Dementia Prediction using Machine Learning Algorithms

Hadeel Fahad Altowairqi Email: ha562@exeter.ac.uk

Dementia

Dementia is a mental condition characterized by a steady decline in cognitive processes that can interfere with everyday life tasks such as memory, problem solving, visual perception, and the capacity to focus on a specific task.

Why should anyone care about it?

Dementia is most common in the elderly and there is no one test to determine if someone has dementia. Doctors diagnose dementia based on a careful medical history assessment , a physical examination, laboratory tests, and the characteristic changes in thinking, day-to-day function and behavior associated with it .

Dementia

Dementia is a mental condition characterized by a steady decline in cognitive processes that can interfere with everyday life tasks such as memory, problem solving, visual perception, and the capacity to focus on a specific task.

Aim of this project

This project aims to use machine learning techniques to generate an automatic model able to predect wither the patient has the probability of dementia or not .Based on a set of features provide by clinical test to overcome diagnostic and treatment challenges .

01 Introduction dataset

The data in this project was collected from a public repository Kaggle .

This dataset includes long-term records of 150 participants aged between 60 to 96.

Each participant had multiple scanning sessions, with a minimum interval of one year between each session, resulting in a total of 373 imaging sessions. For each subject's 3 or 4 T1-weighted MRI scans, collected individually in single scan sessions.

Facts about dataset

64 of the included subjects were characterized as demented at the time of their initial visits and remained so for subsequent scans, including 51 individuals with mild to moderate Alzheimer's disease.

Another 14 subjects were characterized as nondemented at the time of their initial visit and were subsequently characterized as demented at a later visit

Features of Dementia dataset

Feature	Description	
'Subject ID', 'MRI ID'	Identification	
Group	(Converted / Demented / Nondemented)	**Target**
Visit	Number of visit	
M/F	Gender (M if Male, F if Female)	**Demographics Information**
Hand	Handedness (all subjects were right-handed)	
Age	Age in years	
EDUC	Years of education	
SES	Socioeconomic Status (as assessed by the Hollingshead Index of Social Position and classified into categories from 1 (highest status) to 5 (lowest status)	

15 features
373 rows

Features of Dementia dataset

Feature	Description	
MMSE	Mini Mental State Examination (range is from 0 = worst to 30 = best)	**Clinical Information**
CDR	Clinical Dementia Rating (0 = no dementia, 0.5 = very mild AD, 1 = mild AD, 2 = moderate AD)	
eTIV	Estimated Total Intracranial Volume	**Derived anatomic volumes**
nWBV	Normalize 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	
'MR Delay'	Delay	

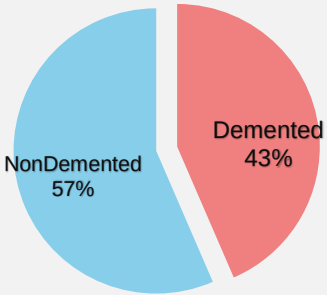
15 features
373 rows

02 Exploratory Data Analysis

Quick insight into Dementia dataset

Summary Description

	Group	Visit	Gender	Age	EDUC	SES	MMSE	CDR	eTIV	nWBV	ASF
count	373.000000	373.000000	373.000000	373.000000	373.000000	373.000000	373.000000	373.000000	373.000000	373.000000	373.000000
mean	0.428954	1.882038	0.428954	77.013405	14.597855	2.436997	27.351206	0.290885	1488.128686	0.729568	1.195461
std	0.495592	0.922843	0.495592	7.640957	2.876339	1.109307	3.675329	0.374557	176.139286	0.037135	0.138092
min	0.000000	1.000000	0.000000	60.000000	6.000000	1.000000	4.000000	0.000000	1106.000000	0.644000	0.876000
25%	0.000000	1.000000	0.000000	71.000000	12.000000	2.000000	27.000000	0.000000	1357.000000	0.700000	1.099000
50%	0.000000	2.000000	0.000000	77.000000	15.000000	2.000000	29.000000	0.000000	1470.000000	0.729000	1.194000
75%	1.000000	2.000000	1.000000	82.000000	16.000000	3.000000	30.000000	0.500000	1597.000000	0.756000	1.293000
max	1.000000	5.000000	1.000000	98.000000	23.000000	5.000000	30.000000	2.000000	2004.000000	0.837000	1.587000



