**VIỆN HÀN LÂM KHOA HỌC VÀ CÔNG NGHỆ VIỆT NAM**

**HỌC VIỆN KHOA HỌC VÀ CÔNG NGHỆ**



**BÁO CÁO TIỂU LUẬN KẾT THÚC HỌC PHẦN THẠC SĨ**

**HỆ HỖ TRỢ QUYẾT ĐỊNH TIÊN TIẾN**

**ĐỀ TÀI:**

**DETECTING EARLY ALZHEIMER'S USING MRI DATA AND MACHINE LEARNING**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | **Học viên thực hiện** | **: HOÀNG HẢI LONG** | | **Giảng viên** | **: TS Vũ Văn Hiệu** | | | **Khoa** | **: CNTT & viễn thông** | | | **Lớp** | **: ITT2022A** | | |  |
| **Hà Nội – 2023** |  |
| **PHIẾU CHẤM ĐIỂM**  Học viên thực hiện:   |  |  |  | | --- | --- | --- | | **Họ và tên** | **Chữ ký** | **Ghi chú** | | Hoàng Hải Long |  |  |   Giảng viên chấm:   |  |  |  | | --- | --- | --- | | **Họ và tên** | **Điểm** | **Chữ ký** | | Giảng viên chấm 1: |  |  | | Giảng viên chấm 2: |  |  | |  |

# 

**DETECTING EARLY ALZHEIMER'S (PHÁT HIỆN SỚM ALZHEIMER)**

**PHÁT HIỆN SỚM ALZHEIMER SỬ DỤNG DỮ LIỆU MRI VÀ MÁY HỌC**

***1. PROBLEM STATEMENT (Mô tả vấn đề)***

+ Bệnh Alzheimer (AD) là một rối loạn thoái hóa thần kinh không rõ nguyên nhân và cơ chế bệnh sinh, chủ yếu ảnh hưởng đến người lớn tuổi và là nguyên nhân phổ biến nhất của chứng mất trí nhớ.

+ Biểu hiện lâm sàng sớm nhất của AD là suy giảm trí nhớ có chọn lọc và trong khi các phương pháp điều trị có sẵn để cải thiện một số triệu chứng, hiện tại không có cách chữa trị nào.

+ Chụp ảnh não qua chụp cộng hưởng từ (MRI), được sử dụng để đánh giá bệnh nhân nghi ngờ mắc AD.

+ Kết quả MRI bao gồm cả sự co rút cục bộ và tổng quát của mô não.

+ Một số nghiên cứu đã gợi ý rằng các tính năng của MRI có thể dự đoán tốc độ suy giảm của AD và có thể hướng dẫn điều trị trong tương lai

+ Tuy nhiên, để đạt được giai đoạn đó, các bác sĩ lâm sàng và nhà nghiên cứu sẽ phải sử dụng các kỹ thuật học máy có thể dự đoán chính xác tiến trình của bệnh nhân từ suy giảm nhận thức nhẹ đến sa sút trí tuệ.

+ Chúng tôi đề xuất phát triển một mô hình âm thanh có thể giúp các bác sĩ lâm sàng làm điều đó và dự đoán sớm bệnh Alzheimer.

***2. DATA (Dữ liệu)***

Nhóm đã tìm thấy dữ liệu liên quan đến MRI được tạo bởi dự án Chuỗi nghiên cứu hình ảnh truy cập mở (OASIS) có sẵn trên trang web của họ và kaggle có thể được sử dụng cho mục đích đào tạo các mô hình máy học khác nhau để xác định bệnh nhân nhẹ. đến chứng mất trí vừa phải.

*2.A DATASET DESCRIPTION (Mô tả dữ liệu)*

+ Chúng tôi sẽ sử dụng dữ liệu MRI theo chiều dọc.

+ Bộ dữ liệu bao gồm dữ liệu MRI theo chiều dọc của 150 đối tượng từ 60 đến 96 tuổi.

+ Mỗi đối tượng được quét ít nhất một lần.

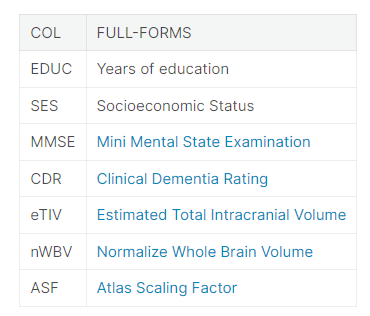
+ Mọi người đều thuận tay phải.

+ 72 trong số các đối tượng được nhóm là 'Không phân biệt' trong suốt nghiên cứu.

+ 64 đối tượng được xếp vào nhóm 'Mất trí nhớ' vào thời điểm họ đến thăm lần đầu và duy trì như vậy trong suốt quá trình nghiên cứu.

+ 14 đối tượng được nhóm là 'Không mất trí nhớ' vào thời điểm họ đến thăm lần đầu và sau đó được mô tả là 'Mất trí nhớ' ở lần thăm khám sau. Chúng thuộc danh mục 'Đã chuyển đổi'.

*2.B COLUMN DESCRIPTORS (Mô tả cột)*



***3. RELATED WORK (Công việc liên quan)***

Ấn phẩm ban đầu chỉ thực hiện một số khám phá sơ bộ về dữ liệu MRI vì phần lớn công việc của họ tập trung vào việc thu thập dữ liệu. Tuy nhiên, trong thời gian gần đây đã có nhiều nỗ lực được thực hiện để phát hiện sớm bệnh nhân Alzheimer bằng cách sử dụng dữ liệu MRI. Một số tác phẩm đã được tìm thấy trong các tài liệu như sau:

1) Khung học máy để dự đoán chuyển đổi bệnh Alzheimer sớm dựa trên MRI trong các môn học MCI. [3]

Trong bài báo này, các tác giả đã quan tâm đến việc xác định suy giảm nhận thức nhẹ (MCI) như một giai đoạn chuyển tiếp giữa suy giảm nhận thức liên quan đến tuổi tác và bệnh Alzheimer. Nhóm đề xuất một nhà sản xuất sinh học mới dựa trên MRI mà họ đã phát triển bằng cách sử dụng các kỹ thuật máy học. Họ đã sử dụng dữ liệu có sẵn từ Cơ sở dữ liệu ADNI của Sáng kiến hình ảnh thần kinh về bệnh Alzheimer. Bài báo tuyên bố rằng dấu ấn sinh học tổng hợp của họ đã đạt được diện tích xác thực chéo gấp 10 lần dưới điểm số đường cong (AUC) là 0,9020 khi phân biệt giữa MCI lũy tiến (pMCI) và MCI ổn định (sMCI).

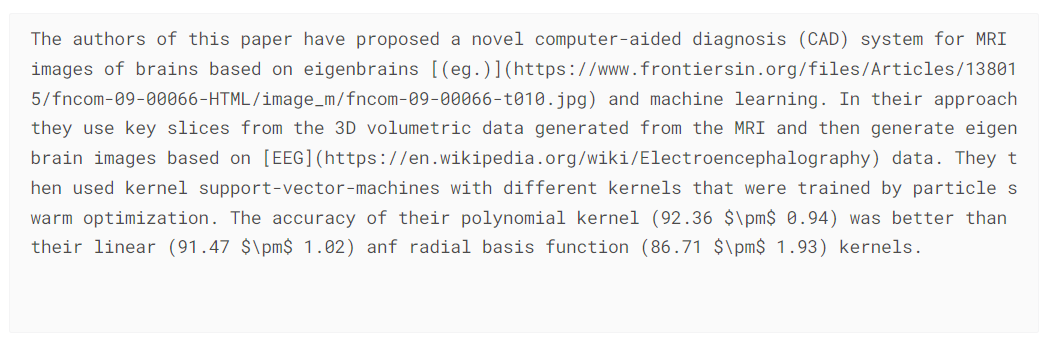
Kỹ thuật đáng chú ý:

1. Học bán giám sát dựa trên dữ liệu có sẵn từ bệnh nhân AD và các biện pháp kiểm soát bình thường, không sử dụng bệnh nhân MCI, để giúp phân loại sMCI/pMCI. Đã thực hiện lựa chọn tính năng bằng cách sử dụng hồi quy logistic chính quy.

