Clemson University

Detection of Alzheimer's disease using Machine learning models.

CPSC 6300: Applied Data Science
Instructor: Dr. Nina Hubig
Semester: Spring 2023
Check Point - 1

BY

Jayanth Talasila
jtalasi@clemson.edu

Sreeram Paladugu
Spaladu@clemson.edu
C11281117

GitHub: https://github.com/sreerampaladugu10/ads-checkpoint-1

In this project, different machine learning models will be implemented to detect Alzheimer's disease. The best-performing model will be presented based on the evaluation metrics used.

About the given data set-

The Alzheimer's dataset used in this project contains information on various studies, including ADNI, HEART, COUPLES, PANACEA, PATRIOT-prelim, PATRIOT, and IAM. The dataset also includes information on the diagnosis of patients, with categories including MCI (mild cognitive impairment), AD (Alzheimer's disease), AUD (alcohol use disorder), and HC (healthy control).

To begin our analysis of the Alzheimer's dataset, we started by examining the data in the file 'demographics-all.xlsx'. This file contains information on the age, sex, and subject group of each patient in the dataset.

	subject	scan-number	HC-AUD-match	subject-group	Study	Diagnosis	Age	age-subject-group	Sex	AUDIT-Total	MMSE
0	002_S_4171	1	NaN	AD-MCI	ADNI	MCI	69.00	elderly	M	NaN	24.0
1	002_S_4229	1	NaN	AD-MCI	ADNI	MCI	66.00	elderly	M	NaN	29.0
2	002_S_4473	1	NaN	AD-MCI	ADNI	MCI	75.00	elderly	M	NaN	27.0
3	002_S_4521	1	NaN	AD-MCI	ADNI	MCI	70.00	elderly	M	NaN	27.0
4	002_S_4799	1	NaN	AD-MCI	ADNI	MCI	68.00	elderly	M	NaN	29.0
...	...	...	...	...	...	...	...	...	...	...	...
310	1125R2	2	NaN	HC	IAM	HC	63.07	mid-life	F	NaN	NaN
311	1126R2	2	NaN	HC	IAM	HC	76.70	elderly	M	NaN	NaN
312	1127R2	2	NaN	HC	IAM	HC	56.98	mid-life	F	NaN	NaN
313	1128R2	2	NaN	HC	IAM	HC	55.37	mid-life	M	NaN	NaN
314	1132R2	2	NaN	HC	IAM	HC	NaN	NaN	F	NaN	NaN

315 rows × 11 columns

Cleaning the dataset-

We used pandas library in Python to load and clean the data a brif summary of what we did is given below-

.

- Drop the columns 'HC-AUD-match', 'AUDIT-Total', and 'MMSE' from the DataFrame, and creates a new DataFrame with the remaining columns.
- Drop any row in the new DataFrame that contains missing values.
- Outputs information on the unique values of several columns in the DataFrame, including 'Study', 'Diagnosis', 'Sex', 'subject-group', and 'age-subject-group'.
- Defines a function to remove outliers from a DataFrame
- Generates a new DataFrame with random data for the columns 'scan-number' and 'Age'.
- Applies the outlier removal function to the 'scan-number' and 'Age' columns in the new DataFrame.

After cleaning the dataset, the datatypes of the columns in the new dataframe 'new_df' were checked. The 'Age' column was a float64, and the rest were either objects or integers.

Outcome after cleaning the dataset-

the result of some data cleaning and transformation operations applied to the original df data frame. The table has the same columns as the original data frame, but with some differences:

- The index values of the rows have been reset to start from 0 to n-1.
- Some missing values (represented as NaN) have been removed or filled with appropriate values, such as the Age column.

- The data in the age-subject-group column has been derived from the Age column by grouping the subjects into age groups.
- The values in the Sex column have been converted to uppercase letters for consistency.

Overall, the outcome table appears to be a cleaned and transformed version of the original data frame that is ready for further analysis or modeling.

The table below shows the dataset after its cleaned-

	subject	scan-number	subject-group	Study	Diagnosis	Age	age-subject-group	Sex
0	002_S_4171	1	AD-MCI	ADNI	MCI	69.00	elderly	M
1	002_S_4229	1	AD-MCI	ADNI	MCI	66.00	elderly	M
2	002_S_4473	1	AD-MCI	ADNI	MCI	75.00	elderly	M
3	002_S_4521	1	AD-MCI	ADNI	MCI	70.00	elderly	M
4	002_S_4799	1	AD-MCI	ADNI	MCI	68.00	elderly	M
...	...	...	...	...	...	...	...	...
309	1124R2	2	HC	IAM	HC	62.85	mid-life	M
310	1125R2	2	HC	IAM	HC	63.07	mid-life	F
311	1126R2	2	HC	IAM	HC	76.70	elderly	M
312	1127R2	2	HC	IAM	HC	56.98	mid-life	F
313	1128R2	2	HC	IAM	HC	55.37	mid-life	M