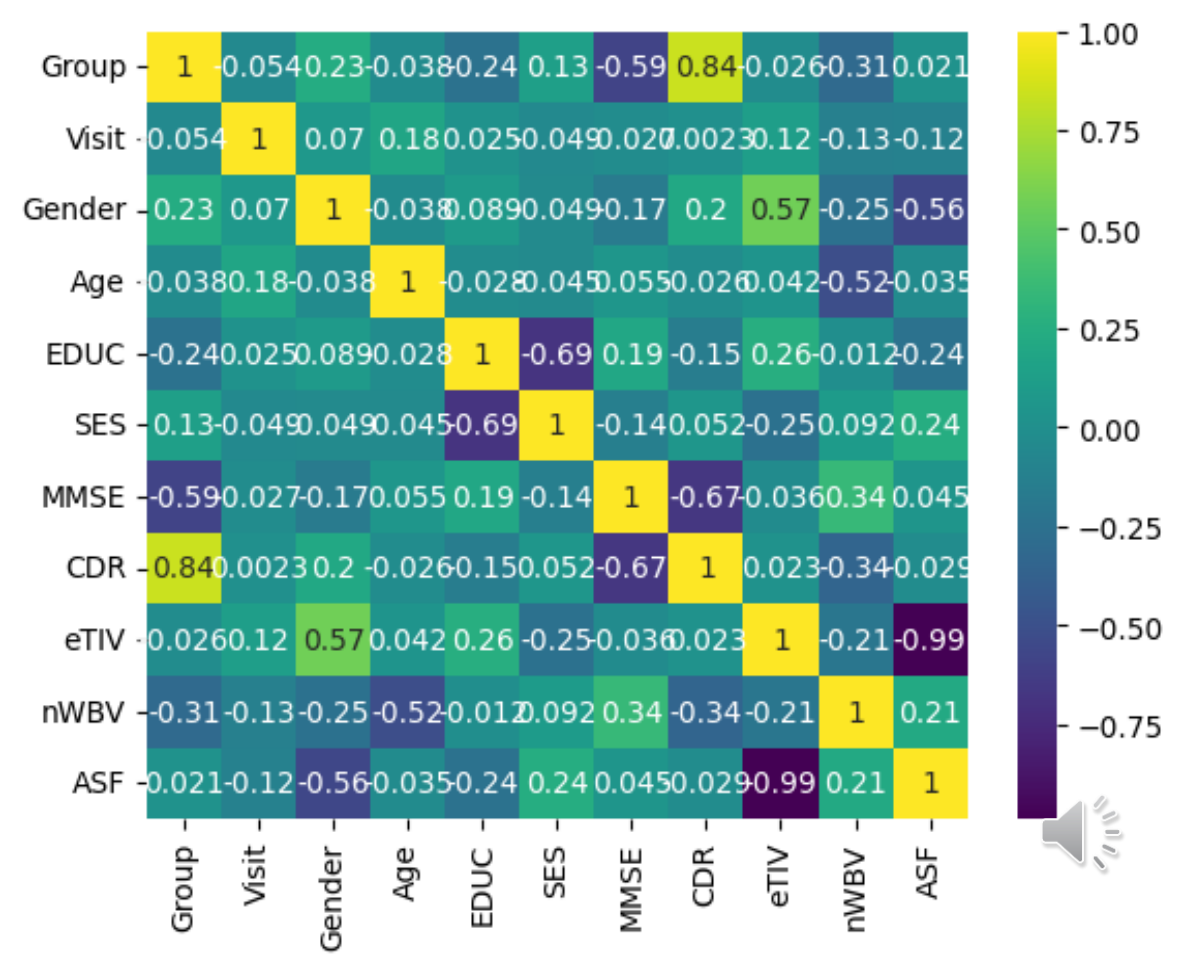
Positive correlation between:

- **CDR** score and **Group** classification [0.84]
- **Gender** and **eTIV** [0.57]

Correlation between Features

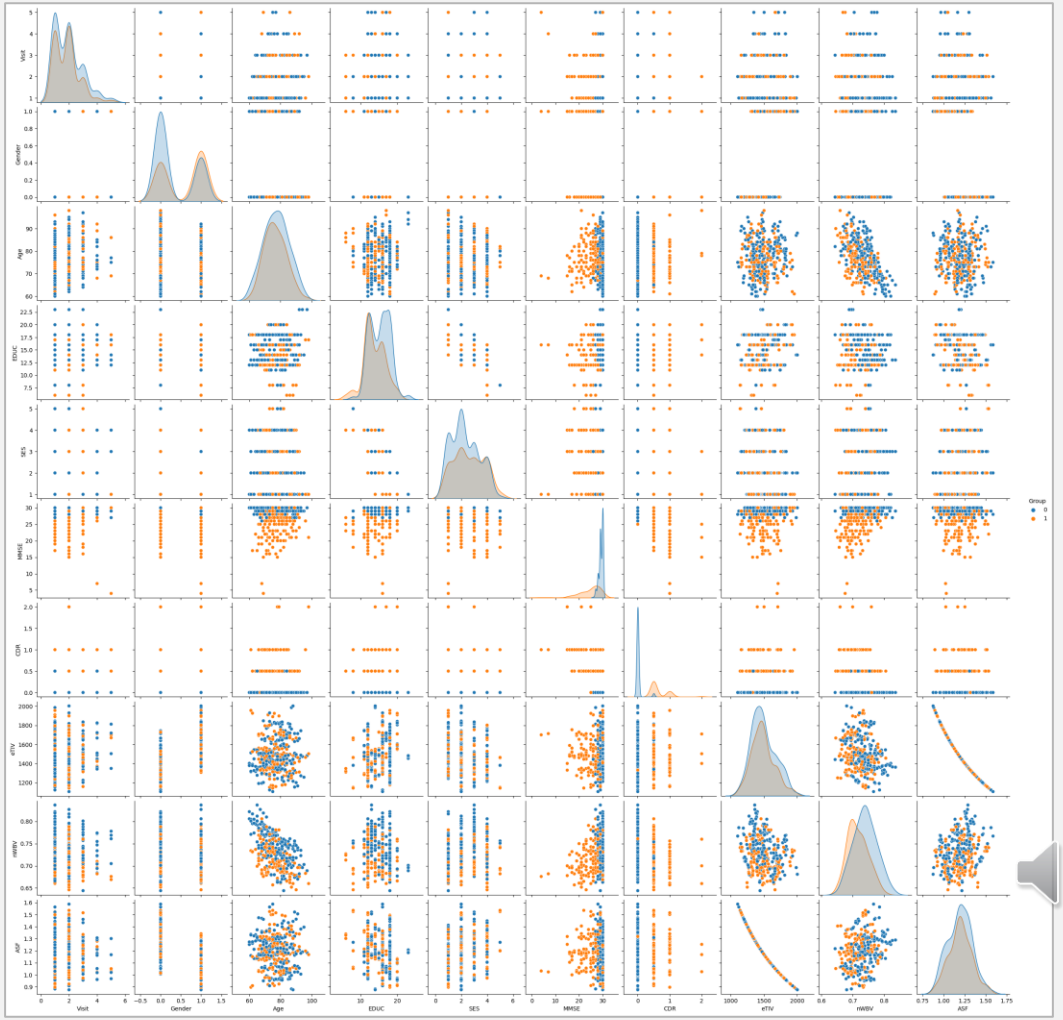
Negative correlation between:

