Predicting Dementia

Machine Learning for Medicine

Caro Alexandre/Balendran Alan

March 31, 2020

What is dementia ?

- **Dementia** is a syndrom which is characterize by a degradation of memory,thinking, behaviour, and the capability of doing daily activities.
  Even tho dementia is mainly observed in eldery people, it can touch other people.
- We count 50 millions people affected by dementia and we account 10 millions new cases each year.
- Alzheimer's disease is the main case of dementia ,60-70% of all the cases are annotated as dementia.
- Dementia is one the main cause of handicap and dependance among elderies.
- Dementia has physical,psychological,social and economical consequences both for the sick ones and for the entourage.

Dataset Presentation

The dataset is taken from *http://www.oasis-brains.org/*

- Dataset of ***373*** observations
- 150 subjects aged 60 to 96. Each subject was scanned on two or more visits, separated by at least one year for a total of 373 observations
- 72 of the subjects were characterized as **nondemented** 64 of the included subjects were characterized as **demented**. Another 14 subjects labelled as **converted** were characterized as nondemented at the time of their initial visit and were subsequently characterized as demented at a later visit.

Features Presentation

Demographic Features

- **M/F**: Gender
- **Age**
- **EDUC**: years of education.
- **SES**: Socio-Economic Status.
- **Hand** : Right/Left Hand

Features Presentation

Clinical Features

- MMSE (Mini Mental State Examination score):from 0 to 30.
- CDR(Clinical Dementia Rating): from 0 to 2.

Features Presentation

Anatomic Features

- **eTIV**: Estimated Intracranial Volume.
- **nWBV**:Normalized Whole-Brain Volume.Percentage of pixels labelled as gray or white matter on the MRI.
- **ASF**: Atlas Scaling Factor. Scale factor allowing to resize the skull size with respect to the volume of the brain in the skull.

Goals

- Firstly we will try to predict the state of dementia of a person,which is encoded in the feature :***’Group’***
- We will see that the previous task is rather easy and that is why we will move on to predict the feature ’***CDR’***.

Data Preprocessing

Removing features

- The feature **’Hand’** is removed because as we can see,only right handed are represented.

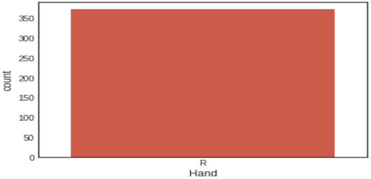
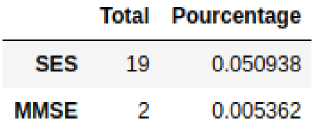


Figure: Right/Left Handed count

- We also remove **’MRI ID’** and **’Subject ID’**, the first one because it has unique value for each observations and the second one due to the fact that we consider each MRI session as a unique observation therefore that feature is not needed.

Data Preprocessing

Missing value



	Total	Pourcentage
SES	19	0.050938
MMSE	2	0.005362

Figure: Missing values on the dataset

- **SES** is filled with the median.
- **MMSE** is filled with the value of the previous observation,since the observations are arranged patient by patient,giving the value of the previous observations is equivalent as filling the MMSE value with the value of the same patient but taken at the previous MRI session

Data Preprocessing

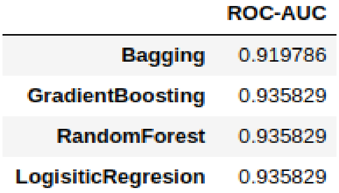
Target

- Since we take each patients MRI session as unique,we get rid of the temporality therefore we binarize our target,a person labelled as **’Converted’** will be considered as ’Demented’.

Modelling

4 models will be used namely:

- Bagging
- Gradient Boosting
- Random Forest
- Logisitic Regression



	ROC-AUC
Bagging	0.919786
GradientBoosting	0.935829
RandomForest	0.935829
LogisiticRegresion	0.935829

Figure: Model Performance

Those are scores on the test data, a train/test split was performed prior to modelling

Conclusion on Group prediction

- We can clearly see that predicting the Dementia state is rather easy without even tuning the hyper-parameters of our models.
- Moreover using *Recursive feature elimination* we can note that **’CDR’** feature is enough in order to well predict ’Group’
- Naturally our goal is now to predict **’CDR’**.

CDR prediction

We do the same preprocessing we did for Group prediction but we just change the target.
We plot **CDR** with respect to other features .

