BIOST 546 Final Project
DIBBYA BISWA
2023-03-10

Scientific Question:

Can we accurately classify subjects as having Alzheimer's Disease or being a Control subject based on their cortical thickness measurements using machine learning techniques?

Abstract:

This project aims to use machine learning techniques to differentiate between patients with Alzheimer's Disease (AD) and healthy elderly (C) by analyzing thickness measurements of their cerebral cortex. The dataset consists of cerebral cortex thickness measurements at 360 brain regions of interest and associated labels indicating whether the subject has Alzheimer's Disease or is a Control subject. Various machine learning models are applied to this dataset, and the performance of these models is evaluated using accuracy metrics for train and test data. The data analysis includes Logistic Regression, Decision Trees, and Random Forests models for both train and test data. The results show that the random forest model outperforms the other two models regarding accuracy for test data prediction. The analysis suggests that cortical thickness measurements can help detect Alzheimer's Disease.

Introduction:

Alzheimer's is a debilitating neurological disorder affecting millions of people worldwide. Early diagnosis of this disease can help in better management of the condition, and hence, there is a need for accurate diagnostic tools. In recent years, there has been a growing interest in using machine learning techniques to diagnose Alzheimer's disease using the data. This project aims to use machine learning tools to determine the difference between patients with Alzheimer's Disease and healthy elderly using thickness measurements of their cerebral cortex.

Data Analysis:

Three machine-learning techniques used are logistic regression, decision trees, and random forest models. Logistic regression is a standard technique for binary classification problems, and it models the relationship between the independent variables and the probability of the dependent variable. For differentiate between patients with Alzheimer's Disease (AD) and healthy elderly (C) by analyzing thickness measurements of their cerebral cortex, logistic regression was used to model the relationship between the predictor variables and the categorical outcome variable, which is whether a subject has Alzheimer's Disease or is a Control subject. This model is a good choice for our data as it can handle binary classification problems and can be easily interpreted.

Decision trees are a popular technique for classification problems, and they structure the data into smaller subgroups based on the values of the independent variables. This model is a good choice for the cortical thickness data as it can handle both categorical and continuous predictors and can perform well in high-dimensional settings. In this study, decision tree models were used to classify the subjects as having Alzheimer's Disease or being a Control subjects based on their cortical thickness measurements. Additionally, this model captures the relationships between different predictors and how they jointly influence the outcome, which is vital in the case of the complex relationships that may exist between cortical thickness measurements and Alzheimer's Disease.

Finally, random forest models are an ensemble technique that combines multiple decision trees to improve classification accuracy and reduce overfitting. This model is a good choice for our data as it can handle both categorical and continuous predictors, perform well in high-dimensional settings, and improve classification accuracy by aggregating multiple decision trees. Random forest models were used in this study to classify the subjects as having Alzheimer's Disease or being a Control subjects based on their cortical thickness measurements. This model is flexible and easy to interpret, providing a visual representation of the decision-making process, which is essential in the case of medical diagnosis, where the ability to interpret and explain the model's predictions is crucial.

Results:

Predictions Accuracy of Train and Test for Three Different Models			
Data	Logistic Regression Model	Decision Tree Model	Random Forest Models
Train Data	75.75%	80%	83.0%
Test Data	72.25%	73.25%	89%

Table 1: For the data, test prediction was performed blindly then later check for accuracy, where else train data was normally performed prediction and checked for accuracy.

The logistic regression model achieved a train prediction accuracy of 75.75% and a test prediction accuracy of 72.25%. This indicates that the model was able to capture some of the relationships between the predictor variables and the dependent variable but was not able to capture all of the complexity of the data. This is likely due to the fact that logistic regression is a linear model and can only capture linear relationships between the variables.

The decision tree model achieved a higher train prediction accuracy of 80% and a test prediction accuracy of 73.25%. This suggests that the decision tree model was able to capture more complex relationships between the variables compared to logistic regression. Decision trees are able to handle both categorical and continuous predictors, and they are able to perform well in high-dimensional settings. However, the model's accuracy on the test data was not significantly higher than that of logistic regression, which may indicate some overfitting.