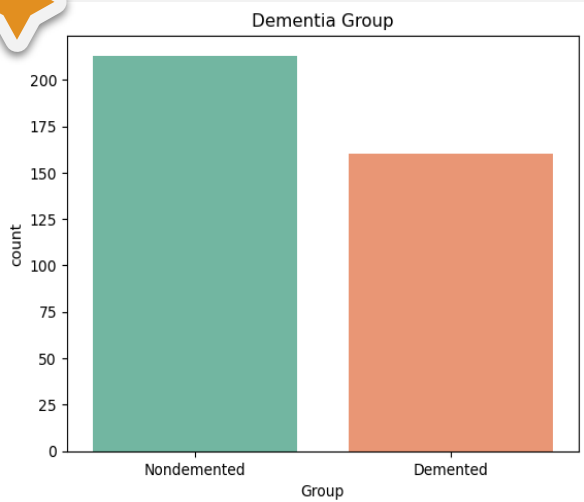
- **Group** and **'MMSE'** [-0.590]
- The **'CDR'** with the **MMSE** [0.674]

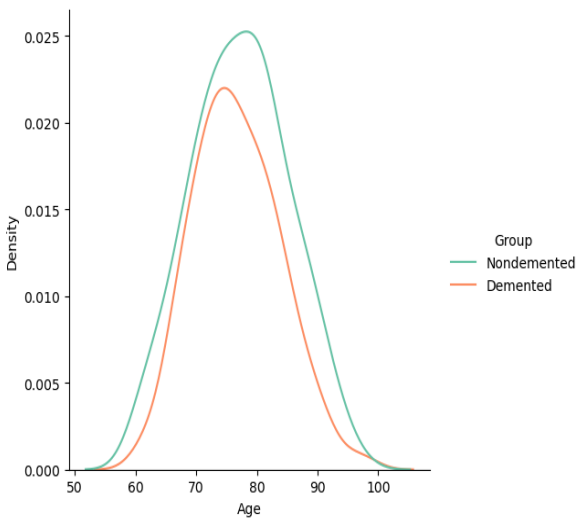


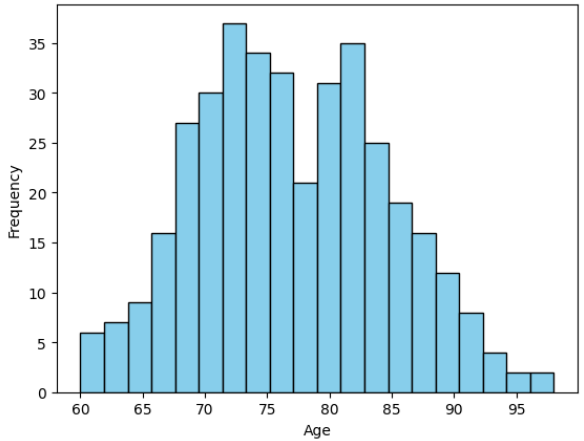
Pair plot

between Features



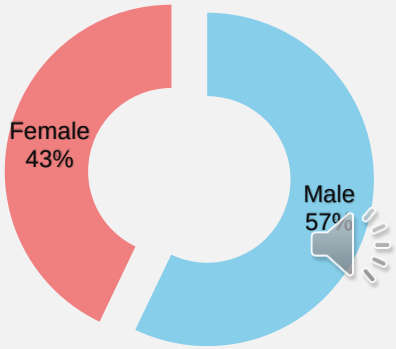


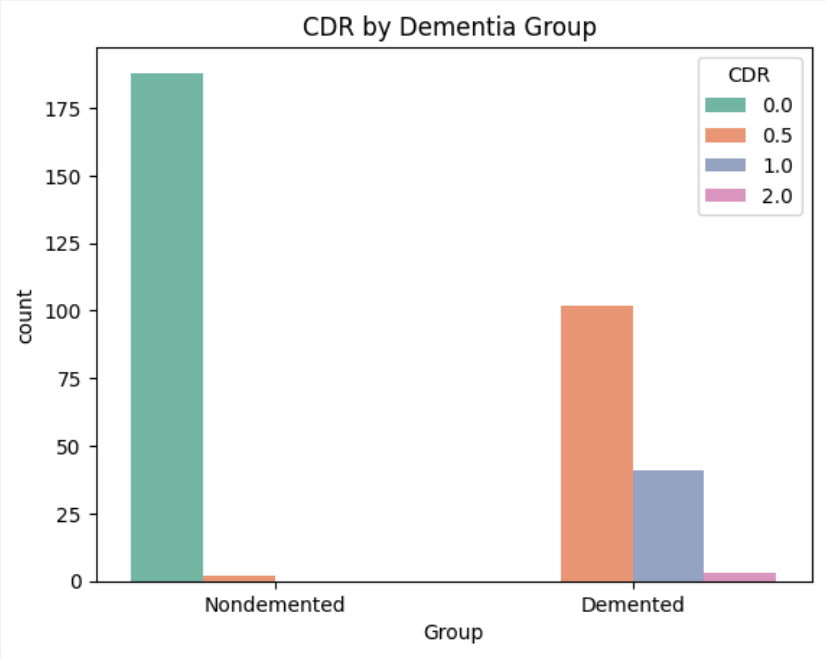


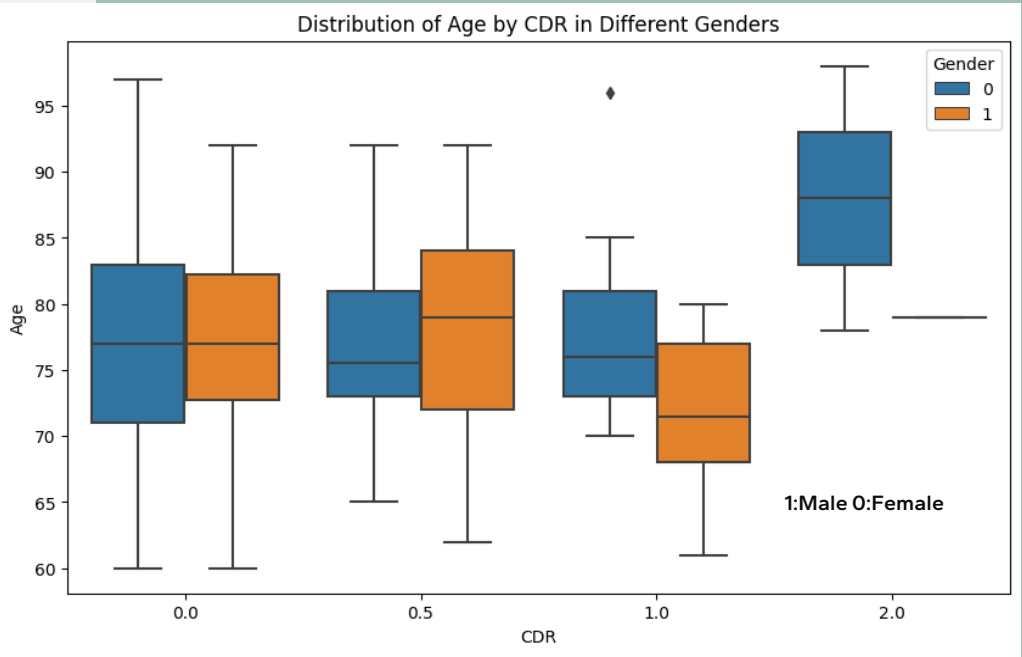


Association of

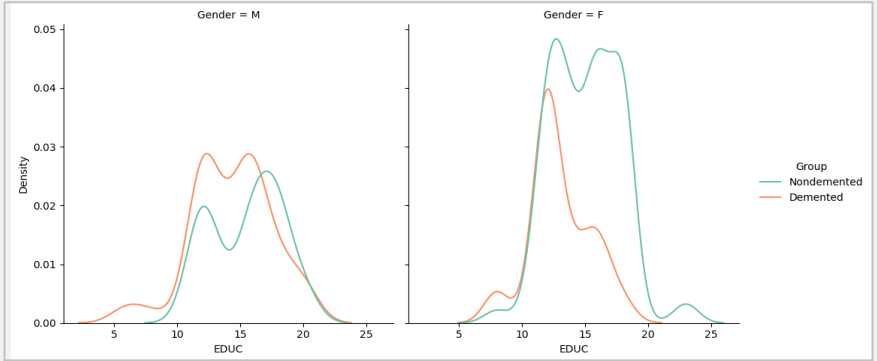
[Age & Gender] with Demented diagnosis

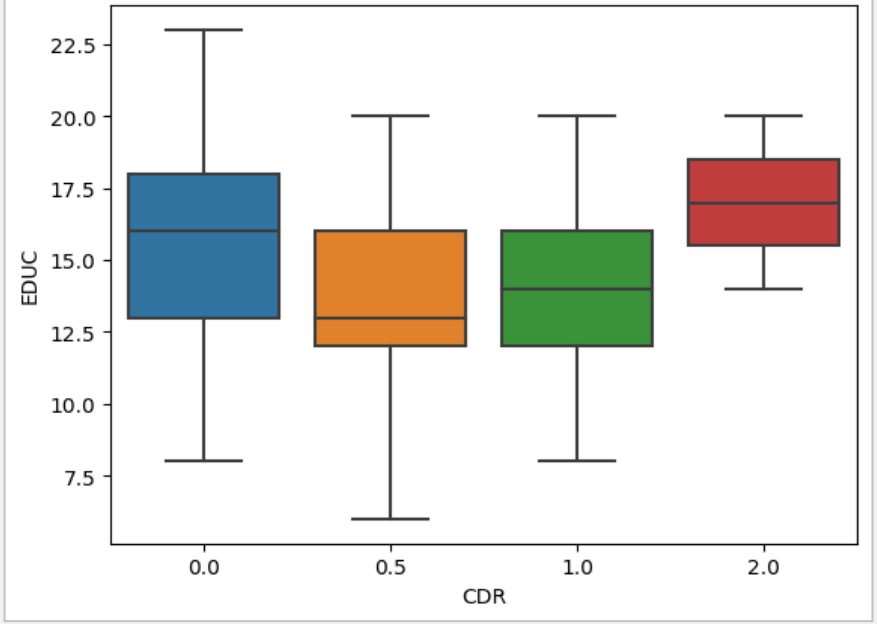


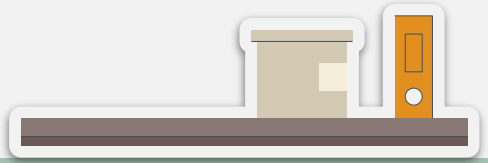




[CDR] Clinical Dementia Rating







[EDUC] Education and Dementia

Data preprocessing

Dementia dataset

Additional columns

The default columns, including Subject ID, MRI ID, and Hand, were removed

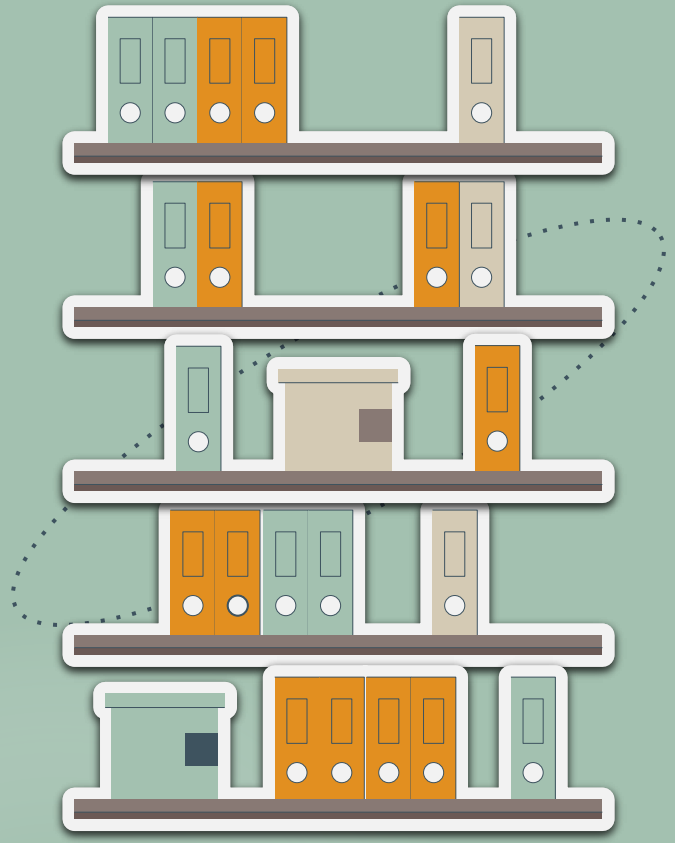
Catagorical columns

Target [Group]
'Nondemented' : 0
'Demented' : 1
Converted :2

Gender = 'M' = 1 and 'F' = 0

Numerical column