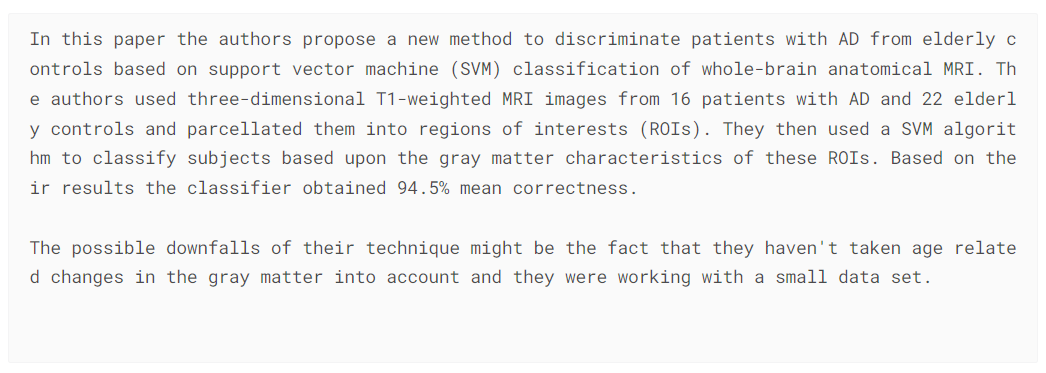
2. Họ đã loại bỏ các hiệu ứng lão hóa khỏi dữ liệu MRI trước khi đào tạo bộ phân loại để ngăn chặn sự nhầm lẫn có thể xảy ra giữa những thay đổi do AD và những thay đổi do lão hóa bình thường.

3. Finally constructed an aggregate biomarker by first learning a separate MRI biomarker and then combining age and cognitive measures about MCI subjects by applying a random foresst classifier.

2. Phát hiện các đối tượng và vùng não liên quan đến bệnh Alzheimer bằng cách sử dụng quét 3D MRI dựa trên não bản địa và học máy. [4]



3. Hỗ trợ phân loại bệnh Alzheimer dựa trên máy vector từ MRI giải phẫu toàn bộ não. [5]

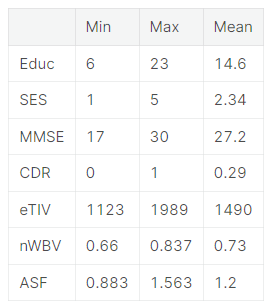


Chúng tôi đã mô tả 3 bài báo ở đây mà chúng tôi thấy thú vị nhất, tuy nhiên, có một số bài khác đã khám phá cùng một câu hỏi. Bất chấp điều đó, điều đáng nói là các bài báo trên đang khám phá dữ liệu MRI thô và mặt khác, chúng tôi đang xử lý 3 đến 4 dấu ấn sinh học được tạo ra từ hình ảnh MRI.

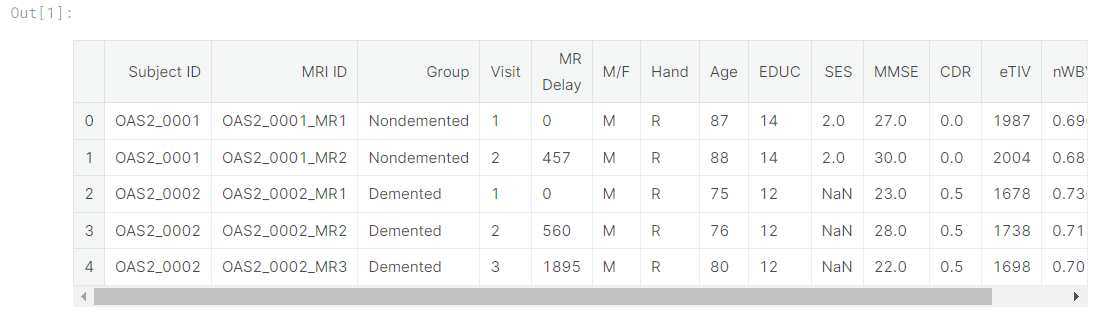
***4. EXPLORATORY DATA ANALYSIS (EDA) (Phân tích dữ liệu khám phá)***

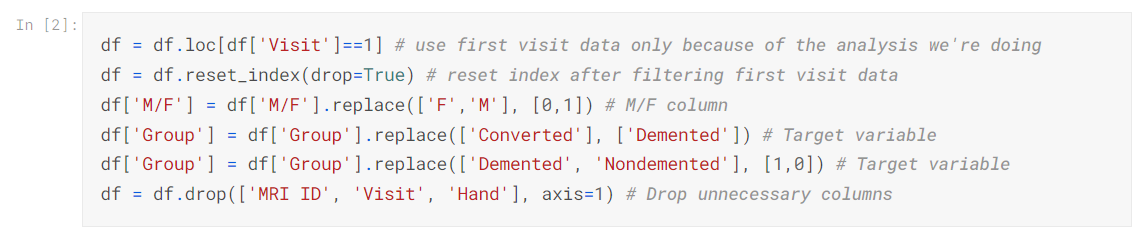
Trong phần này, chúng tôi tập trung khám phá mối quan hệ giữa từng đặc điểm của xét nghiệm MRI và chứng sa sút trí tuệ của bệnh nhân. Lý do chúng tôi tiến hành quy trình Phân tích dữ liệu khám phá này là để nêu rõ mối quan hệ của dữ liệu thông qua biểu đồ để chúng tôi có thể giả định các mối tương quan trước khi trích xuất dữ liệu hoặc phân tích dữ liệu. Nó có thể giúp chúng ta hiểu được bản chất của dữ liệu và lựa chọn phương pháp phân tích thích hợp cho mô hình sau này.

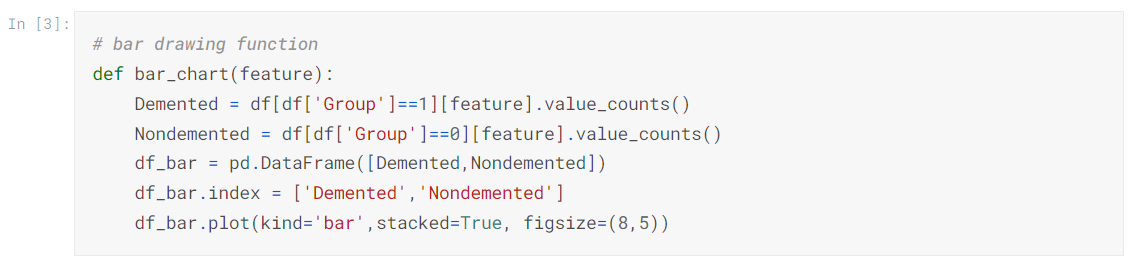
Các giá trị tối thiểu, tối đa và trung bình của từng tính năng để triển khai biểu đồ như sau.

