Alzheimer’s disease\_Analysis Project Vikas

Vikas Munasu

2023-05-03

## Summary

The project involve prediction of the Alzheimer’s disease at its ealiest stages. It has included the predictive models random forest, decision tree model, and logistic regression model. The project intended to identify the most accurate model that could be used in predicting the Alzheimer’s disease at early stages. Therefore, the project has utilized data downloaded from the <https://dementia.talkbank.org/access/> website for the prediction. The data contains 3,012 rows and 8 observations. As Diagnosed is later created in the project to represent identified Alzheimer’s disease as 1 and not identified as 0. The predictions have shown that decision tree model is the most accurate model for predicting Alzheimer’s disease at early stages.

## Introduction

Alzheimer’s disease is a progressive disease that affects people’s brains and has been noted to have diverse impacts on millions of individuals globally. The disease has been identified to severely impact memory loss and affected individuals being unable to talk in a conversation. Health professionals have noted a challenge in identifying Alzheimer’s disease in its early stages due to a lack of accurate and simplified tests. Therefore, the project’s main focus include the creation of a machine learning method that is essential and can be used by a health professional to identify Alzheimer’s disease early stages. The project is important since it provides the capability of diagnosing Alzheimer’s disease at its early stages. Therefore, the health professionals are provided with an opportunity of identifying advanced outcomes on treatment and also ensure that there is no progression of Alzheimer’s disease due to early diagnosis of the disease.

## Literature Review

The project is based on previous works that have tried to understand how speech analytics can be utilized to identify early signs of Alzheimer’s disease. In their research, Tanveer et al. (2020), the authors utilized machine learning algorithms in performing analysis on speech patterns, where they made some meaningful insights into various changes associated with Alzheimer’s disease. In another study by Konig et al. (2018), there was research on using speech features for detecting the early signs of individuals suffering from cognitive impairment. In these two papers, there has been an illustration of having a high chance of using speech analysis to detect early signs in individuals who have Alzheimer’s disease.

## Data

Data used for this project is downloaded from the <https://dementia.talkbank.org/access/> website. The data for this project include:

## Rows: 3,012  
## Columns: 8  
## $ X <int> 0, 2, 3, 5, 6, 7, 8, 10, 11, 12, 14, 16, 17, 18, 20, 21…  
## $ subject <int> 101, 101, 101, 101, 101, 101, 101, 101, 101, 101, 101, …  
## $ counter.pauses <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, …  
## $ tone.level <dbl> 1.524, 3.141, 4.034, 5.624, 7.140, 9.497, 11.388, 13.85…  
## $ fluency.level <dbl> 1.817, 3.232, 4.629, 6.884, 7.957, 10.789, 12.019, 14.3…  
## $ hard.Pause <dbl> 0.2921, 0.0909, 0.5954, 1.2597, 0.8169, 1.2927, 0.6314,…  
## $ Type <chr> "IntraClausal Pause", "Pre-ClaSub InterClasal Pause", "…  
## $ Diagnosis <chr> "AD", "AD", "AD", "AD", "AD", "AD", "AD", "AD", "AD", "…

Confirm if the dataset has any missing values for each variable.

## X subject counter.pauses tone.level fluency.level   
## 0 0 0 0 0   
## hard.Pause Type Diagnosis   
## 0 0 0

Recode the Diagnosis variable to a new variable Diagnosed to 1s and 0s for being diagnosed with Alzheimer’s Disease (AD) and others, respectively. This is done by indexing the vectors.

## Rows: 3,012  
## Columns: 9  
## $ X <int> 0, 2, 3, 5, 6, 7, 8, 10, 11, 12, 14, 16, 17, 18, 20, 21…  
## $ subject <int> 101, 101, 101, 101, 101, 101, 101, 101, 101, 101, 101, …  
## $ counter.pauses <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, …  
## $ tone.level <dbl> 1.524, 3.141, 4.034, 5.624, 7.140, 9.497, 11.388, 13.85…  
## $ fluency.level <dbl> 1.817, 3.232, 4.629, 6.884, 7.957, 10.789, 12.019, 14.3…  
## $ hard.Pause <dbl> 0.2921, 0.0909, 0.5954, 1.2597, 0.8169, 1.2927, 0.6314,…  
## $ Type <chr> "IntraClausal Pause", "Pre-ClaSub InterClasal Pause", "…  
## $ Diagnosis <chr> "AD", "AD", "AD", "AD", "AD", "AD", "AD", "AD", "AD", "…  
## $ Diagnosed <fct> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1…

View the created Diagnosed variable.

##   
## 0 1   
## 2305 707

## Methodology

The data is split into 70% training set and 30% testing set.

## split\_dataset  
## 0 1   
## 2108 903

Create both the training set and testing set as dataset\_train and dataset\_test respectively.

dataset\_train <- dataset[split\_dataset == 0, ]  
dataset\_test <- dataset[split\_dataset== 1, ]

## Build Random Forest Model

Create the random forest using Diagnosed as target variable and tone.level, fluency.level, and hard.Pause predictor variables.