Missing values in both 'SES' and 'MMSE' Impute with mode and median respectively



03
Modeling
Classification problem

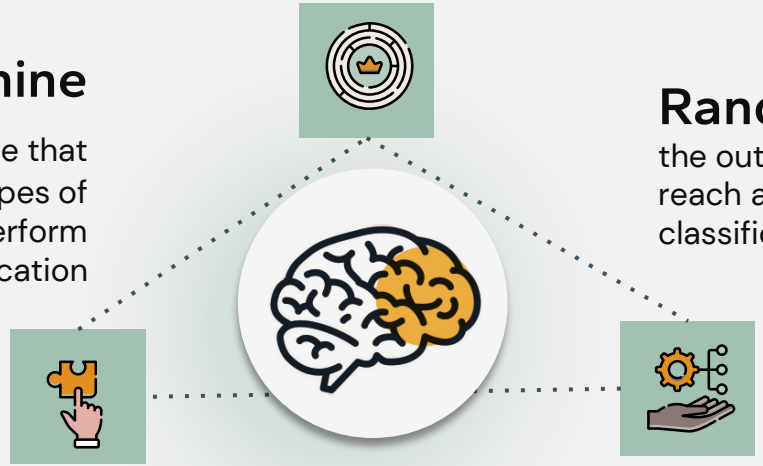
Classification algorithms

Logistic Regression:

It is a classification algorithm estimates the probability of an event occurring based on a given dataset of independent variables

Support Vector Machine (SVM) :

It finds a hyper-plane that creates a boundary between the types of data. Inherently, SVM can only perform binary classification



Random Forest:

combines the output of multiple decision trees to reach a single result. It handles both classification and regression problems.

04 overview of data analysis and results

Evaluation matrix

In the **Dementia dataset** and a predictive model, every data point will be on one of the below four categories.

- **True Positive (TP):** The individual having Demented and is correctly predicted as Demented.
- **True Negative (TN):** The individual not having Demented and was correctly predicted as not Demented.