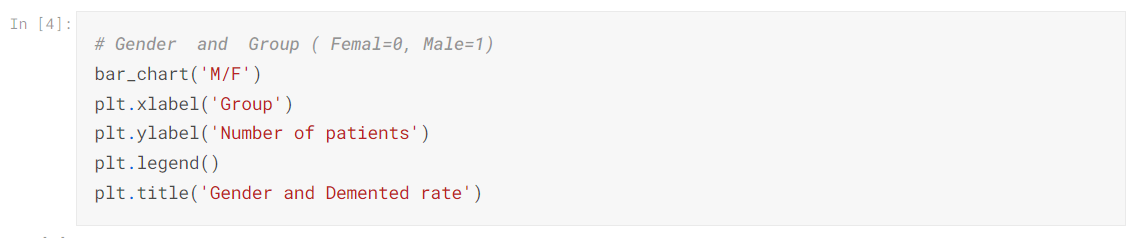


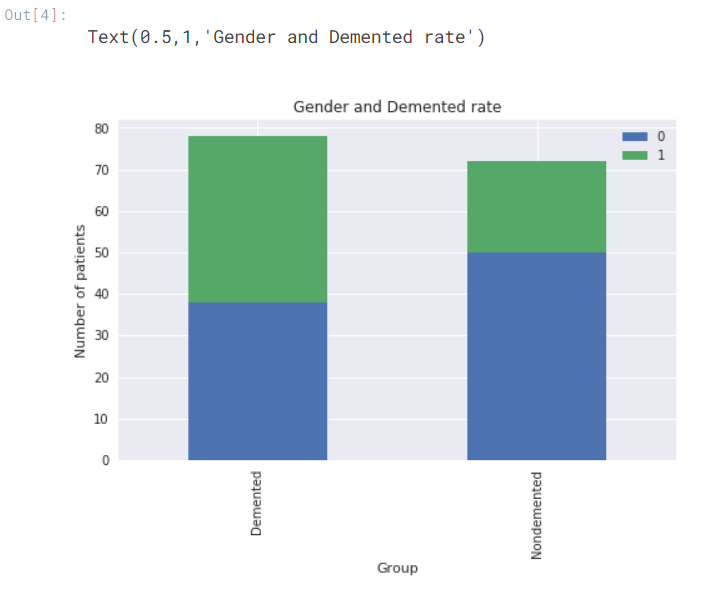




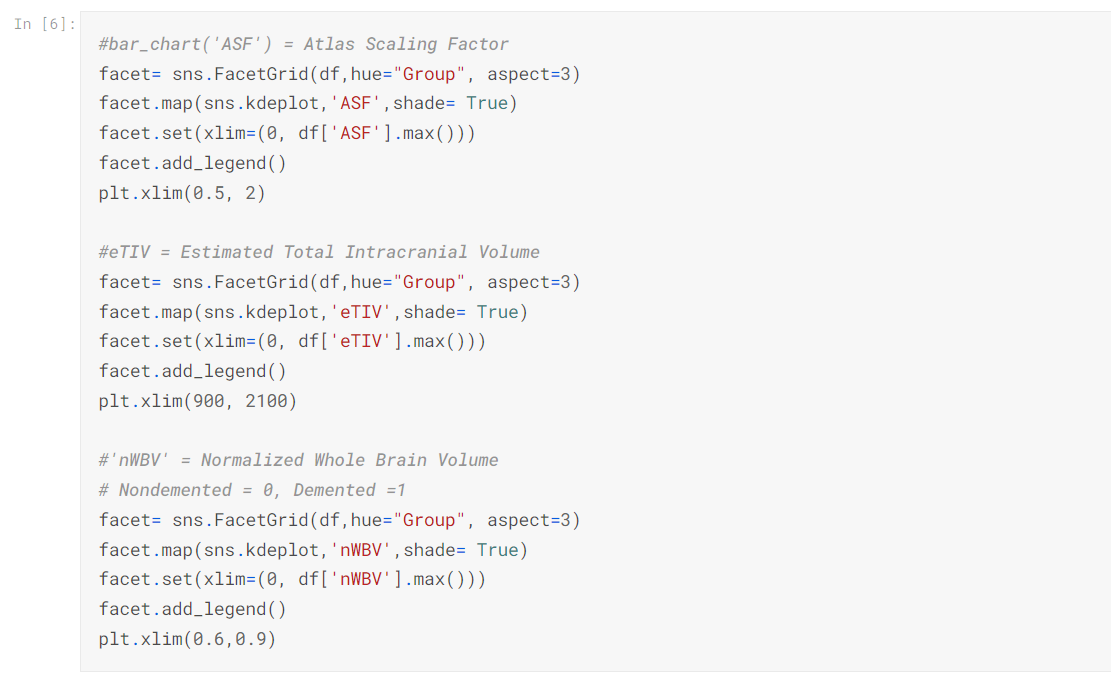


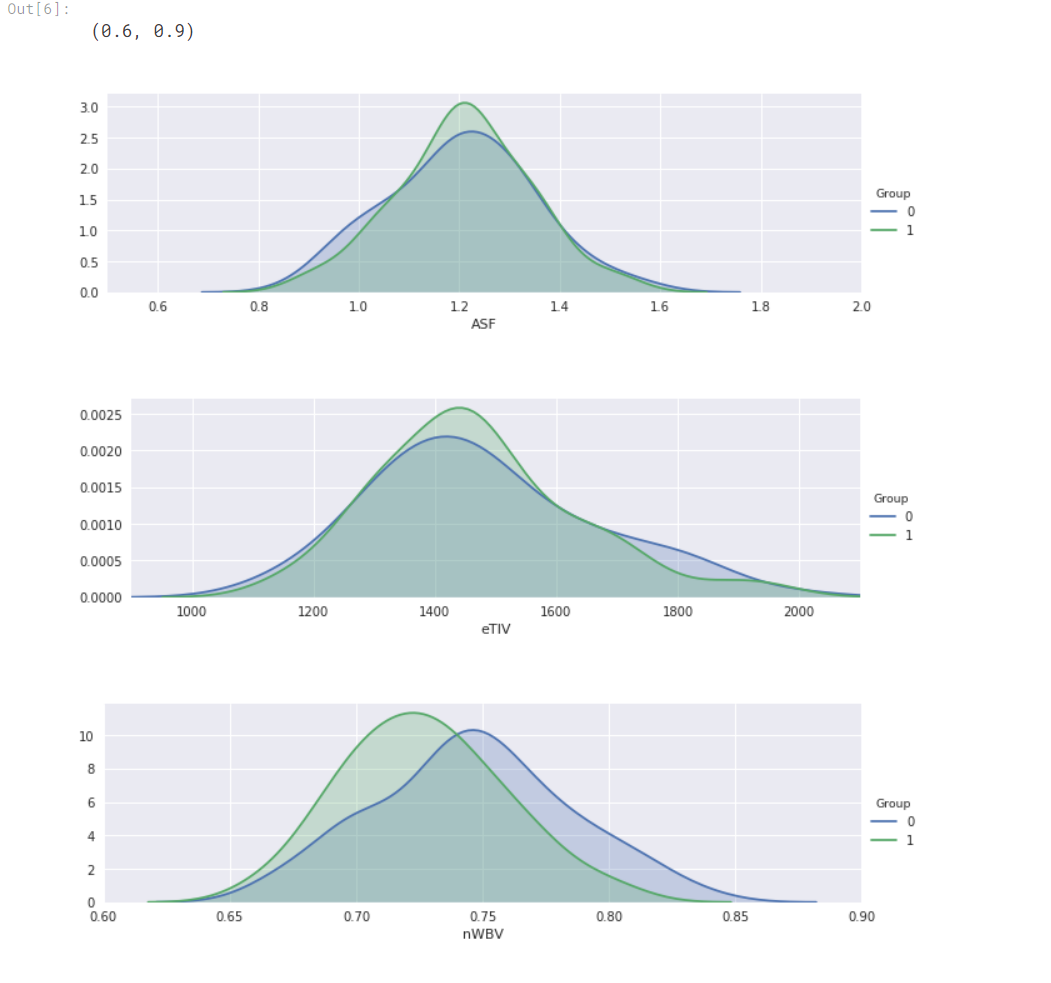




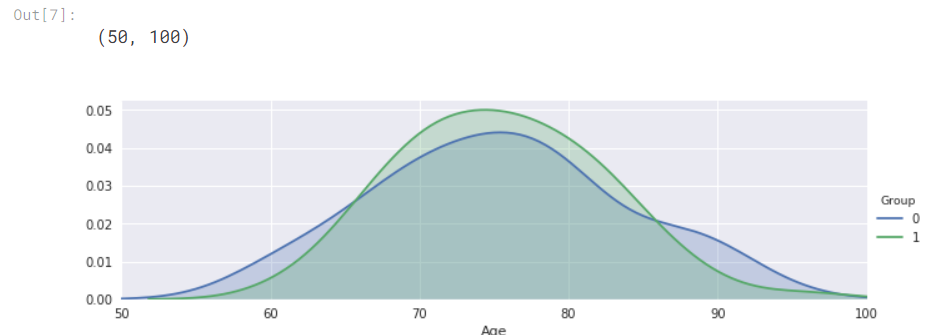




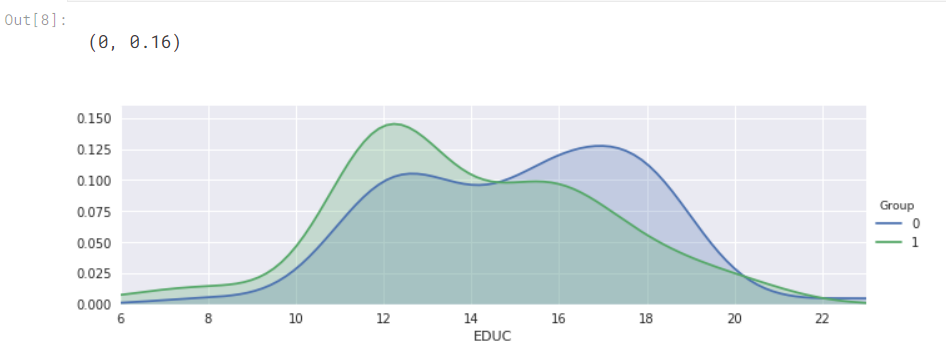












*Intermediate Result Summary (Tóm tắt kết quả trung gian)*

1. Đàn ông có nhiều khả năng mắc bệnh mất trí nhớ, bệnh Alzheimer hơn phụ nữ

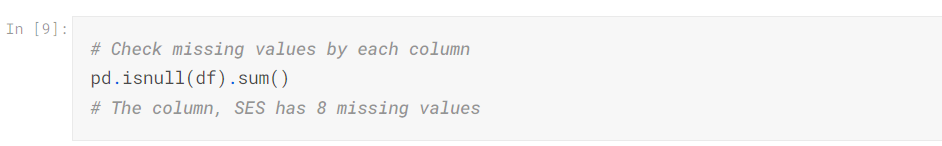
2. Những bệnh nhân mất trí nhớ ít được giáo dục hơn về số năm học.

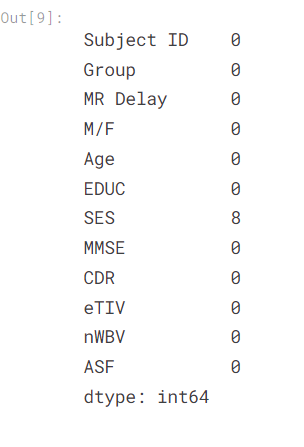
3. Nhóm không mất trí nhớ có khối lượng não cao hơn nhóm mất trí nhớ.

