HDSC 2022 Winter Internship Premiere Project

CLASSIFICATION & PREDICTION OF DEMENTIA (Team PCA)

Presenters

Omotosho Olamilekan Abdullateef Ogundipe

Project Code **: PP22/H606**

Content Synopsis

1. What is Dementia ?
2. Problem Statement
3. Dataset Description
4. Exploratory Data Analysis (EDA)
5. Feature Engineering
6. Modeling
7. Conclusion

What is Dementia ?

- Not a specific disease, dementia is a group of conditions characterised by impairment of at least two brain functions, such as memory loss and judgement.
- Symptoms include forgetfulness, limited social skills and thinking abilities so imparied that it interferes with daily functioning
- Treatment can help, but this condition can’t be cured.

Problem Statement

- The deterioration of cognitive function of a person is in some way, one of the many syndromes that one wouldn’t want to be diagnosed with.

- The chances that a person will have dementia and the probable type can have is what we aim to solve in our project.

Dataset Description

```
[ ] data.head(5)
```

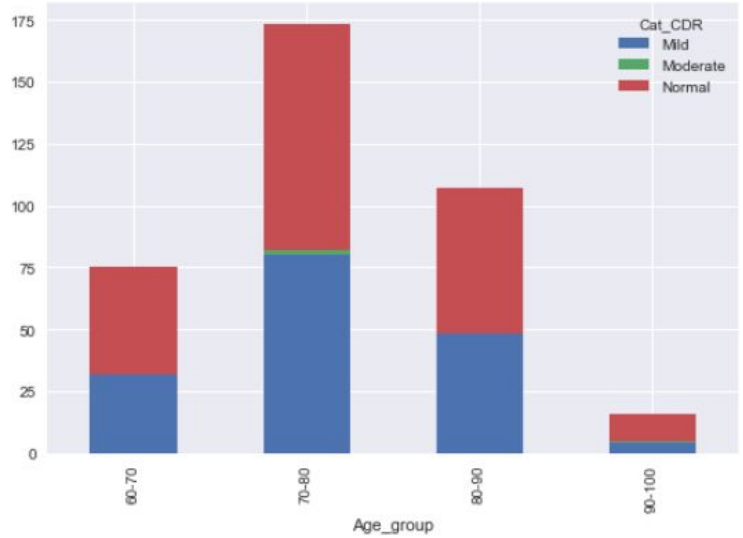
	Subject ID	MRI ID	Group	Visit	MR Delay	M/F	Hand	Age	EDUC	SES	MMSE	CDR	eTIV	nWBV	ASF
0	OAS2_0001	OAS2_0001_MR1	Nondemented	1	0	M	R	87	14	2.0	27.0	0.0	1987	0.696	0.883
1	OAS2_0001	OAS2_0001_MR2	Nondemented	2	457	M	R	88	14	2.0	30.0	0.0	2004	0.681	0.876
2	OAS2_0002	OAS2_0002_MR1	Demented	1	0	M	R	75	12	NaN	23.0	0.5	1678	0.736	1.046
3	OAS2_0002	OAS2_0002_MR2	Demented	2	560	M	R	76	12	NaN	28.0	0.5	1738	0.713	1.010
4	OAS2_0002	OAS2_0002_MR3	Demented	3	1895	M	R	80	12	NaN	22.0	0.5	1698	0.701	1.034

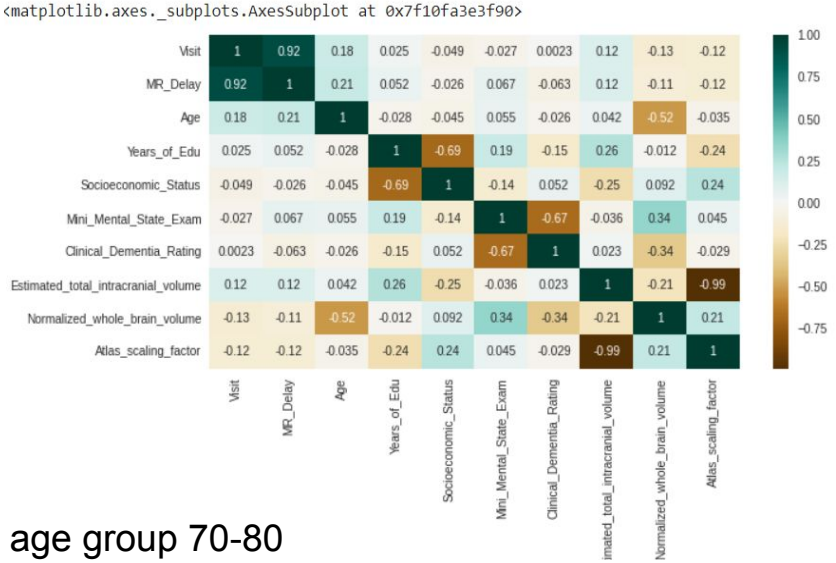
- Dataset Source : Battineni, Gopi; Amenta, Francesco; Chintalapudi, Nalini (2019), “Data for: MACHINE LEARNING IN MEDICINE: CLASSIFICATION AND PREDICTION OF DEMENTIA BY SUPPORT VECTOR MACHINES (SVM)”, Mendeley Data, V1, doi: 10.17632/tsy6rbc5d4.1

```
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 373 entries, 0 to 372
Data columns (total 15 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Subject ID                      373 non-null    object
 1   MRI ID                          373 non-null    object
 2   Group                           373 non-null    object
 3   Visit                           373 non-null    int64
 4   MR_Delay                        373 non-null    int64
 5   Gender                          373 non-null    object
 6   Handedness                      373 non-null    object
 7   Age                             373 non-null    int64
 8   Years_of_Edu                    373 non-null    int64
 9   Socioeconomic_Status            354 non-null    float64
 10  Mini_Mental_State_Exam          371 non-null    float64
 11  Clinical_Dementia_Rating        373 non-null    float64
 12  Estimated_total_intracranial_volume 373 non-null    int64
 13  Normalized_whole_brain_volume   373 non-null    float64
 14  Atlas_scaling_factor            373 non-null    float64
dtypes: float64(5), int64(5), object(5)
memory usage: 43.8+ KB
```