- **False Positive (FP):** The individual not having Demented, is incorrectly predicted as Demented.
- **False Negative (FN):** The individual having Demented, is incorrectly predicted as not having Demented.

Evaluation report

01

Accuracy

provides an overall measure of how well the model is performing. calculated as : (TP + TN) / (TP + TN + FP + FN).

02

Precision

The ability of the classifier not to label as positive a sample that is negative calculated as : TP / (TP + FP).

03

Recall

The ability of the classifier to find all the positive samples. calculated as: TP / (TP + FN).

04

F1 Score

The relation of precision and recall. calculated as : 2 * (Precision * Recall) / (Precision + Recall)

05

The AUC-ROC

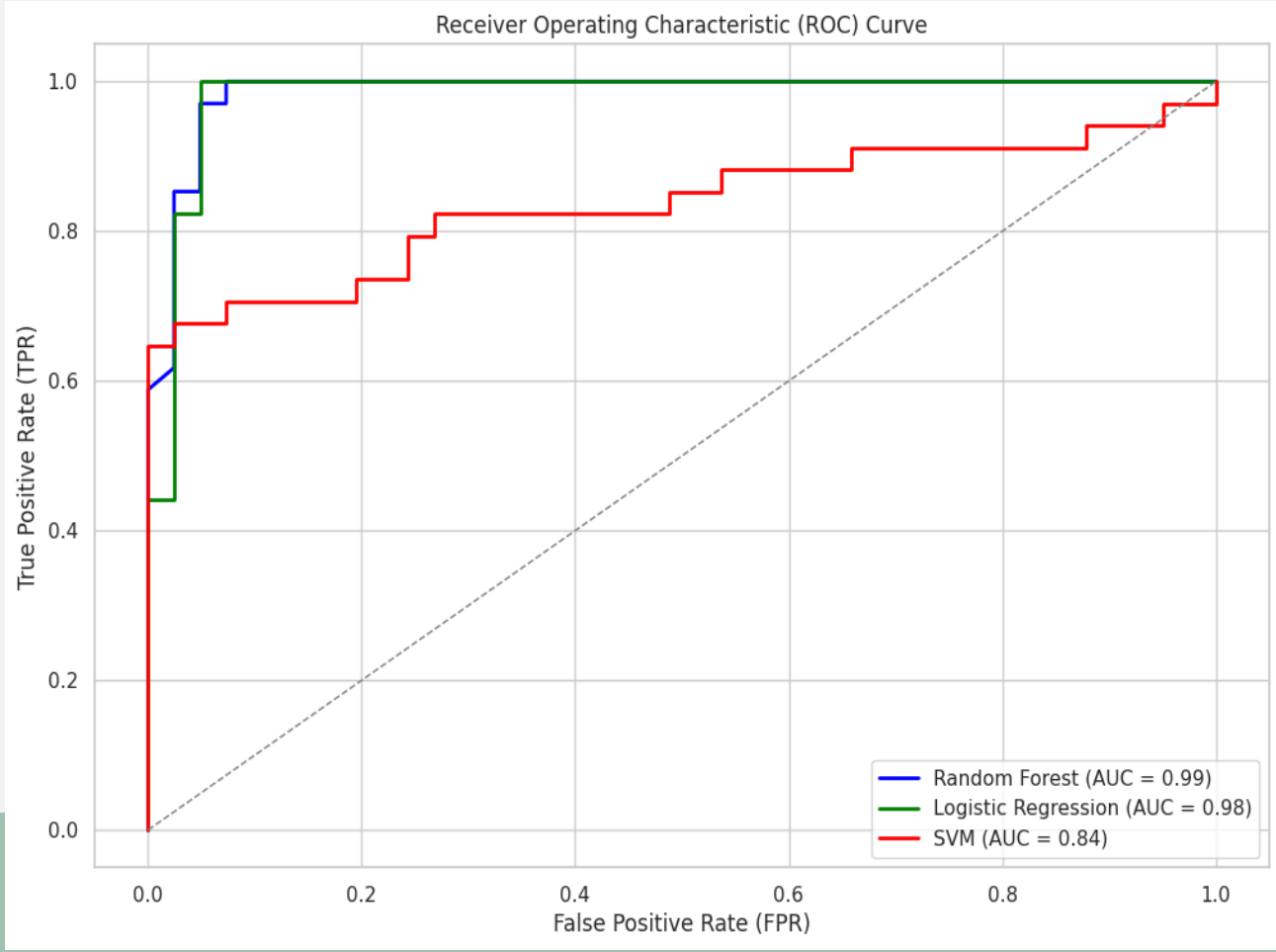
is a scalar value representing the area under the ROC curve. A perfect classifier has an AUC-ROC score of 1.0, while a random classifier has a score of 0.5.

