Predicting Alzheimer’s in Patients

Final Project

William Acorda

DATA 824

Purpose

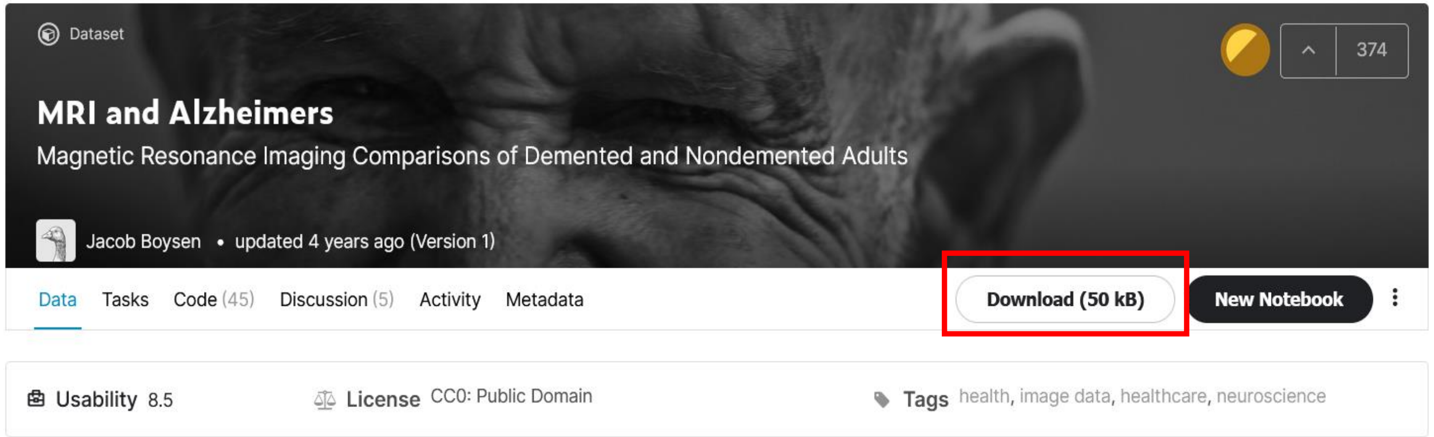
- Conduct an exploratory data analysis (EDA) on a dataset containing MRI and Alzheimer’s patient data
- Conduct an analysis to predict (classify) whether a patient has dementia or not
- Look at available features in the data to determine importance in predicting whether patient has dementia

Overview: OASIS

- Open Access Series of Imaging Studies (OASIS)
  - Website: https://www.oasis-brains.org/
  - A project that focuses on providing neuroimaging datasets for scientific use
  - Project responsible for making dataset freely available
  - Data is generated by the Knight ADRC and its affiliated studies
    - Website: https://knightadrc.wustl.edu/

Data Acquisition

- Source: https://www.kaggle.com/jboysen/mri-and-alzheimers
  - *Note: Data is made available and credited to the OASIS project*
- How to Download via Kaggle:



MRI and Alzheimers

Magnetic Resonance Imaging Comparisons of Demented and Nondemented Adults

Jacob Boysen • updated 4 years ago (Version 1)

Usability 8.5

License CC0: Public Domain

Tags health, image data, healthcare, neuroscience

- File Type: ZIP file
  - Contains 2 CSV files
    - File 1: oasis_cross-sectional.csv
    - File 2: oasis_longitudinal.csv

Data: Features

- Number of Features: 12 (see table)

Feature	Description	Type
ID	Identification (Patient)	Character
M.F	Gender (Male or Female)	Character/Nominal
Hand	Dominant Hand	Character
Age	Age in Years	Numeric
Educ	Education Level	Numeric/Ordinal
SES	Socioeconomic Status	Numeric/Ordinal
MMSE	Mini Mental State Examination	Numeric
CDR	Clinical Dementia Rating	Numeric/Ordinal
eTIV	Estimated Total Intracranial Volume	Numeric
nWBV	Normalize Whole Brain Volume	Numeric
ASF	Atlas Scaling Factor	Numeric
Delay	Time (Feature Unknown)	Numeric

