Classifying Dementia Using the Open Access Series of Imaging Studies (OASIS) Longitudinal Dataset

Sarah Bao
Brown Department of Mathematics & Computer Science
October 23, 2024

https://github.com/s-bao/data1030-project.git

Introduction

Problem Statement:

- aim to classify individuals as either "Demented", "Nondemented", or “Converted” based on features such as age, cognitive test scores, and brain volume measurements
- importance: early detection of dementia is vital for timely intervention and management, potentially improving patients' quality of life
    - doctors make clinical diagnoses based off comprehensive assessment of patient's condition, since a definitive diagnosis can only be made after an autopsy of the brain
    - model can assist clinicians in identifying high-risk individuals

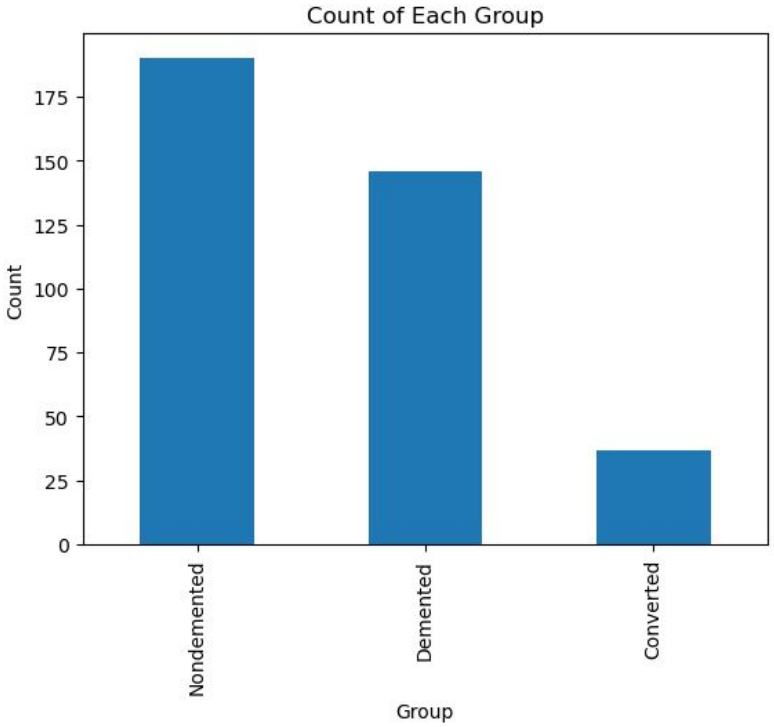
Dataset Description:

- Kaggle MRI and Alzheimer's (https://www.kaggle.com/datasets/jboysen/mri-and-alzheimers/data)
- OASIS longitudinal dataset: contains information on 150 subjects ages 60 to 96; longitudinal because it has information on at least 2 visits for each subjects, where visits are separated by at least 1 year, where all cognitive and MRI measurements are updated during each visit