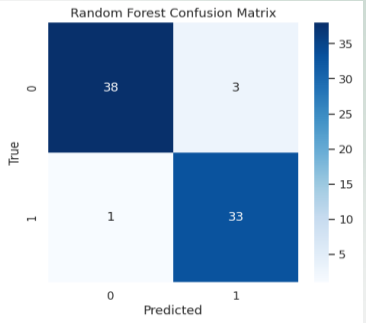
- In classification, we consider all valid attributes of Dementia dataset the following Table presents classification results.

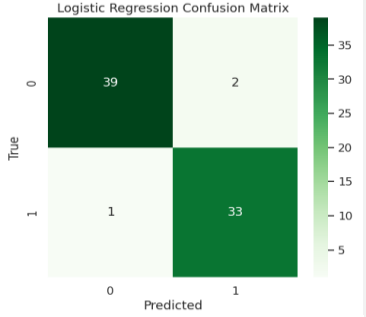
Classifier	Accuracy	Precision	Recall	F1 Score	Accuracy
Logistic Regression	98%	0 = 0.97 1 = 0.94	0.95 0.97	0.96	0.96
Random Forest	99%	0 = 0.97 1 = 0.94	0.93 0.97	0.95 0.94	0.95
Support Vector Machine (SVM)	84%	0 = 0.55 1 = 0.00	0.95 0.00	0.71 0.00	0.55

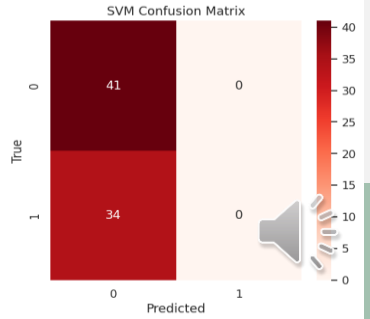
- The results show that 2 out of 3 classifiers demonstrate competitive performance as they result above 96% accuracy for the project datasets.

Case 1

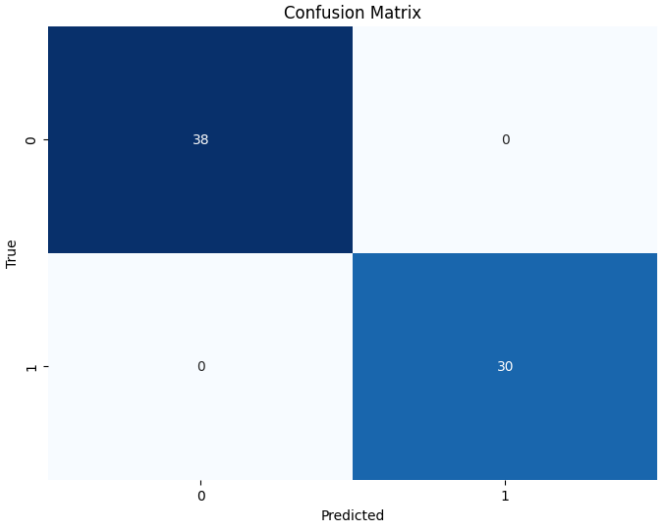

Receiver Operating Characteristic (ROC) Curve
True Positive Rate (TPR)
False Positive Rate (FPR)
0.0
0.2
0.4
0.6
0.8
1.0
Random Forest (AUC = 0.99)
Logistic Regression (AUC = 0.98)
SVM (AUC = 0.84)

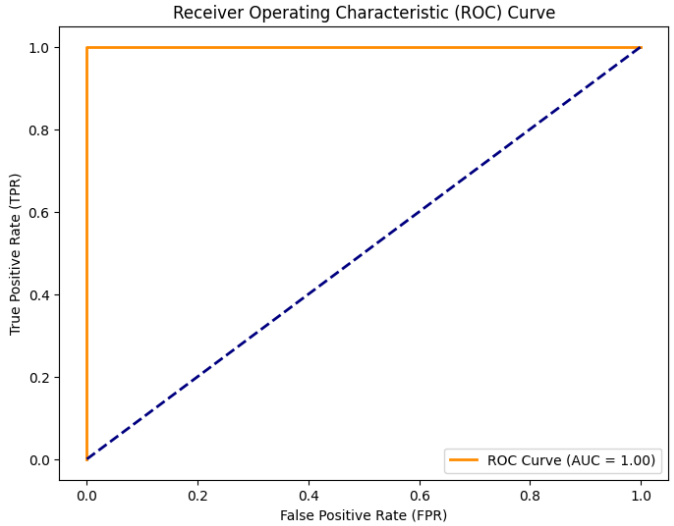

Random Forest Confusion Matrix
38
3
1
33
True
Predicted
0
1


Logistic Regression Confusion Matrix
39
2
1
33
True
Predicted
0
1


SVM Confusion Matrix
41
0
34
0
True
Predicted
0
1

- The result shows that accuracy of Logistic Regression model improve significantly to 100% by using lasso mode.