313 rows × 8 columns

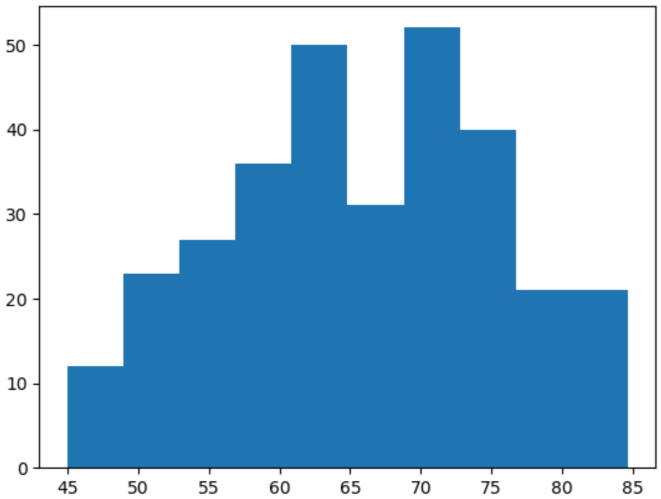
Observations and key predictors-

There are 313 observations and 8 variables in the outcome table. The "scan-number" variable has a mean of 1.37 and a standard deviation of 0.48, indicating that there are two scans per subject and the majority of subjects have only one scan. The "Age" variable has a mean of 65.70 and a standard deviation of 9.78, indicating that the subjects' ages range from

45 to 84 years, with an average age of 65.70 years.

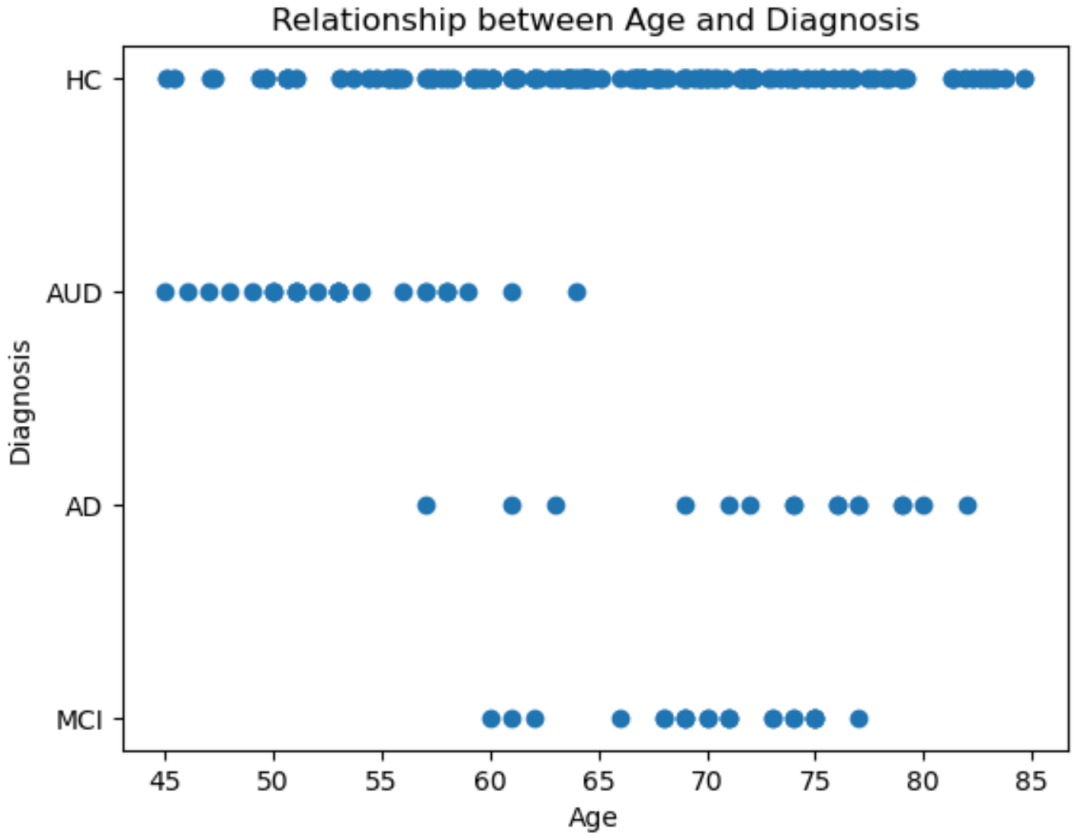
	scan-number	Age
count	313.000000	313.000000
mean	1.373802	65.703003
std	0.484587	9.781700
min	1.000000	45.000000
25%	1.000000	58.020000
50%	1.000000	66.750000
75%	2.000000	73.000000
max	2.000000	84.660000

To visualize the age distribution of the subjects, a histogram was plotted using the matplotlib library. The histogram shows that the majority of subjects are between the ages of 60 and 80 years, with a peak around 70 years of age.