Exploratory Data Analysis

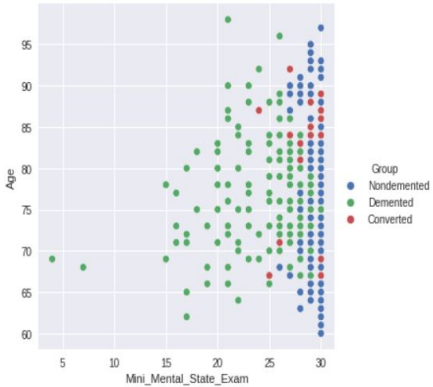




From the plot, the majority of the Dementia cases are in the age group 70-80 years.

Exploratory Data Analysis

<seaborn.axisgrid.FacetGrid at 0x7f10f99b0c50>

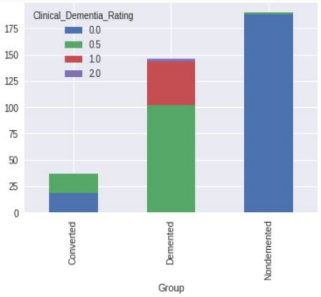


INFERENCE

We can see that most of the non demented individuals have high MMSE score across all ages hence,a good mental state

Also, we have more demented individuals within the age of 75 -85,and their dementia is mild (since MMSE score is within 20-25 here)

[ ]



INFERENCE

from the Table and Visualization

All Demented Subjects have a CDR value of 0.5, 1.0 and 2.0 which is in line with the clinical definition of those CDR values 188 out 0f 190 Non-demented subjects have CDR value of 0! Just 2 Non-demented subject have a CDR value of 0.5, which truly is inconsistent and confusing as a CDR of 0.5 is a clinical indicaion of Very mild dementia (More reason why we should use CDR as our Target variable and Not have both in the Dataset) Conclusions:

Group column should be removed whilst developing the features CDR column should be grouped, the cases having 0 score as Normal and all other score >= 0.5 as dementia CDR should be Used as the Target column

Feature Engineering

- OneHotEncoding for categorical variable
- Creation of cat_CDR column from CDR column,
- Dropping unimportant features (Feature Selection)
- Missing Values Imputation (Median Imputation),
- Resampling and
- Standardization / Normalization (Feature Scaling)

Feature Engineering

```
X_train.cov()
```

	Visit	MR Delay	Age	EDUC	SES	MMSE	eTIV	nWBV	ASF	SES_imputed	MMSE_imputed
Visit	0.811968	515.336271	0.886358	0.087678	-0.042503	-0.104776	20.697015	-0.003238	-0.017401	-0.041828	-0.096604
MR Delay	515.336271	395815.868676	730.822102	97.291427	-3.478181	188.181688	15685.625201	-1.388775	-13.121528	-1.712336	189.603518
Age	0.886358	730.822102	56.575565	-1.316939	-0.309508	1.980039	28.469550	-0.136960	-0.015140	-0.284094	1.952851
EDUC	0.087678	97.291427	-1.316939	8.409565	-2.435173	1.883406	143.208912	-0.001161	-0.106553	-2.259191	1.840339
SES	-0.042503	-3.478181	-0.309508	-2.435173	1.331213	-0.571969	-54.256076	0.003587	0.041696	1.331213	-0.571969
MMSE	-0.104776	188.181688	1.980039	1.883406	-0.571969	14.179375	-12.680486	0.050360	0.014042	-0.516285	14.179375
eTIV	20.697015	15685.625201	28.469550	143.208912	-54.256076	-12.680486	30800.405215	-1.176455	-23.828714	-49.979775	-16.410684
nWBV	-0.003238	-1.388775	-0.136960	-0.001161	0.003587	0.050360	-1.176455	0.001368	0.000923	0.003550	0.050113
ASF	-0.017401	-13.121528	-0.015140	-0.106553	0.041696	0.014042	-23.828714	0.000923	0.018866	0.038284	0.017635
SES_imputed	-0.041828	-1.712336	-0.284094	-2.259191	1.331213	-0.516285	-49.979775	0.003550	0.038284	1.270332	-0.517999
MMSE_imputed	-0.096604	189.603518	1.952851	1.840339	-0.571969	14.179375	-16.410684	0.050113	0.017635	-0.517999	14.104773