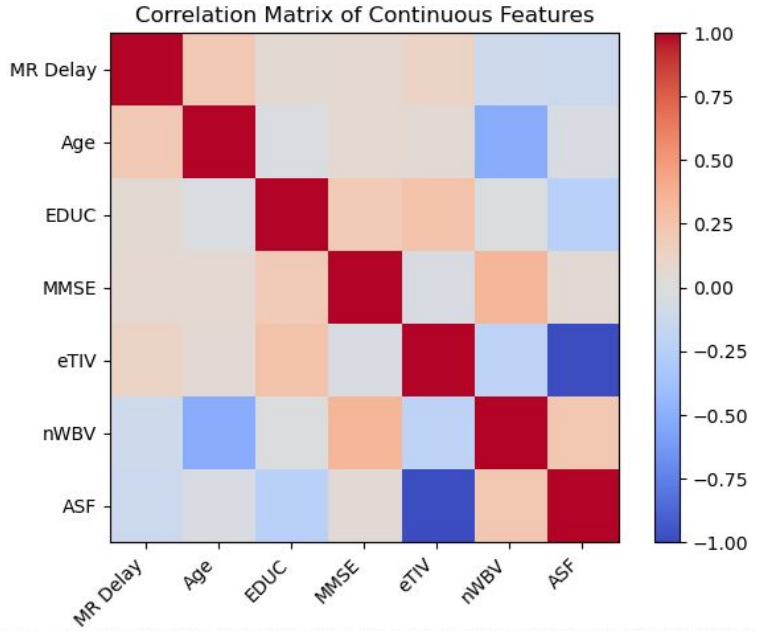
Qualities of Dataset

- 373 by 15
- Non-IID data since multiple data points provide information on the same person
- Limited missing data (will go more into this later)
- 15 features: Subject ID, MRI ID, Group, Visit, MR Delay, M/F, Hand, Age, EDUC, SES, MMSE, CDR, eTIV, nWBV, ASF
    - Target variable: Group
    - MMSE, CDR (mini-mental state examination, clinical dementia rating) are scores that come from two different cognitive tests
    - eTIV, nWBV, ASF (estimated total volume in skull, normalized volume of whole brain, Atlas scaling factor) are metrics that come from the open access MRI scans for each person during each visit

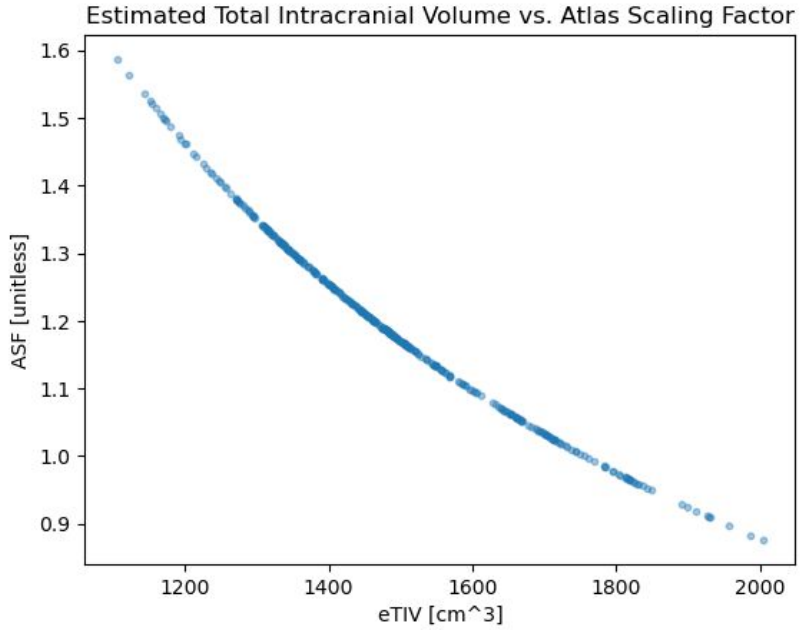
EDA: Target Variable Distribution



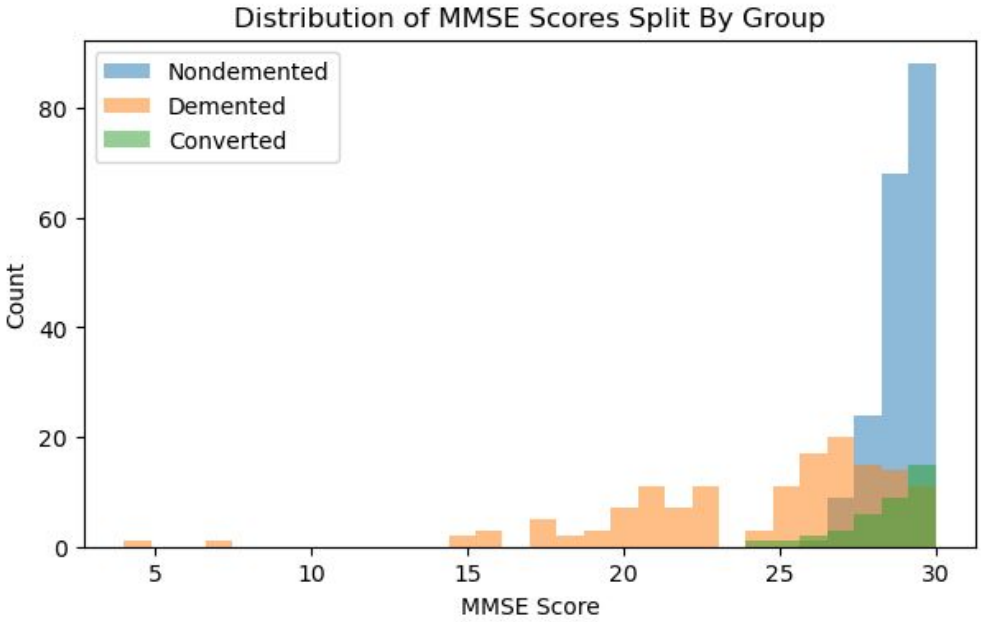
EDA: Correlation Map of Continuous Features



Plot of the correlation (Pearson) matrix of continuous features in the OASIS Longitudinal Dataset.