Visualisation

CDR against MMSE

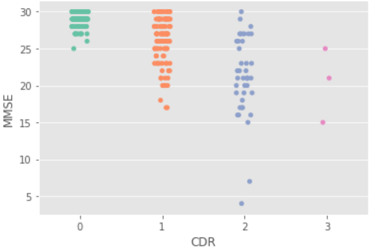


Figure: MMSE with respect to CDR

We see that people in the class 0 and 1 have a **MMSE** value *lower* than the two other group.

Visualisation

ASF against MMSE

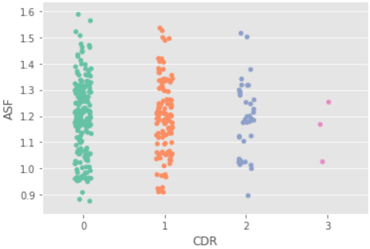


Figure: ASF with respect to CDR

We notice that the **ASF** is rather uniformly distributed in each classes and also we remark that the last class is under-represented.

Visualisation

AGE against MMSE

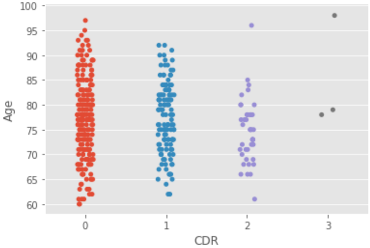
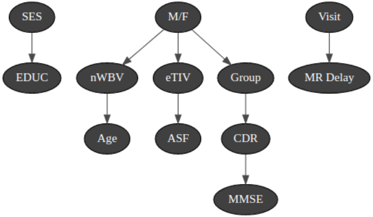


Figure: Age with respect to CDR

The age also seems not to be correlated with **CDR** exept for the last class where older people are more likely to be diagnosed as demented but since we don't have much data for the last class we can't really conclude.

Features selection

- As for Group prediction,we use Recursive Feature Elimination and we note that all the features are kept
- A bayesian net is also performed in order to comprehend the relation between the different features



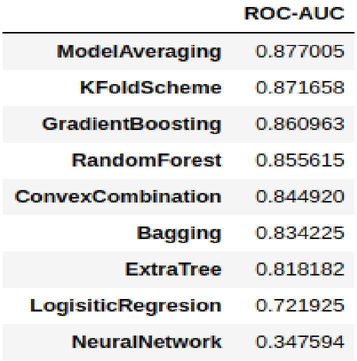
Modelling

On top of the models we saw earlier we will also use:

- ***Extra Tree*** (variant of Random Forest)
- ***Model Averaging*** We takes various models and take as predictions a majority vote
- ***Convex Combination*** Similar to the Model Averaging except we give differents weights to each classifier
- ***Kfold Scheme*** We take one classifier and we train it on differents subsamples of the dataset,we then take a majority vote,this model is robust to outliers
- ***Neural Network*** with 2 hiddens layers

Remark:See the notebook for the different implementation of those models

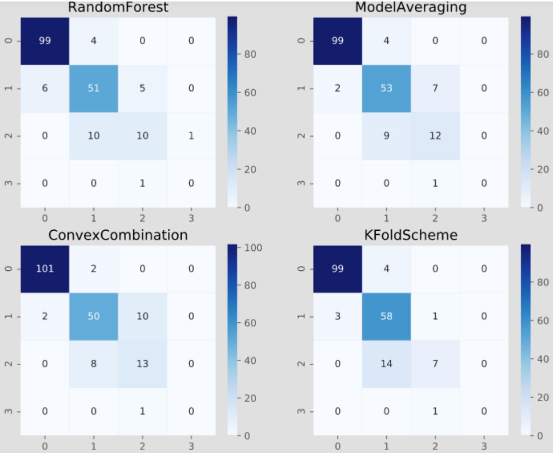
Performance



	ROC-AUC
ModelAveraging	0.877005
KFoldScheme	0.871658
GradientBoosting	0.860963
RandomForest	0.855615
ConvexCombination	0.844920
Bagging	0.834225
ExtraTree	0.818182
LogisiticRegresion	0.721925
NeuralNetwork	0.347594

We clearly see that the **averaging model** yields good results followed by the **KFoldScheme.**

Confusion Matrix



Impediements

- MRI file were provided but they were at the same format for each patient therefore could be used in our studies.
- An another thing was that almost no data prepocessing was needed,therefore we had to work on ’Raw’ data,no new features could have been created.
- Unsupervised learning didn’t yield promising results.
- The size of dataset behind too low.

Conclusion

We learned during this project the different steps from analysing a dataset to modelling according to our dataset problematic. We also learned the difficulties brought by medical dataset which is its low dimension. We applied different methods that we saw during our courses in order to get the most informations out of the dataset.