Classification Model for Early Signs of Dementia

Sol Jung
Halıcıoğlu Data Science Institute | University of California San Diego class of 2024

Overview

The “oasis_longitudinal_demographics” data file consists of 15 variables with 373 observations. Of the 373 observations there is 150 patients, i.e. there are multiple records for each patient ranging from 1 to 5 visits. Using the Decision Tree Classification Model, the goals is to classify whether the patient is diagnosed with dementia or not, in order to further accurately categorize new observations. Reference the Note session at the bottom for the variables and their descriptions..

Methodology

In the data preprocessing stage, variables ‘Hand’, and ‘Visit’, were removed as all patients in this dataset were right-handed, and the number of visits correlated with the number of MRI ID scans. ”. Then instead of dropping the “N/A” values in SES and MMSE, the MICE function is used to replace the missing observations with multiple imputations using a regression model. Because there were multiple observations for each Subject ID (patient), for better accuracy, it was best to focus on the last MRI ID scan listed (the last visitation). The target variable, “Group”, consisted of the following : Demented, Nondemented, and Converted. In Model 1 “Converted” is replaced with ‘Demented” to increase the accuracy of the model, accuracy of 1. In Model 2, “Converted” is kept and had an accuracy of 0.88 in addition to more nodes in the decision tree reference Figure 3.
The model construction is straight forward, by creating a x (dependent variable) and y (target variable) to end up with a x and y train and test set, here the test size is 30%. Figure 1 is the visualization of the decision tree for the trained Model 1. Figure 2 is a joint plot to further visualize the relationship between CDR and MMSE in Model 1.

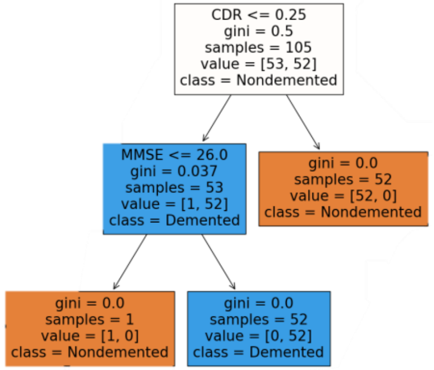

Figure 1: Model 1 Decision Tree, Accuracy =1

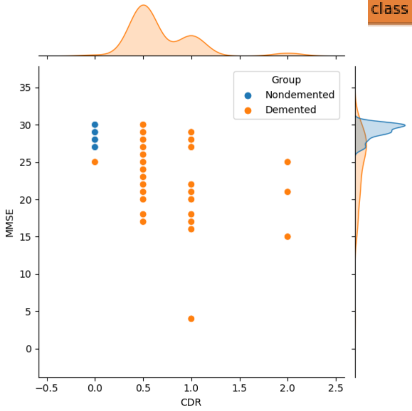

Figure 2: Joint plot of MMSE and CDR scores

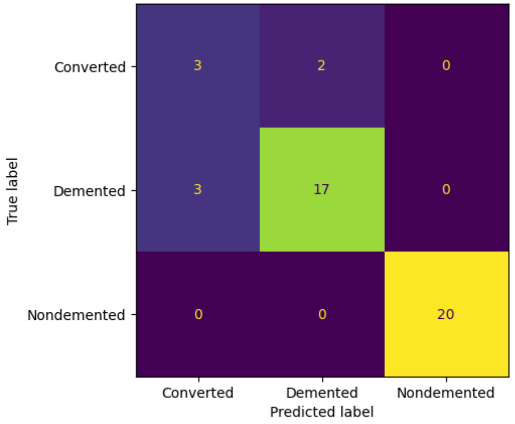

Figure 3: Confusion Matrix displaying the model classes, predictions and true values. Accuracy = 0.88

Analysis

Compared to Model 1 where the “Converted” observation is replaced with “Demented”, the accuracy of Model 2, 0.88, is lower. However, it is important to point out that the predicted dementia and converted are in fact true. In other words, in Figure 3, all nondemented patients were true positives. This model can be interpreted with an accuracy of 1 given the fact that converted patients in their last visit are diagnosed with Dementia even if they were considered Nondemented in the initial visits. Through Model 2 (reference Figure 4), the nodes in the decision tree bring light to the importance of the MR Delay, SES, eTIV, and Age, in addition the CDR and MMSE. Higher the CDR the more likely one will be classified with Dementia, however when it comes to the MMSE, a higher score does not necessarily imply that one is nondemented. So, by observing the other variables such as the SES and eTIV, it is safe to assume that the lower the social economic standing and an eTIV<= 1329, one will be classified as demented. Given that Dementia is correlated with brain shrinkage, it makes sense that someone with a smaller eTIV and the lack of resources such as medication, healthy food, etc. will be classified with Dementia.

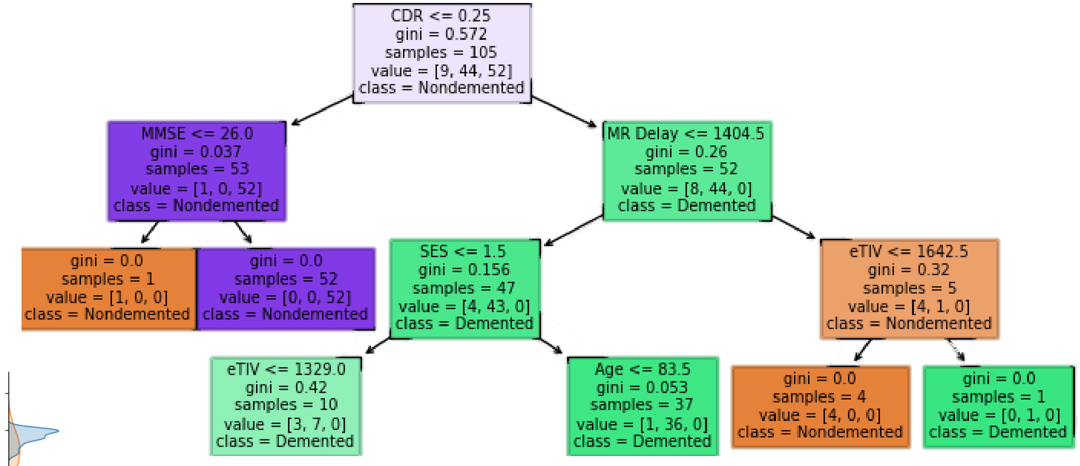

Figure 4: Model 2 Decision Tree with Converted, Demented, and Nondemented

Conclusion

Other supervised machine learning models such as SVM and KNN should be used to construct models to further accurately diagnose patients with Dementia earlier in their lives to start the prevention care as soon as possible as there are no definite cure for Dementia. Even if there are many variables, focusing on the MMSE, eTIV, and the CDR, for a SVM 3d model may be a better alternative to the decision tree, Model 1 and Model 2 may have different accuracies, however they both emphasize the importance of the CDR and MMSE scores. When creating a model without the CDR the accuracy of the model dramatically decreases. Based on the models, the CDR is the most reliable source when classifying patients with Dementia.

References

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

Notes

EDUC (Education), SES (Social Economic Status), CDR (Clinical Dementia Rating) 0 no dementia to 3 severe dementia,
MMSE (Mini-Mental State Examination) lower score can indicate cogitative impairment, eTIV (est. Total Intracranial Volume) total volume inside the skull ->average adult eTIV 1428, nWBV (normalized Whole Brain Volume) proportion of brain volume to the overall cranial volume, ASF (Atlas Scaling Factor) technique that measures that standardized total intracranial volume