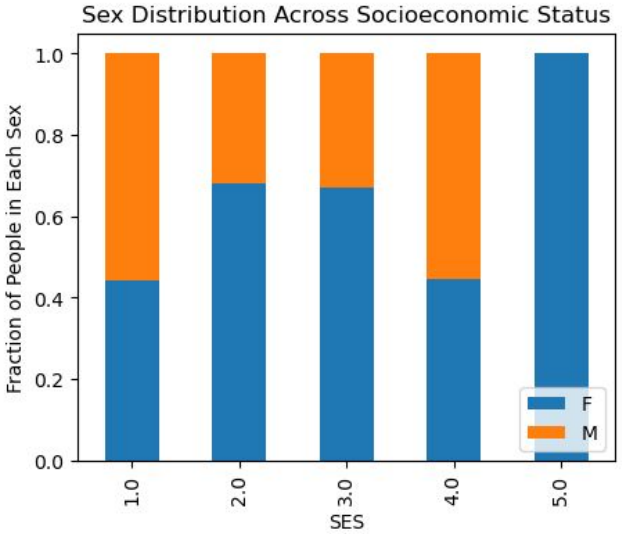


EDA: Distribution of MMSE Scores by Group

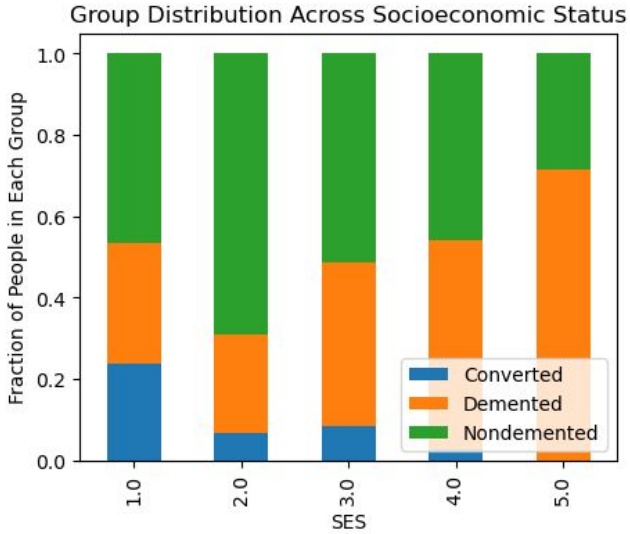


MMSE Score: 9 or lower = Severe cognitive impairment, 10-18 = Moderate cognitive impairment, 19-23 = Mild cognitive impairment, 24+ = Normal cognition

EDA: Socioeconomic Status (SES) Exploration



Note that for the socioeconomic status (SES) variable, 1.0 is highest status, and 5.0 is lowest status.



Note that for the socioeconomic status (SES) variable, 1.0 is highest status, and 5.0 is lowest status.

Splitting

- has group structure ➜ group split on “Subject ID”
  - samples are not iid because each subject appears multiple times in the dataset
- imbalanced ➜ stratified split on “Group” target variable
- relatively small (< 100k) data points ➜ kfold for cross validation
- **used StratifiedGroupKFold**

Preprocessing

- dropped “Hand”, “Subject ID”, “MRI ID”, and “Group” leaving 11 features
- decided which encoder to use on each feature
  - minmax : age, educ, MMSE
  - one-hot: M/F
  - ordinal: visit, SES, CDR
  - standard: MR delay, eTIV, nWBV, ASF

```
Before preprocessing...
X train: (228, 11)
X val: (69, 11)
X test: (76, 11)

After preprocessing...
X train: (228, 12)
X val: (69, 12)
X test: (76, 12)
```

Preprocessing

```
Missing values in each column:
Subject ID     0
MRI ID         0
Group          0
Visit          0
MR Delay       0
M/F            0
Hand           0
Age            0
EDUC           0
SES           19
MMSE           2
CDR            0
eTIV           0
nWBV           0
ASF            0
```

- Missing values only exist for SES and MMSE, both continuous features
  - ~5% missing for SES
  - ~0.5% missing for MMSE
- ~5% of visits have missing values

Thank you for listening!