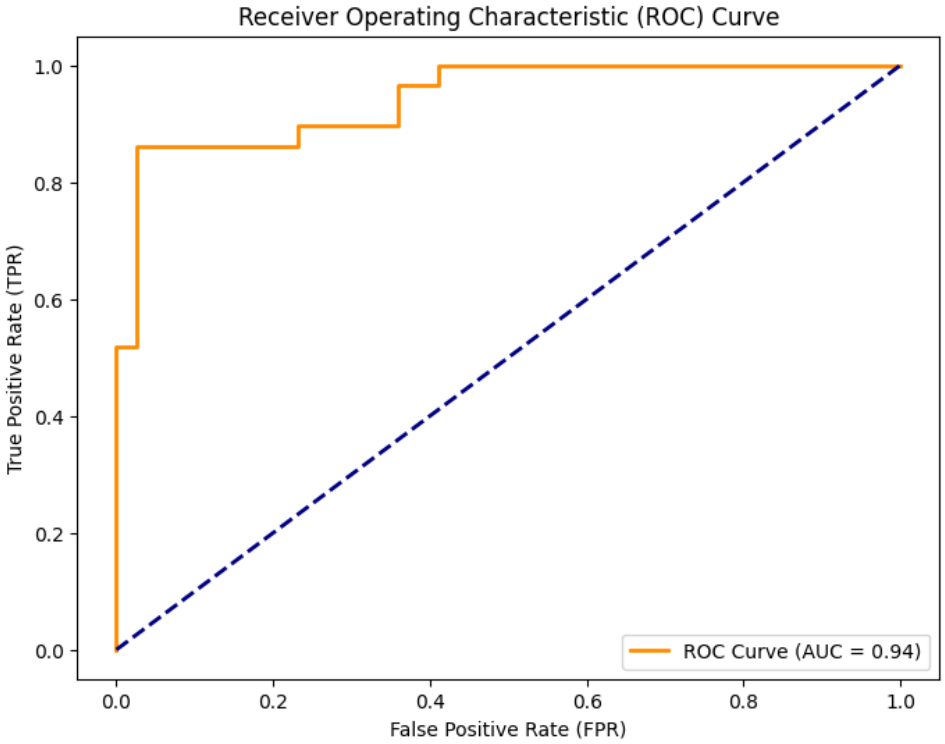
Case 2

Logistic Regression

[Using LASSO]

L1 penalty

- In this case I adjust the parameters of this classification model the accuracy improve to 0.94



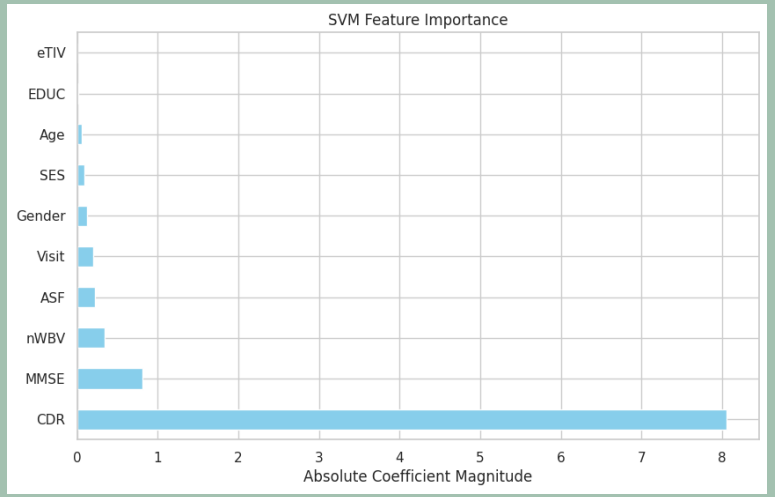
Case 3

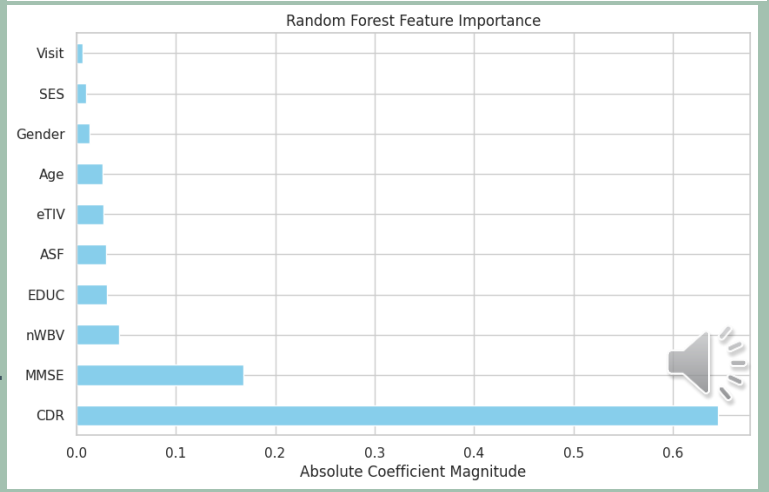
SVM

[Optimal C and Gamma]

