Machine Learning Approaches for Alzheimer's Disease Analysis

Suong Tran, Quang Nguyen

Data Analytics Department, Dickinson College

Dickinson

Introduction

Alzheimer's disease (AD) presents a challenge with its progressive cognitive decline. Previous machine learning attempts to predict AD onset faltered due to complex feature interpretation.[1] This study tackles this issue using unsupervised machine learning on a dataset from the National Alzheimer's Coordinating Center (NACC).[2] By using easily interpretable features spanning demographics, lifestyle, health, cognition, and functionality of AD individuals, this research aims to deepen understanding of AD progression for early diagnosis and targeted interventions.

Data

The original dataset in this project was provided by the NACC. It contains data about 13689 patients collected since 2005 during standardized annual evaluations conducted at the NIA-funded Alzheimer's Disease Research Centers (ADRCs) across the country. The dataset used for analysis only contains 22 features selected based on literature review and easiness-to-understand for normal people.

The 22 features were selected and categorized into 5 groups:

1) Demographic Factors:
- Age (NACCAGE)
- Sex

2) Lifestyle Factors:
- BMI
- Smoking Habits

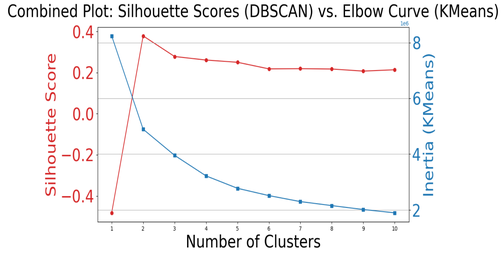
3) Health Conditions:
- Alcohol-related issues
- Diabetes (presence or absence)
- Hypertension (History/Presence)
- Cardiovascular Health (Heart attack or cardiac arrest)

4) Cognitive Assessment:
- CDRSUM (Standard CDR sum of boxes)
- DECIN (Informant report of memory decline)
- MOSLOW (Observations of slowed movement, expression changes)
- Memory (CDR-based assessment)

5) Functional Activities:
- SPEECH (Speech impairment)
- BILLS (Difficulty in handling bills)
- TRAVEL (Difficulty in traveling)
- MOFALLS (Frequency of falls)
- MOTREM (Rhythmic shaking)

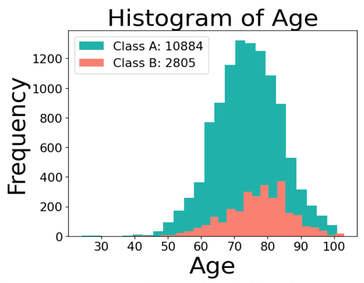
K-means clustering

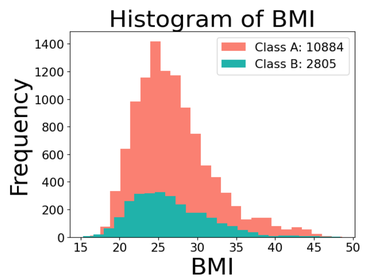
With no actual disease label for any patient, the most intuitive approach was doing unsupervised learning, and k-Means clustering was the selected method.



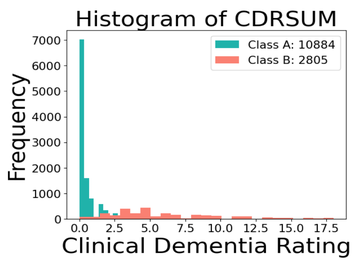
Interpretation of categorized classes

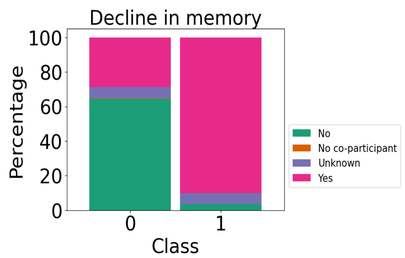
After clustering, it can be seen that most features in the first three groups have the same distribution for both clusters.