Data: Missingness

- Table shows the missingness by number of observations
- 5 variables have significant amount of missingness:
  - Educ, SES, MMSE, CDR and Delay
- What can be done?
  - Go back and try to get values from the study
  - Omit missing records/observations
  - Imputation

Feature	Missingness (# of obs)
ID	0
M.F	0
Hand	0
Age	0
Educ	201
SES	220
MMSE	201
CDR	201
eTIV	0
nWBV	0
ASF	0
Delay	416

Data: Missingness

- What was done?
  - Trying to retrieve missing data from the previous study was not possible. Therefore, the option chosen was to omit records/observations with significant missingness.
    - This is likely not the best option
    - The option for imputation was not utilized but could be used in the future

Data: Preparation

- Based on the features there are several that will be dropped for the analysis:
  - ID – Is a unique identifier and is unnecessary (not helpful) in prediction
  - Hand – There is only one option in the dataset and that’s R (right-handed)
  - Delay – Missingness in the data is approximately 95%
- Based on the features there are several that are ordinal and need to be converted to a factor:
  - M.F, Educ, SES, CDR
  - CDR will be converted to a two-level factor (dementia/Yes or no dementia/No) for prediction, i.e. binary variable

Data: Privacy & Security

- The data has been redacted such that information that can identify the patient is unavailable in this dataset
- Since this is actual patient information as part of the OASIS project it is important to follow the data agreement when using the data
  - Visit website: https://www.oasis-brains.org/ for details

Exploratory Data Analysis: Visualization

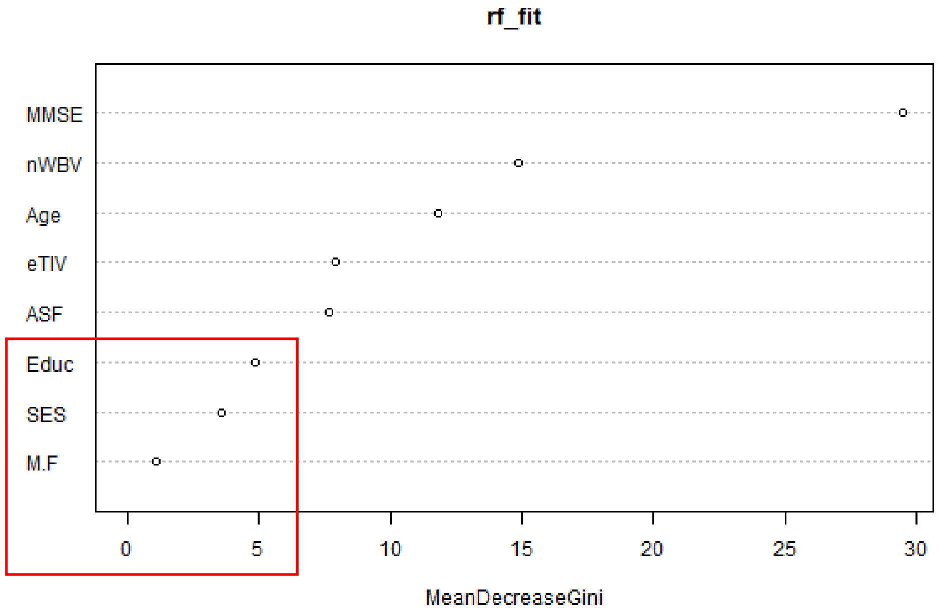
- Several data visualizations (plots/charts) were created:
  - Density plot showing Age vs. Dementia
  - Correlation Plot for numerical variables
  - Histogram for MMSE
  - Bar plots for categorical variables
    - Educ (Education Score)
    - SES (Socioeconomic Status Score)
    - M.F (Gender)
  - Boxplots for numerical variables
    - ASF
    - MMSE
    - eTIV
    - nWBV
  - Correlation plot for numerical variables (interactive)