4. Nồng độ cao hơn ở nhóm 70-80 tuổi ở nhóm Sa sút trí tuệ so với nhóm bệnh nhân không sa sút trí tuệ.

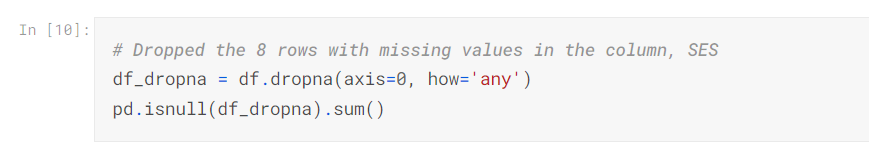
***5. Data Preprocessing (Tiền xử lý dữ liệu)***

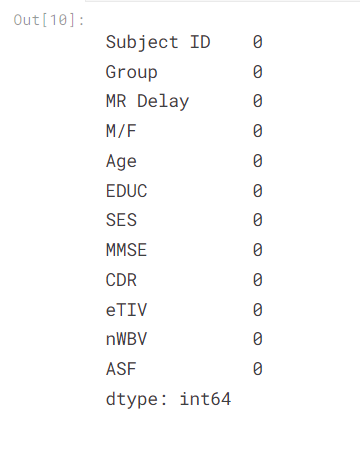
Chúng tôi đã xác định 8 hàng có giá trị bị thiếu trong cột SES. Chúng tôi giải quyết vấn đề này bằng 2 cách tiếp cận. Một là bỏ các hàng có giá trị bị thiếu. Cách khác là thay thế các giá trị bị thiếu bằng các giá trị tương ứng, còn được gọi là 'Đánh giá'. Vì chúng tôi chỉ có 150 dữ liệu, tôi cho rằng việc quy nạp sẽ giúp ích cho hiệu suất của mô hình của chúng tôi.

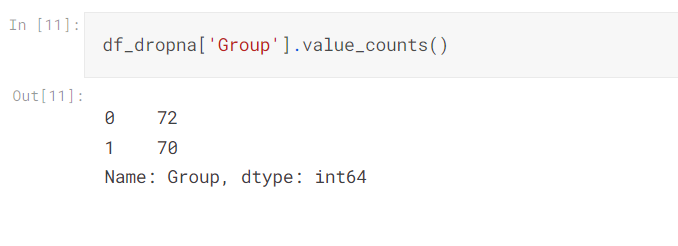




*5.A Removing rows with missing values (Xóa các hàng có giá trị bị thiếu)*

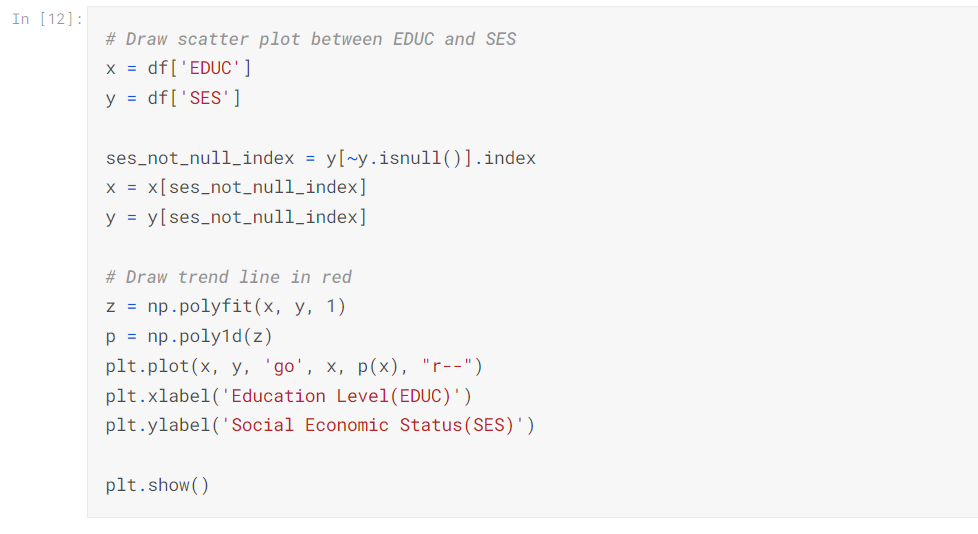


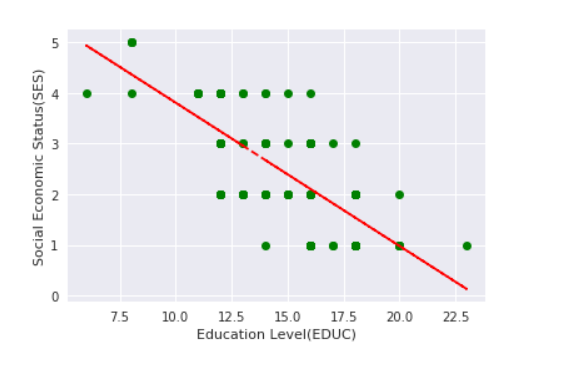


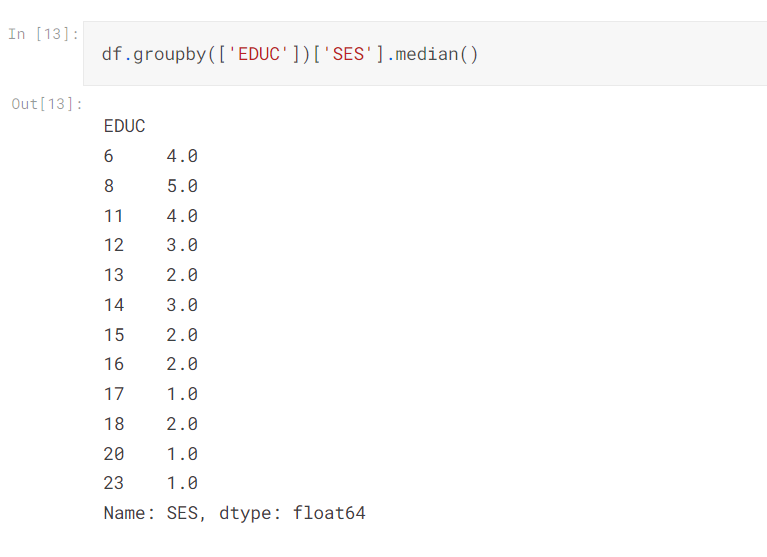


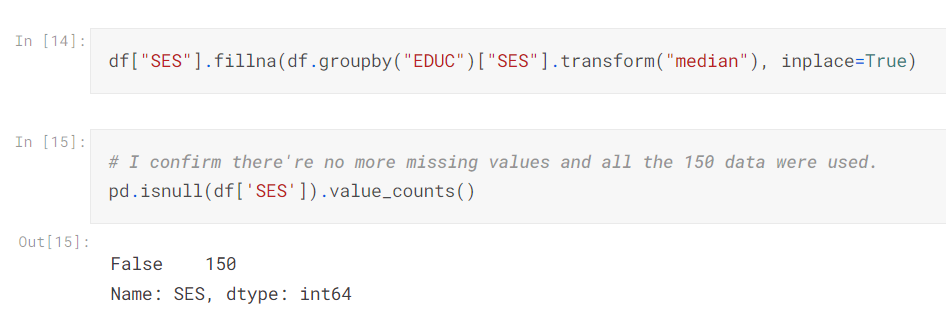
*5.B Imputation*

Scikit-learning cung cấp gói quy nạp [6], nhưng chúng tôi thực hiện thủ công. Vì SES là một biến rời rạc nên chúng tôi sử dụng giá trị trung bình để tính toán.