The random forest model achieved the highest train prediction accuracy of 83.0% and a significantly higher test prediction accuracy of 89%. This indicates that the model effectively reduced variance and overfitting. Random forest models are an ensemble technique that combines multiple decision trees to improve classification accuracy and reduce overfitting. The success of the random forest model can be attributed to its ability to reduce variance by aggregating multiple decision trees. Random forests use bootstrapped training datasets to build a separate prediction model for each dataset and average the resulting predictions, limiting the overfitting in decision trees. Additionally, random forests randomly sample a subset of predictors to build each tree, decorrelating the trees and reducing variance even further. In contrast, decision trees and logistic regression models do not use such techniques and are more prone to overfitting.

Conclusions:

Based on the prediction accuracy results, it is possible to classify subjects accurately as having Alzheimer's Disease or being a Control subject based on their cortical thickness measurements using machine learning techniques. The random forest model outperformed the decision tree model and logistic regression model regarding accuracy. These results suggest that machine learning models can help diagnose Alzheimer's Disease and may have applications in other medical fields.

Advanced Analysis:

The performance of the random forest model was evaluated using confusion matrix statistics, including accuracy, sensitivity, specificity, positive predictive value, and negative predictive value. The confusion matrix shows that the random forest model correctly classified 332 out of 400 samples. The results show that the random forest model achieved a train predicted accuracy of 0.83, which is significantly higher than the no information rate of 0.7575. The 95% confidence interval ranges from 0.7895 to 0.8655, which means we can be confident that the true accuracy of the random forest model lies within this range. The Kappa value of 0.4192 indicates a moderate level of agreement between the predicted and actual classes. The sensitivity of 0.3505 indicates that the random forest model correctly identifies 35% of the individuals with Alzheimer's disease. The specificity of 0.9835 indicates that the random forest model correctly identifies 98.35% of the individuals without Alzheimer's disease. The positive predictive value of 0.8718 indicates that among those predicted to have Alzheimer's disease, 87.18% have the disease. The negative predictive value of 0.8255 indicates that among those predicted not to have Alzheimer's disease, 82.55% actually do not have the disease. Overall, the results of this advanced analysis suggest that the random forest model performs reasonably well in predicting Alzheimer's disease, but there is still room for improvement in its accuracy.

Recommendations:

Neural networks can be a powerful tool for analyzing cortical thickness data. The structure of a neural network allows it to automatically learn complex relationships between the input (predictor) variables and the output (response) variable. For example, in the case of cortical thickness, the input variables could be various features of brain scans, and the output variable could be the cortex's thickness in different brain regions.

By training a neural network on a large dataset of brain scans and associated cortical thickness measurements, we can develop a model that can accurately predict cortical thickness in new scans. This could be particularly useful for identifying individuals who may be at risk for specific neurological conditions or monitoring cortical thickness changes over time.

Additionally, neural networks can handle high-dimensional data, essential for analyzing brain scans with thousands of data points per subject. They can also handle missing and noisy data, which is common in medical imaging. Overall, neural networks can provide a powerful and flexible approach to analyzing cortical thickness data and can be a valuable tool in research aimed at understanding brain structure and function.

Appendix

Confusion Matrix and Statistics of Random Forest Model

	Reference	
Prediction	C	AD
C	34	63
AD	5	298

Accuracy : 0.83
95% CI : (0.7895, 0.8655)
No Information Rate : 0.7575
P-Value [Acc > NIR] : 0.0002861
Kappa : 0.4192
Mcnemar's Test -Value : 4.77e-12
Sensitivity : 0.3505
Specificity : 0.9835
Pos Pred Value : 0.8718
Neg Pred Value : 0.8255
Prevalence : 0.2425
Detection Rate : 0.0850
Detection Prevalence : 0.0975
Balanced Accuracy : 0.6670