Exploratory Data Analysis: Takeaways

- Patients with dementia have ages between 60 and 90 years of age
- Patients with a high MMSE score (approximately 25) tend to not have dementia
  - MMSE seems to be an indicator of dementia
- Patients with higher nMBV values tend to not have dementia
- Factors such as Education and Socioeconomic Scores don’t seem to indicate dementia
- Correlated variables include:
  - eTIV and ASF: -0.99 correlation
  - nWBV and Age: -0.74 correlation
  - MMSE and nWBV: 0.48 correlation

Model Building: Feature Selection

- There were 8 features in the dataset, but reduced to 5 features after conducting a Random Forest feature selection method
  - The criteria of a mean decrease in Gini value of 5 or less was used arbitrarily
    - Educ, SES and M.F were dropped from the model based on this criteria



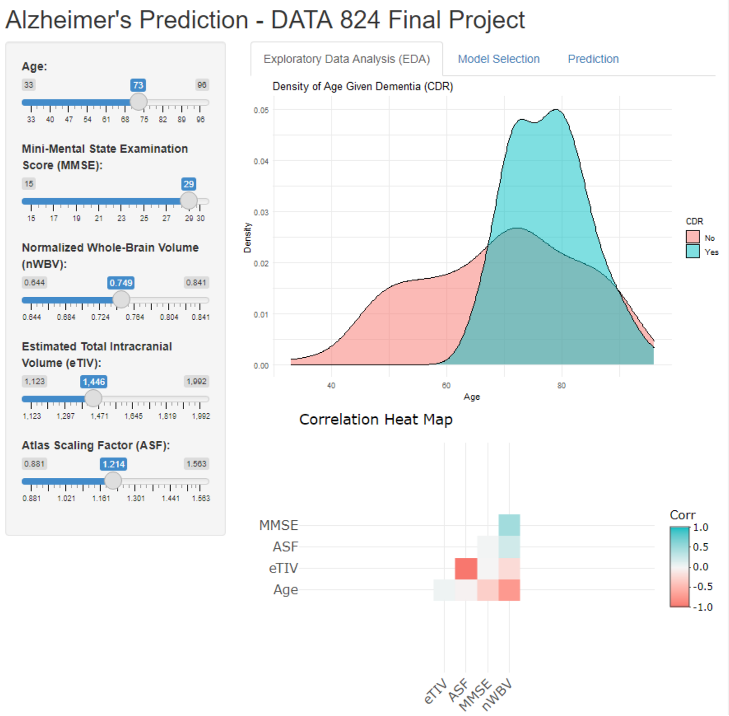
Model Building: Final Model

- Data was split into test and validation datasets and cross-validation was utilized to determine final model
  - The metric for assessment was the F1 Score
- Models selected for comparison were:
  - Logistic Regression
  - Single Decision Tree
  - KNN
  - Flexible Discriminant Analysis (FDA)
  - Support Vector Machine (SVM) using a radial kernel
- The “best” model i.e. final model was the SVM model with an F1 Score of 0.91

Shiny App: Description & Location

- The Alzheimer’s Prediction app was developed using the R package, *Shiny*
- The layout is structured using 1 side panel for inputs and 3 tabs where information is displayed
  - Tab 1 shows visualizations as part of the exploratory data analysis (EDA)
  - Tab 2 shows visualizations as part of the model selection process
  - Tab 3 shows a gauge visualization and utilizes the slider inputs to predict the probability that a patient has dementia
- The code can be located at the following GitHub repository
  - https://github.com/wacorda/MSASADS-DATA-824-Final-Project