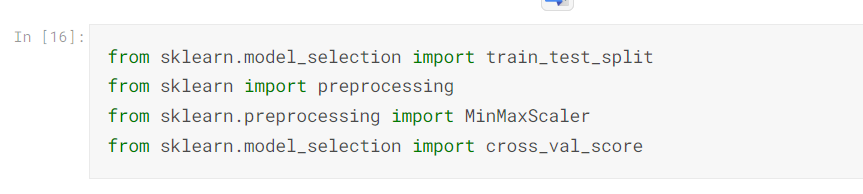




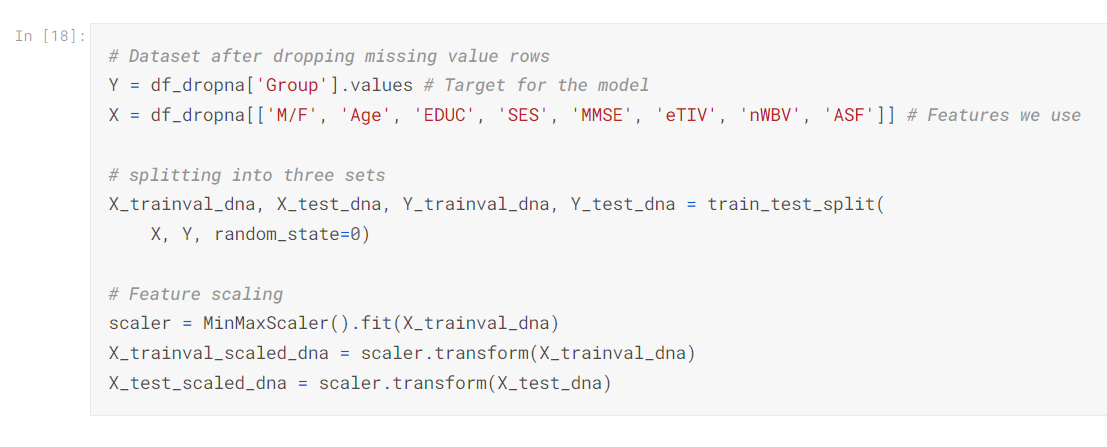




*5.C Splitting Train/Validation/Test Sets*







*5.D Cross-validation (Xác thực chéo)*

Chúng tôi tiến hành xác thực chéo 5 lần để tìm ra các tham số tốt nhất cho từng mô hình, Hồi quy logistic, SVM, Cây quyết định, Rừng ngẫu nhiên và AdaBoost. Vì chỉ số hiệu suất của chúng tôi là độ chính xác, nên chúng tôi tìm các thông số điều chỉnh tốt nhất theo độ chính xác. Cuối cùng, chúng tôi so sánh độ chính xác, khả năng thu hồi và AUC cho từng kiểu máy.

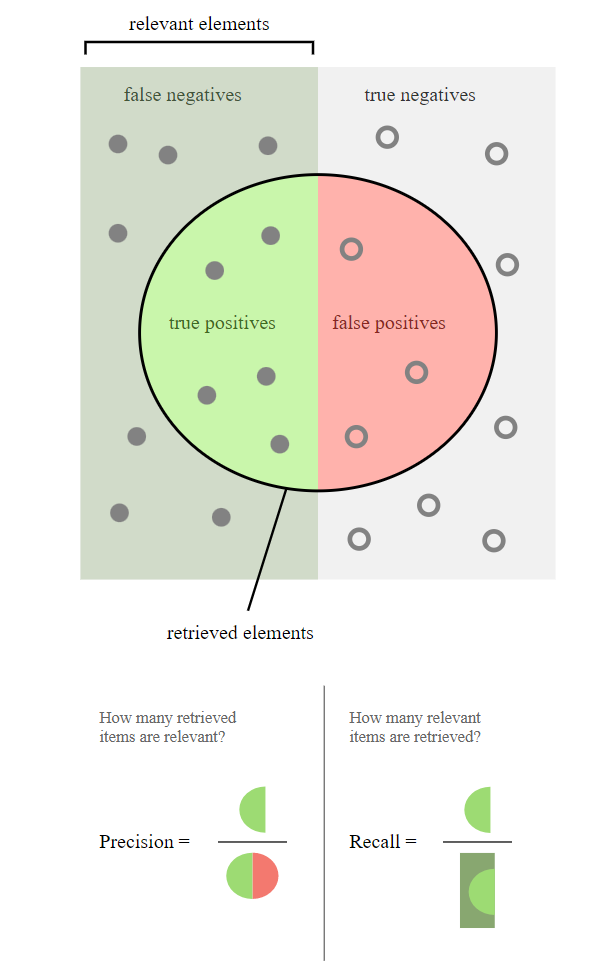
***6. MODEL***

*6.A Performance Measures (Các biện pháp thực hiện)*

Chúng tôi sử dụng diện tích bên dưới đường cong đặc tính hoạt động của máy thu (AUC) làm thước đo hiệu suất chính của chúng tôi. Chúng tôi tin rằng trong trường hợp chẩn đoán y tế đối với các bệnh giai đoạn cuối không đe dọa đến tính mạng như hầu hết các bệnh thoái hóa thần kinh, điều quan trọng là phải có tỷ lệ dương tính thực sự cao để tất cả bệnh nhân mắc bệnh alzheimer được xác định càng sớm càng tốt. Nhưng chúng tôi cũng muốn đảm bảo rằng tỷ lệ dương tính giả càng thấp càng tốt vì chúng tôi không muốn chẩn đoán nhầm một người trưởng thành khỏe mạnh là người mất trí nhớ và bắt đầu điều trị y tế. Do đó AUC dường như là một lựa chọn lý tưởng cho thước đo hiệu suất.

Chúng tôi cũng sẽ xem xét độ chính xác và thu hồi cho từng kiểu máy.

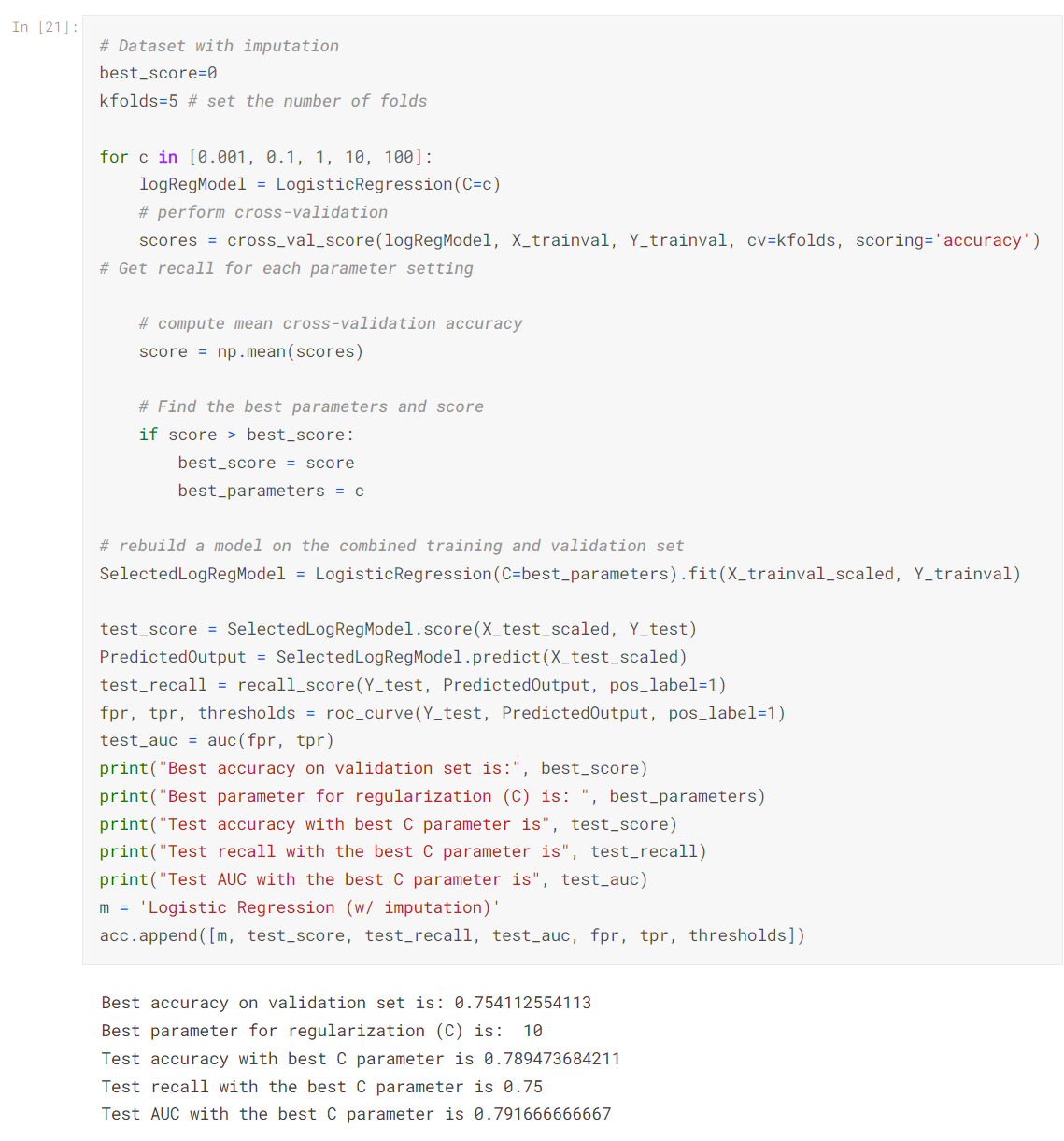
Trong hình bên dưới, bạn có thể coi các yếu tố liên quan là những chủ thể thực sự mất trí nhớ. Độ chính xác và Nhớ lại [12]