We can see that Variable covariance is also changed for the `SES` variable but for the `MMSE` is almost same there is no change

So after all the inferences derived after the imputation . We come to a conclusion to make the additional missing indicator for the variable `SES`

Modeling

- Cross Validation
- Hyperparameter tuning with GridSearch
- Tested different Classification models

Base Model :
→ We have taken Logistic Regression as our Base model

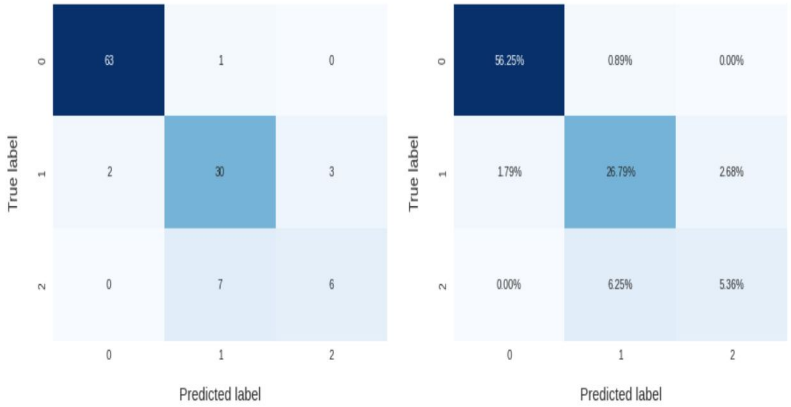
```
clf = [ LogisticRegression(random_state=2), DecisionTreeClassifier(random_state=2), SVC (random_state=2),
       RandomForestClassifier(random_state=2), GradientBoostingClassifier(random_state=2) ]
models = [ 'Logistic Regression', 'Tree', 'Support vector machine', 'RFC', 'Gradient boost' ]

for clf, model in zip(clf,models):
  clf.fit ( X_train_std, y_train )
  y_pred = clf.predict ( X_test_std )
  print ( f'Cross validation score of {model}: %.3f \n' %cross_val_score (clf, X_train_std, y_train, cv=5).mean() )
```

```
Cross validation score of Logistic Regression: 0.836
```

Modeling

Best Fit Model :
→ We have got Random Forest model as our best fit model.



PRECISION, RECALL AND F1-SCORE USING RANDOM FOREST CLASSIFIER (NORMALIZED SET)

```
from sklearn.metrics import classification_report

rfc = RandomForestClassifier()
rfc.fit (X_train_norm, y_train)
print(classification_report(y_test, rfc.predict(X_test_norm)))
```

```
              precision    recall  f1-score   support

           0       0.97      0.98      0.98        64
           1       0.83      0.86      0.85        35
           2       0.73      0.62      0.67        13

    accuracy                           0.90       112
   macro avg       0.84      0.82      0.83       112
weighted avg       0.90      0.90      0.90       112
```

Modeling

Predict probability of having dementia.

we will be using binary classification predict_proba and we will go with the model with the best performance (rfc).

```
rfc.fit(X_train_std, y_train1)  # y_train1 is the second target variable for the binary classification.
rfc.predict_proba(X_test_std)
```

```
array([[0.  , 1.  ],
       [0.06, 0.94],
       [0.99, 0.01],
       [0.98, 0.02],
```

Since our target is (0,1), then the classifier output a probability matrix of dimension (N,2). The first index refers to the probability that the data belong to class 0 (Normal), and the second refers to the probability that the data belong to class 1 (Dementia).

These two would sum to 1.

We can then output the result by:

```
# Probability that a selected patient wll have demenatia.

rfc.predict_proba([X_test.iloc[67,]])[:,1]
```

```
array([0.83])
```

We can change the value (67) to see the change in probability.

Limitation

- Small dataset
- Imbalanced classes

Conclusion

Accuracy may not be a good measure since the dataset is not balanced (classes have different number of data instances). Therefore, we use F1 score which is the weighted average of precision and recall.

The best fit model i.e Random Forest Classifier gives 0.90 F1 score which better than the base model.

Abdulwasiu Tiamiyu (PL)
V Sreenidhi Reddy (APL)
Omotosho Olamilekan (QA)
Abdullateef Ogundipe
Abdulsalam Suleiman Abiodun
Abiodun-Olojede Joshua
Amit Budhiraja
Abiola Ogunbajo
Nofisat Abiodun Ayanlola
Onanuga Damilola Daniel
Oyentunji Rebecca
Idaresit Okposen
Usman Sada
Mutala Sibdoo
Dolapo Towolawi

Team Members

Thank You