Shiny App: Screenshots



Alzheimer's Prediction - DATA 824 Final Project

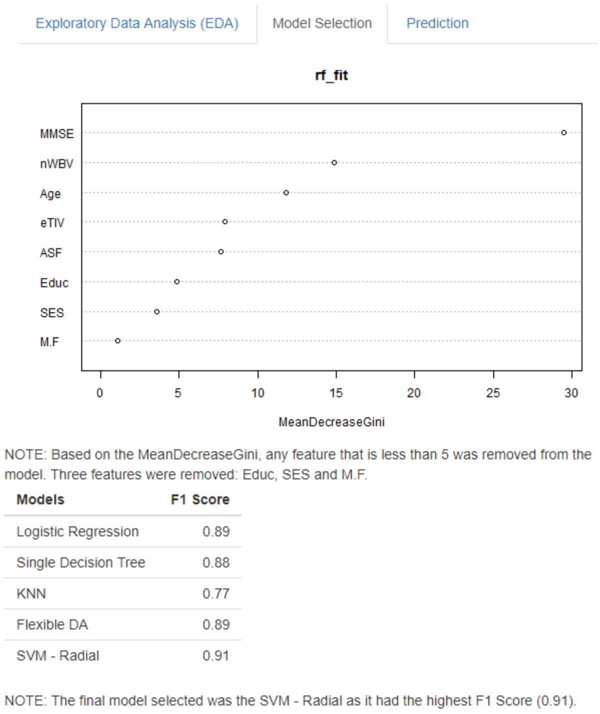
Age: 73

Mini-Mental State Examination Score (MMSE): 29

Normalized Whole-Brain Volume (nWBV): 0.749

Estimated Total Intracranial Volume (eTIV): 1,446

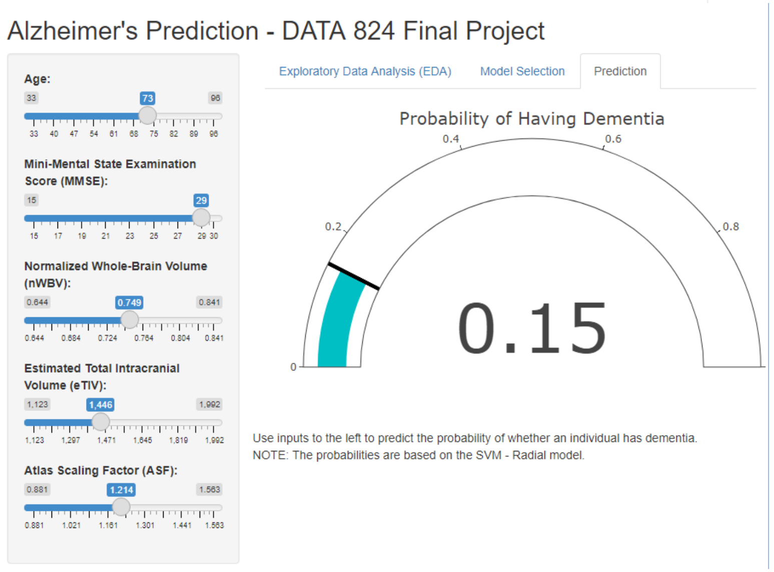
Atlas Scaling Factor (ASF): 1.214



NOTE: Based on the MeanDecreaseGini, any feature that is less than 5 was removed from the model. Three features were removed: Educ, SES and M.F.

Models	F1 Score
Logistic Regression	0.89
Single Decision Tree	0.88
KNN	0.77
Flexible DA	0.89
SVM - Radial	0.91

NOTE: The final model selected was the SVM - Radial as it had the highest F1 Score (0.91).



Alzheimer's Prediction - DATA 824 Final Project

Age: 73

Mini-Mental State Examination Score (MMSE): 29

Normalized Whole-Brain Volume (nWBV): 0.749

Estimated Total Intracranial Volume (eTIV): 1,446

Atlas Scaling Factor (ASF): 1.214

Use inputs to the left to predict the probability of whether an individual has dementia.
NOTE: The probabilities are based on the SVM - Radial model.

Conclusion

- I hope you enjoyed my Shiny app that was used to predict the probability that a patient has dementia
- If you have any questions, please feel free to contact me
  - E-mail: wacorda@kumc.edu