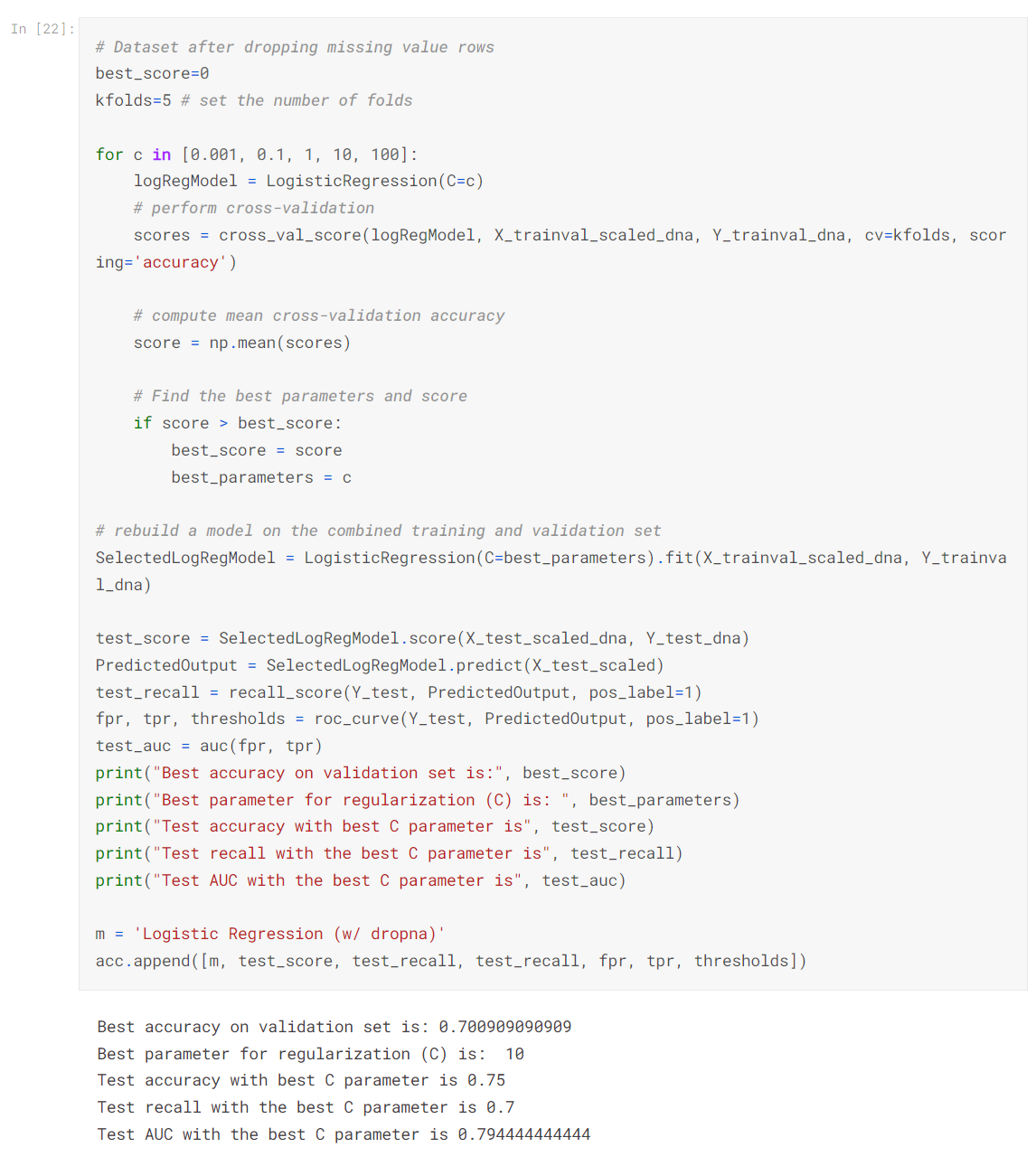


*6.B Logistic Regression*

Tham số C, nghịch đảo của cường độ chính quy hóa. Tuning range: [0.001, 0.1, 1, 10, 100]







Nhìn chung, tập dữ liệu có phép loại bỏ tốt hơn tập dữ liệu không có phép loại trừ. Đối với các mô hình sau này, chúng tôi sử dụng tập dữ liệu mà không cần quy nạp

*6.C SVM*

C: Penalty parameter C of the error term. [0.001, 0.01, 0.1, 1, 10, 100, 1000]

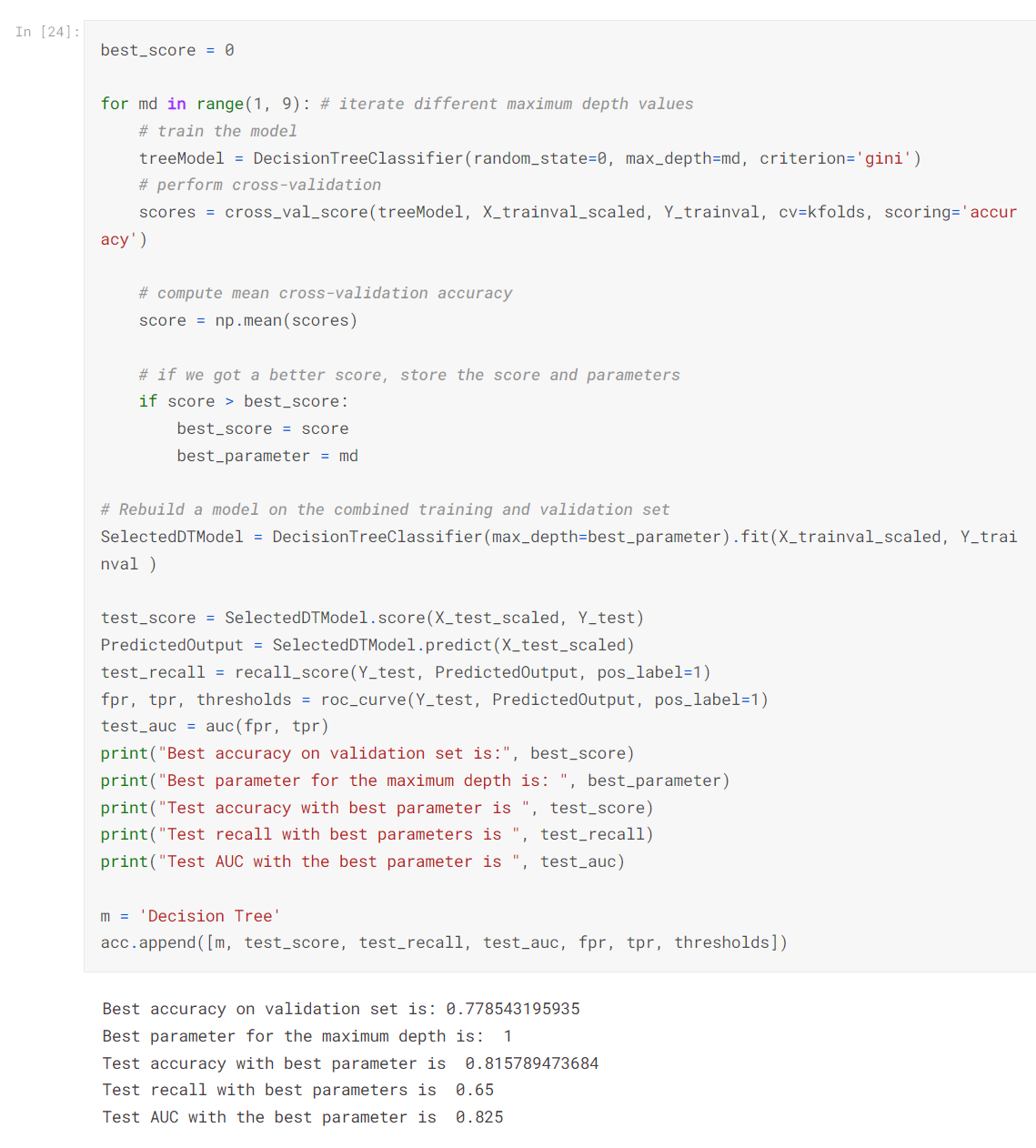
gamma: kernel coefficient. [0.001, 0.01, 0.1, 1, 10, 100, 1000]

kernel: kernel type. ['rbf', 'linear', 'poly', 'sigmoid']

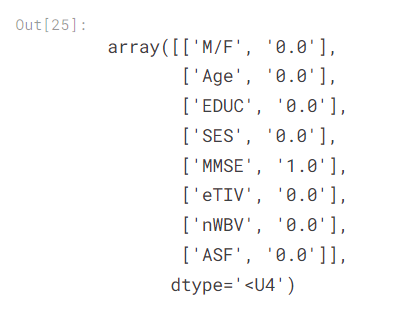


*6.D Decision Tree*

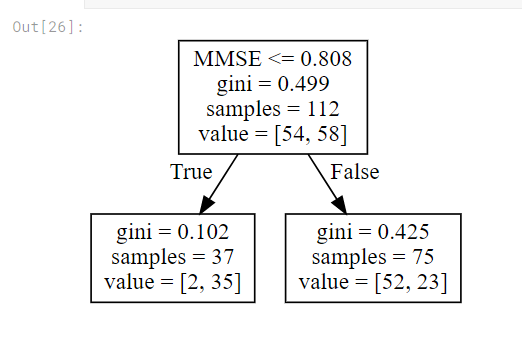
Maximum depth. [1, 2, ..., 8] 8 is the number of features