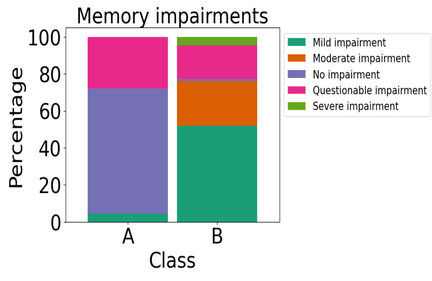


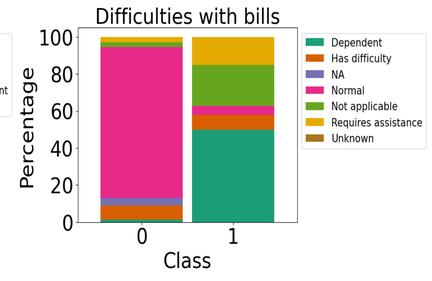


However, significant distinctions exist in the last two groups of feature.









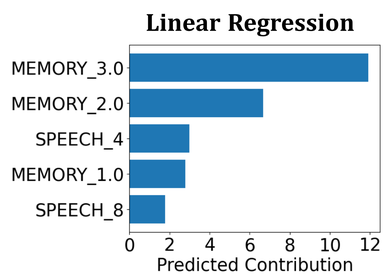
Class A appears to consist of individuals who are generally healthier, with fewer reported impairments in memory, motor function, speech, and activities of daily living.

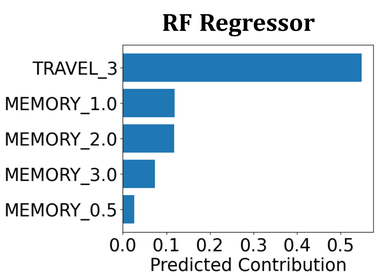
Class B, on the other hand, seems to have a higher prevalence of memory decline, motor function issues, falls, tremors, and difficulties with daily activities.

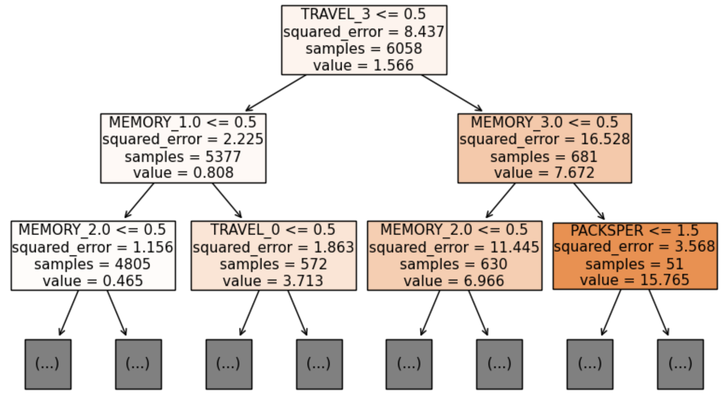
Supervised learning models

For supervised learning, the selected methods are linear regression and random forest. To do this, CDRSUM (clinical dementia rating score) is used as the dependent variable and the rest of the previously selected features are the independent variables. The metrics selected to assess the accuracy of these models are R-squared, mean squared error and mean absolute error.

Model	Linear Regression	LASSO Regression	Support Vector Regressor	Random Forest Regressor
Mean Squared Error	0.75	1.97	8.48	0.42
R-squared Score	0.91	0.77	0.01	0.95
Mean Absolute Error	0.46	0.55	1.95	0.25







References & Acknowledgement

[1] Lin, M.; Gong, P.; Yang, T.; Ye, J.; Albin, R. L.; Dodge, H. H. Big Data Analytical Approaches to the NACC Dataset: Aiding Preclinical Trial Enrichment. *Alzheimer Dis. Assoc. Disord.* **2018**, 32 (1), 18-27.
[2] National Alzheimer's coordinating center. *https://maccdata.org*

We extend my gratitude to Data Analytics department for this wonderful opportunity, and Professor Eren Bilen for his guidance. We also would like to express our sincere gratitude to NACC for giving access to the data.