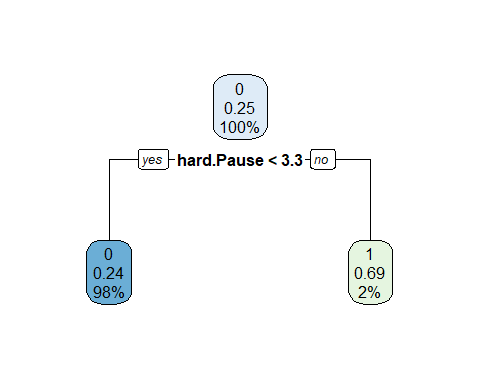
##   
## Call:  
## randomForest(formula = Diagnosed ~ tone.level + fluency.level + hard.Pause, data = dataset\_train, proximity = TRUE)   
## Type of random forest: classification  
## Number of trees: 500  
## No. of variables tried at each split: 1  
##   
## OOB estimate of error rate: 28.07%  
## Confusion matrix:  
## 0 1 class.error  
## 0 1459 167 0.1027060  
## 1 425 58 0.8799172

We use the created random forest to make predictions on test data through confusion matrix approach (Finnstats, 2021).

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction 0 1  
## 0 618 193  
## 1 61 31  
##   
## Accuracy : 0.7187   
## 95% CI : (0.6882, 0.7478)  
## No Information Rate : 0.7519   
## P-Value [Acc > NIR] : 0.9899   
##   
## Kappa : 0.0605   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.9102   
## Specificity : 0.1384   
## Pos Pred Value : 0.7620   
## Neg Pred Value : 0.3370   
## Prevalence : 0.7519   
## Detection Rate : 0.6844   
## Detection Prevalence : 0.8981   
## Balanced Accuracy : 0.5243   
##   
## 'Positive' Class : 0   
##

From the predictions, it is clear that the model has accuracy of 71.98% in predicting the diagnosis of AD (Alzheimer’s Disease) using the variables tone.level, fluency.level, and hard.Pause.

## Build Decision Tree Model

Using the rpart library (Finnstats, 2021b), the decision tree model is built. 

Using the predict() function in R, the test dataset is predicted. Then a table illustrating the count of patients classified to be diagnosed with AD and those not is created.

## predictions\_2  
## 0 1  
## 0 674 5  
## 1 213 11

To evaluate the classification performance of the decision tree model, the confusion matrix method has been applied. The prediction accuracy of the test data is depicted to be 75.85%.

## [1] "Accuracy for decision tree is: 0.758582502768549"

## Build Logistic Regression Model

Create the Logistic Regression Model using Diagnosed as target variable and tone.level, fluency.level, and hard.Pause predictor variables

##   
## Call:  
## glm(formula = Diagnosed ~ tone.level + fluency.level + hard.Pause,   
## family = "binomial", data = dataset\_train)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6679 -0.7164 -0.6756 -0.6240 1.8878   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -1.42052 0.09671 -14.688 <2e-16 \*\*\*  
## tone.level -135.60998 129.07258 -1.051 0.293   
## fluency.level 135.60886 129.07260 1.051 0.293   
## hard.Pause -135.33011 129.07381 -1.048 0.294   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 2269.7 on 2108 degrees of freedom  
## Residual deviance: 2236.7 on 2105 degrees of freedom  
## AIC: 2244.7  
##   
## Number of Fisher Scoring iterations: 4

Test the Accuracy of the logistic regression model

## [1] "Accuracy for logistic regression analysis: 0.753045404208195"

## Results and Conclusion

From the prediction analysis conducted in this project, random forest, decision tree, and logistic regression predictive models have been developed. The analysis have been a comparative analysis for the three models. The goal is to identify the best model that can be implemented for identifying the Alzheimer’s Disease at its earliest stages. The results of the three models have depicted that decision tree model has the highest accuracy of prediction, at 75.85%. Therefore, it is concluded that the best model to use in prediction of the Alzheimer’s Disease is Decision Tree model.

## References

Konig, A., Satt, A., Sorin, A., Hoory, R., Derreumaux, A., David, R., & Robert, P. H. (2018). Use of speech analyses within a mobile application for the assessment of cognitive impairment in elderly people. Current Alzheimer Research, 15(2), 120-129.

Tanveer, M., Richhariya, B., Khan, R. U., Rashid, A. H., Khanna, P., Prasad, M., & Lin, C. T. (2020). Machine learning techniques for the diagnosis of Alzheimer’s disease: A review. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 16(1s), 1-35. Finnstats. (2021b). Decision Trees in R | R-bloggers. R-bloggers. <https://www.r-bloggers.com/2021/04/decision-trees-in-r/>

Finnstats. (2021). Random Forest in R | R-bloggers. R-bloggers. <https://www.r-bloggers.com/2021/04/random-forest-in-r/>