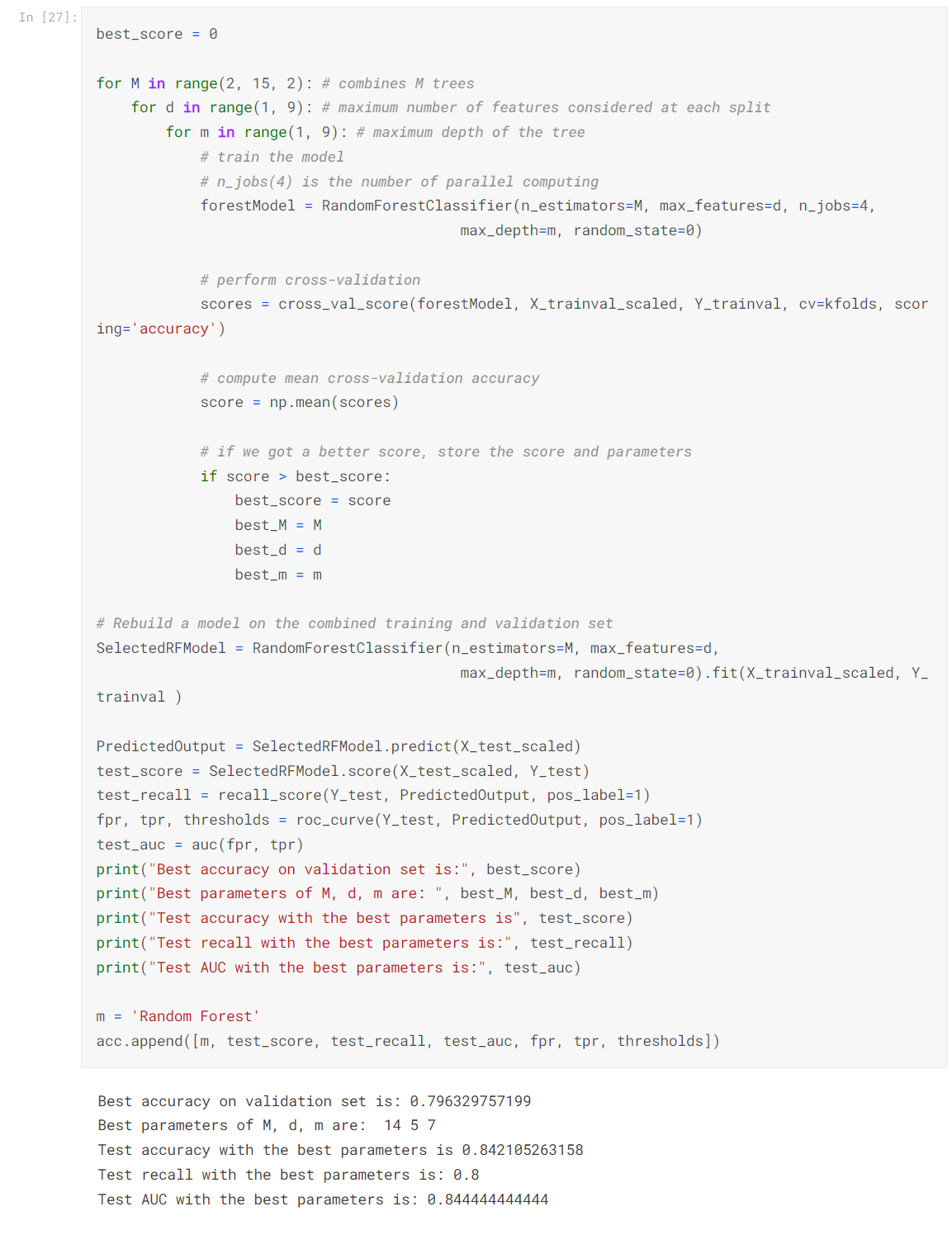


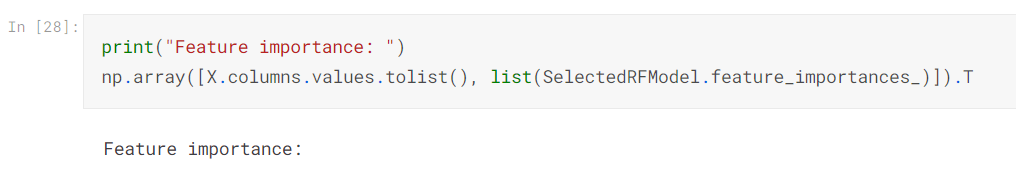
*6.E Random Forest Classifier (Bộ phân loại rừng ngẫu nhiên)*

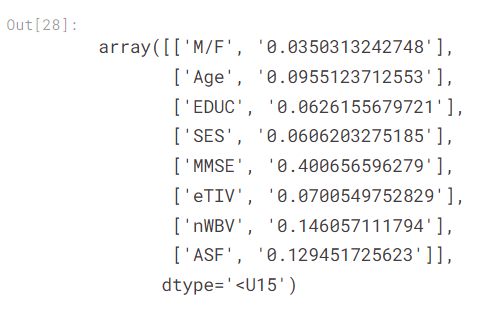
n\_estimators(M): số cây trong rừng

max\_features(d): số lượng các tính năng cần xem xét khi tìm kiếm sự phân chia tốt nhất

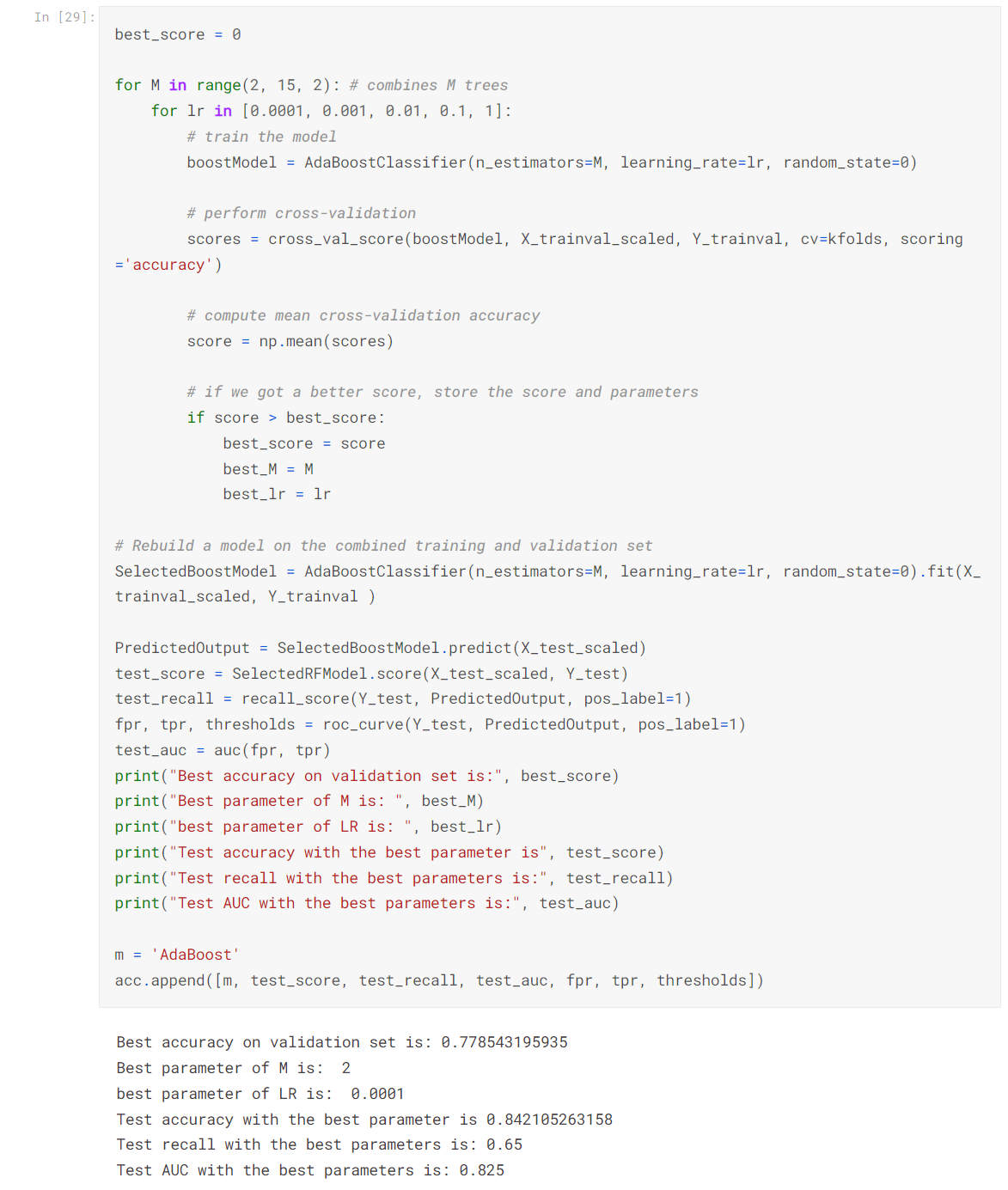
max\_depth(m): độ sâu tối đa của cây.



