean :14.70 Mean :2.46 Mean :27.41 Mean :0.2712 Mean :1490   
## 3rd Qu.:16.75 3rd Qu.:3.00 3rd Qu.:30.00 3rd Qu.:0.5000 3rd Qu.:1595   
## Max. :23.00 Max. :5.00 Max. :30.00 Max. :2.0000 Max. :2004   
## nWBV ASF   
## Min. :0.6444 Min. :0.8755   
## 1st Qu.:0.6987 1st Qu.:1.1002   
## Median :0.7291 Median :1.1935   
## Mean :0.7299 Mean :1.1938   
## 3rd Qu.:0.7569 3rd Qu.:1.2923   
## Max. :0.8368 Max. :1.5873

No visible outliers

### Remove colums with no meaning

 When looking at the Handedness parameter we can see that the only variable is R for right handed people, because everybody is right handed we can remove the column

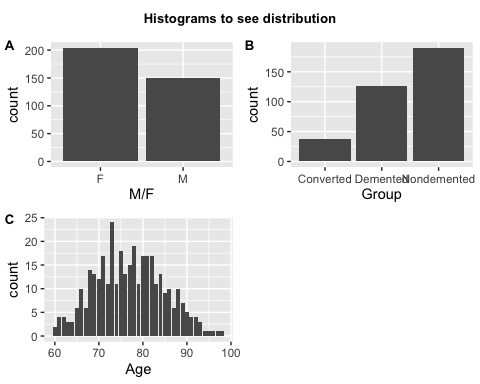
removing the MRI id and delay because these are not parameters that have influence on the outcome if a subject has dementia

CDR is basically a parameter telling if a patient has dementia or not so it will not be taking in for machine learning but testing if algorithm is correct.

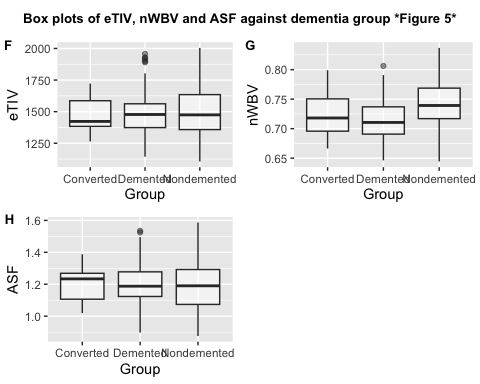
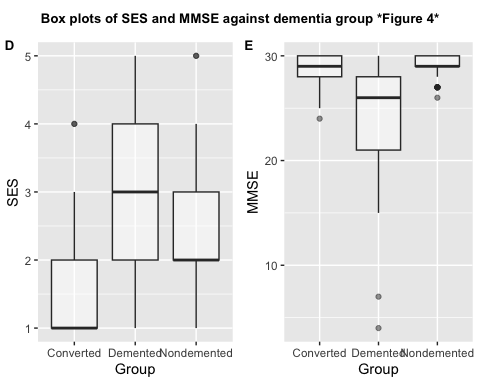
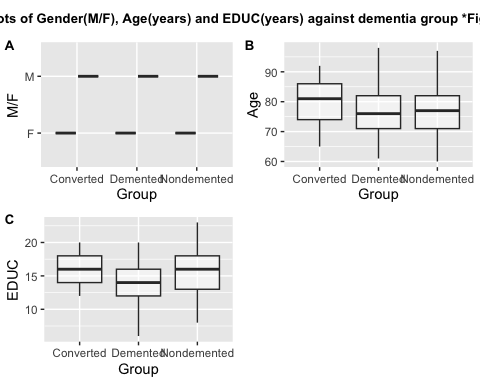
## Testing the dataset

Second thing to do is to look at the underlying distribution and the variation within the dataset

### equal distribution



## Finding correlation

 Waardes die waarschijnlijk van invloed zijn Educatie, SES, MMSE, nWBV, klein beetje eTIV en ASFs