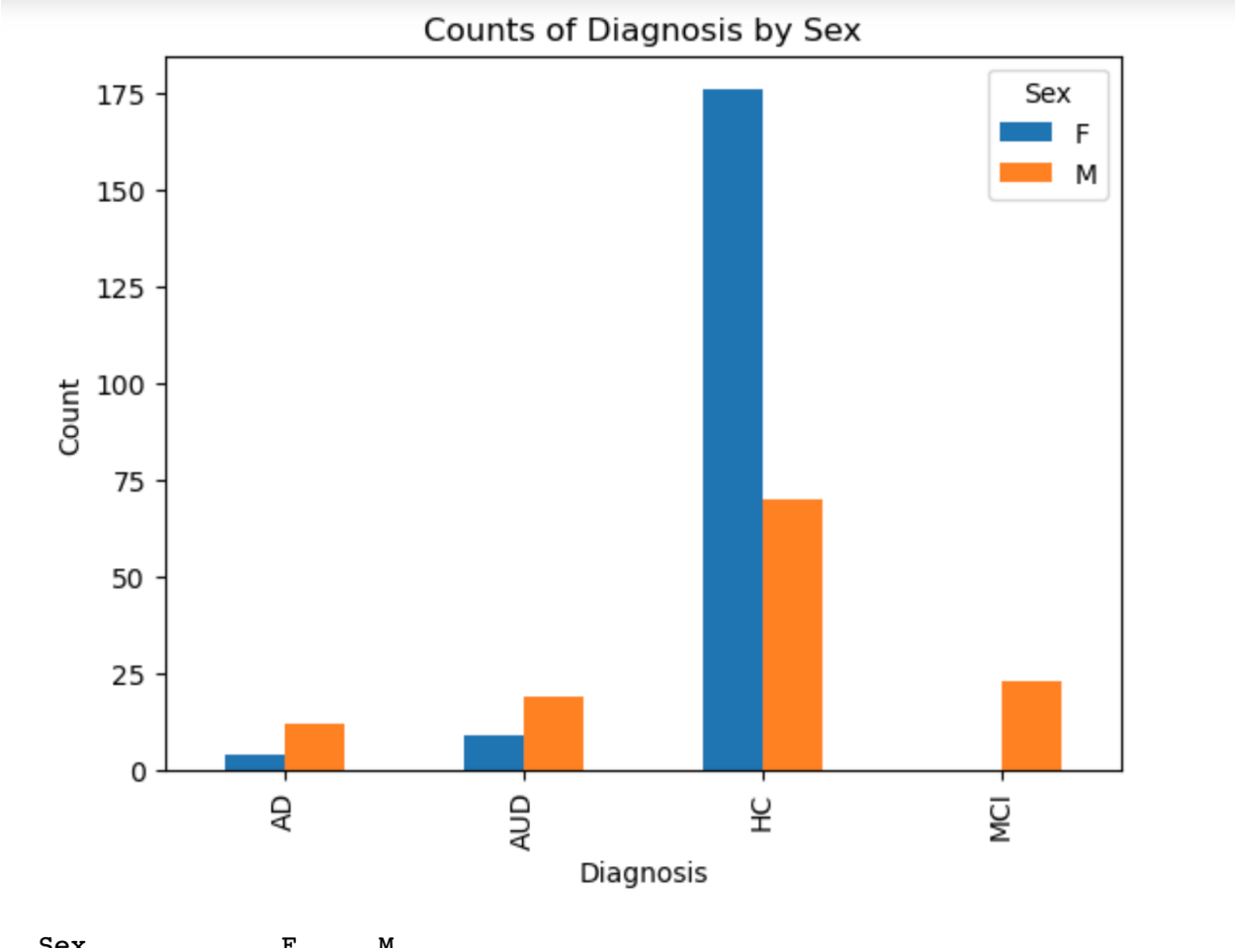


Based on the data, we can use age, sex, and diagnosis as **predictors**. The scan number is not useful as a predictor because it is just an identifier for the scans and does not contain any information that can help predict the diagnosis or other variables of interest.

For this project, we can select "Age" as the predictor variable. We can use a scatter plot to visualize the relationship between "Age" and "Diagnosis". The scatter plot shows that there is no clear relationship between age and diagnosis.



Alternatively, we can select "Sex" as the predictor variable. We can use a bar plot to visualize the relationship between "Sex" and "Diagnosis". The bar plot shows that there are more females than males in the MCI group, while the HC group has more males than females. This suggests that sex may be an important predictor of diagnosis.



Sex F M

Conclusion-

In summary, Age, sex, and diagnosis are the main variables for this study, and the cleaned dataset has been prepared for future analysis. A bar plot between "Sex" and "Diagnosis" suggested that sex may be a significant predictor of diagnosis, however a scatter plot between "Age" and "Diagnosis" did not demonstrate any clear association. Overall, the cleaned dataset offers insightful information about the critical indicators for diagnosing Alzheimer's disease, which can guide future research and modeling.