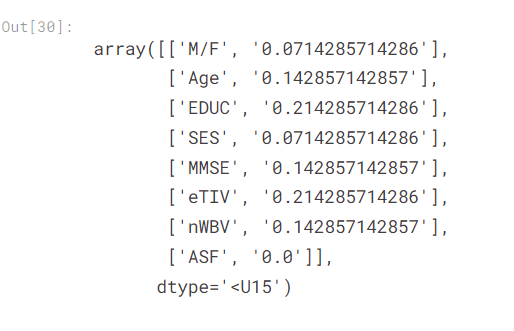




*6.F AdaBoost*



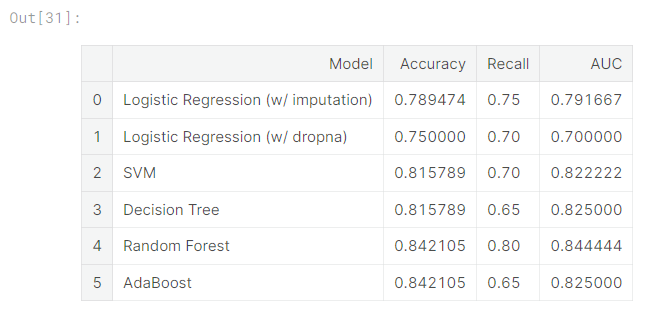




**7. CONCLUSION (Kết luận)**

*7.A RESULTS (Kết luận)*





Dưới đây là so sánh kết quả của chúng tôi với kết quả từ các bài báo đã được liệt kê trước đó:



Có thể nhận thấy rằng kết quả của chúng tôi có thể so sánh được và trong một số trường hợp nhất định tốt hơn so với kết quả từ công việc trước đó. Trình phân loại rừng ngẫu nhiên của chúng tôi là một trong những mô hình hoạt động tốt nhất.

*7.B UNIQUE APPROACH (Cách tiếp cận độc đáo)*

Điểm độc đáo trong cách tiếp cận của chúng tôi là chúng tôi sẽ bao gồm cả các số liệu như MMSE và Giáo dục trong mô hình của mình để huấn luyện mô hình phân biệt giữa những người trưởng thành khỏe mạnh bình thường và những người mắc bệnh Alzheimer. MMSE là một trong những tiêu chuẩn vàng để xác định chứng sa sút trí tuệ và do đó chúng tôi nghĩ rằng đây là một tính năng quan trọng cần đưa vào.

Thực tế tương tự cũng làm cho cách tiếp cận của chúng tôi đủ linh hoạt để áp dụng cho các bệnh thoái hóa thần kinh khác được chẩn đoán bằng cách sử dụng kết hợp các tính năng của MRI và các bài kiểm tra nhận thức.

*7.C IMPLEMENTATION (Thực hiện)*

Mục đích chính của các nhóm là khám phá cách học máy có thể tạo ra sự khác biệt trong môi trường lâm sàng. Để làm được điều đó, chúng tôi đã phát triển một chương trình web bằng thuật toán của mình, chương trình này có thể được sử dụng cho bất kỳ ai bất kể kinh nghiệm lập trình của họ. Bằng cách sử dụng mô-đun CGI, chúng tôi muốn mọi người tận dụng nỗ lực của bạn!

*7.D LIMITATIONS (Giới hạn)*

Có những hạn chế trong việc triển khai một mô hình phức tạp do số lượng của bộ dữ liệu. Mặc dù bản chất của từng tính năng là rõ ràng, phạm vi giá trị thử nghiệm của từng nhóm không được phân loại tốt. Nói cách khác, lẽ ra chúng ta nên xác định rõ ràng hơn sự khác biệt trong các biến có thể đóng vai trò trong kết quả. Giá trị dự đoán khi sử dụng mô hình rừng ngẫu nhiên cao hơn so với các mô hình khác. Nó ngụ ý rằng có khả năng đạt được tỷ lệ dự đoán cao hơn nếu chúng ta chú ý hơn đến việc phát triển quy trình phân tích và làm sạch dữ liệu. Hơn nữa, điểm thu hồi hoàn hảo 1.0 của SVM 1.0. Cho biết rằng chất lượng và độ chính xác của phân loại có thể giảm đáng kể khi chúng tôi sử dụng các tập dữ liệu khác nhau.

*7.E FURTHER RESEARCH (Nghiên cứu thêm)*

Bài học chính đối với chúng tôi là có một số yếu tố chính gây ra chứng Sa sút trí tuệ và chúng tôi nên tiếp tục kiểm tra và làm rõ quy trình theo những cách khác nhau. Để nghiên cứu sâu hơn, chúng tôi cần nâng cao hiểu biết của mình thông qua EDA phức tạp hơn xử lý với cỡ mẫu lớn hơn. Chẳng hạn, chúng tôi sẽ thử không chỉ độ tuổi mà còn nhóm nó thành thế hệ, hoặc phân loại khối lượng mô não hoặc điểm thi. Nếu kết quả từ quá trình này được phản ánh trong quá trình làm sạch dữ liệu và ảnh hưởng tích cực đến việc ra quyết định của mô hình, thì độ chính xác của mô hình dự đoán có thể được cải thiện hơn nữa.

***8. ACKNOWLEDGEMENTS (Lời cảm ơn)***

Tài liệu tham khảo

1. Marcus DS, Fotenos AF, Csernansky JG, Morris JC, Buckner RL. Open Access Series of Imaging Studies (OASIS): Longitudinal MRI Data in Nondemented and Demented Older Adults. Journal of cognitive neuroscience. 2010;22(12):2677-2684. doi:10.1162/jocn.2009.21407.

2. Marcus, DS, Wang, TH, Parker, J, Csernansky, JG, Morris, JC, Buckner, RL. Open Access Series of Imaging Studies (OASIS): Cross-Sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults. Journal of Cognitive Neuroscience, 19, 1498-1507. doi:10.1162/jocn.2007.19.9.1498

3. Elaheh Moradi, Antonietta Pepe, Christian Gaser, Heikki Huttunen, Jussi Tohka, Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects, In NeuroImage, Volume 104, 2015, Pages 398-412, ISSN 1053-8119, doi.org/10.1016/j.neuroimage.2014.10.002.

4. Zhang Y, Dong Z, Phillips P, et al. Detection of subjects and brain regions related to Alzheimer’s disease using 3D MRI scans based on eigenbrain and machine learning. Frontiers in Computational Neuroscience. 2015;9:66. doi:10.3389/fncom.2015.00066.

5. Magnin, B., Mesrob, L., Kinkingnéhun, S. et al. Support vector machine-based classification of Alzheimer’s disease from whole-brain anatomical MRI. Neuroradiology (2009) 51: 73. doi.org/10.1007/s00234-008-0463-x

6. http://scikit-learn.org/stable/modules/preprocessing.html#imputation

7. Ye, D.H., Pohl, K.M., Davatzikos, C., 2011. Semi-supervised pattern classification: application to structural MRI of Alzheimer's disease. Pattern Recognition in NeuroImaging(PRNI), 2011 International Workshop on. IEEE, pp. 1–4. http://doi:10.1109/PRNI.2011.12.

8. Filipovych, R., Davatzikos, C., 2011. Semi-supervised pattern classification of medical images: application to mild cognitive impairment (MCI). Neuroimage 55 (3), 1109–1119. https://doi.org/10.1016/j.neuroimage.2010.12.066

9. Zhang, D., Shen, D., 2012. Predicting future clinical changes ofMCI patients using longitudinal and multimodal biomarkers. PLoS One 7 (3), e33182. https://doi.org/10.1371/journal.pone.0033182

10. Batmanghelich, K.N., Ye, D.H., Pohl, K.M., Taskar, B., Davatzikos, C., 2011. Disease classification and prediction via semi-supervised dimensionality reduction. Biomedical Imaging: From Nano to Macro, 2011 IEEE International Symposium on. IEEE, pp. 1086–1090. http://10.1109/ISBI.2011.5872590

11. Ardekani,B.A.,Bachman,A.H.,Figarsky,K.,andSidtis,J.J.(2014).Corpus callosum shape changes in early Alzheimer’s disease: an MRI study using the OASISbraindatabase. BrainStruct.Funct. 219,343–352.doi:10.1007/s00429-013-0503-0

12. https://en.wikipedia.org/wiki/Precision\_and\_recall