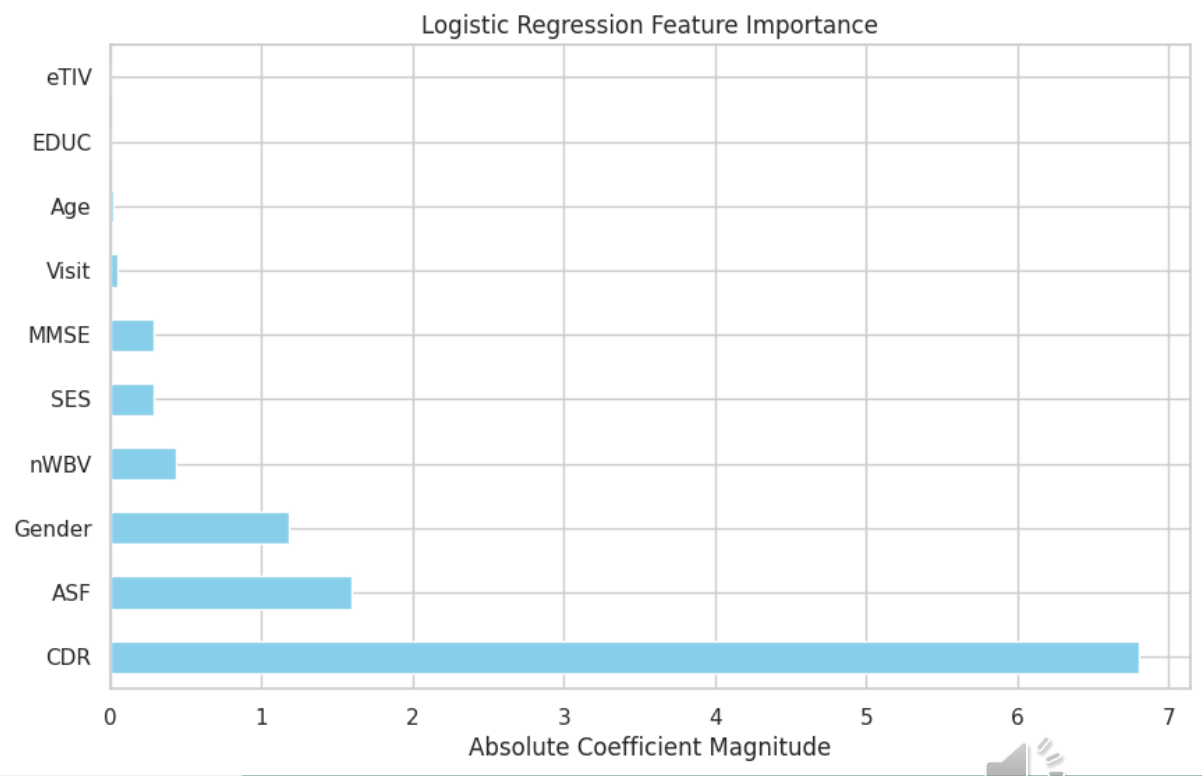
Feature Importance	
Random Forest	**SVC**
CDR 0.645463	CDR 8.056240
MMSE 0.168217	MMSE 0.805382
nWBV 0.042710	nWBV 0.336700
EDUC 0.030589	ASF 0.220431
ASF 0.030107	Visit 0.193837
eTIV 0.027638	Gender 0.122556
Age 0.026204	SES 0.093485
Gender 0.013383	Age 0.052511
SES 0.009527	EDUC 0.015916
Visit 0.006161	eTIV 0.000754

[MMSE', 'CDR', 'nWBV] are the most important features that affect models' accuracy.





Feature Importance
Logistic Regression
CDR 6.801441 ASF 1.592692 Gender 1.177310 nWBV 0.436057 SES 0.291746 MMSE 0.290973 Visit 0.045448 Age 0.017183 EDUC 0.002713 eTIV 0.000307

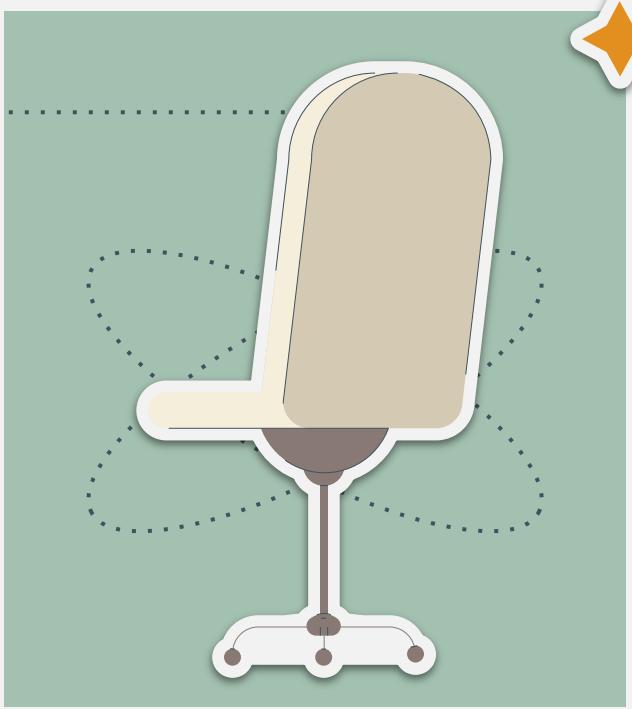


05

Conclusion

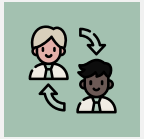
Based on the results and values of performance, it can be concluded that **Random Forest algorithm** can be implemented in real-time scenarios to classify subjects based on diagnoses.

It is crucial to acknowledge that Dementia and Alzheimer's disease are intricate mental conditions, so complete dependence on ML algorithms for diagnosis should complemented by clinical ability to identify potential risk factors and patterns, providing valuable support to healthcare professionals in early detection and risk assessment.



Dataset

When machine learning model applied in real time, the converted category of the group column should be modified manually to either "demented" or "nondemented" depending on the CDR outcome.



Models

From feature importance we can conclude using **Random forest** that the Clinical Dementia Rating **CDR** is the most important factor to generate trustworthy estimation.

06 Limitations

Thanks

Please keep this slide for attribution

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**