Jonathan Ramos
COMP 4441 Final Pape
Predicting Alzheimer's Disease Diagnoses

**Figures and figure legends for reference in the final paper.**

Figure 1

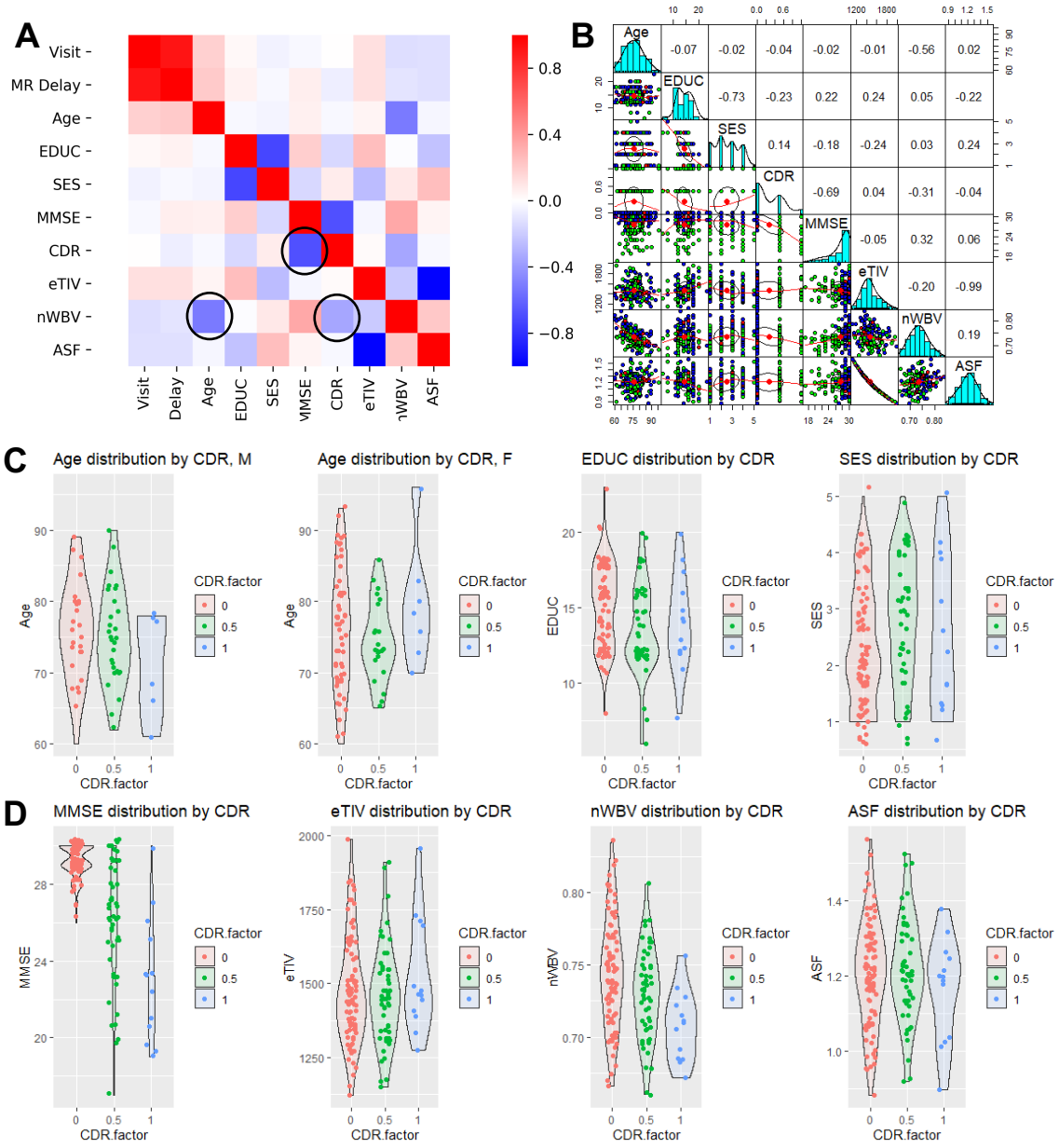



Figure 1. A) Correlation heatmap of demographic and continuous variables from the dataset. Correlations of note are circled. B) Pair plot of the same, reduced to columns of interest. C) Violin plots of demographic variables. D) Violin plots of measured and derived variables in set.

Figure 2

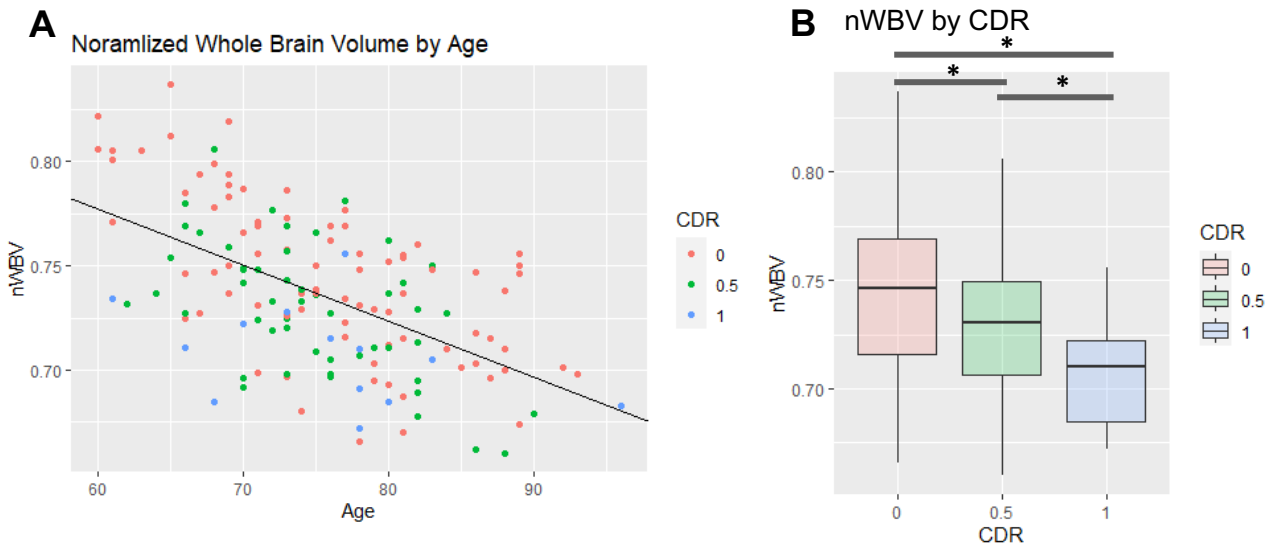



Figure 2. A) Scatter plot of nWBV by Age with regression line, points color coded by CDR. Regression line intercept = 0.9395, p=2e-16; slope = -0.002696, p=1.18e-13, adjusted R squared = 0.3067. nWBV and Age are well correlated, Pearson's r=-0.558, p=1.183e-13. B) Boxplots of nWBV grouped by CDR. All three columns are significantly different from each other, one-way ANOVA p=0.000117, post-hoc Welch's t-tests: CDR 0 and CDR 0.5 p=0.01128, CDR 0.5 and CDR 1 p=0.1186, CDR 1 and CDR 0 p=9.115e-5. Note Bonferroni corrected p value for chosen alpha level is p=0.01667 to account for multiple comparisons.

Figure 3

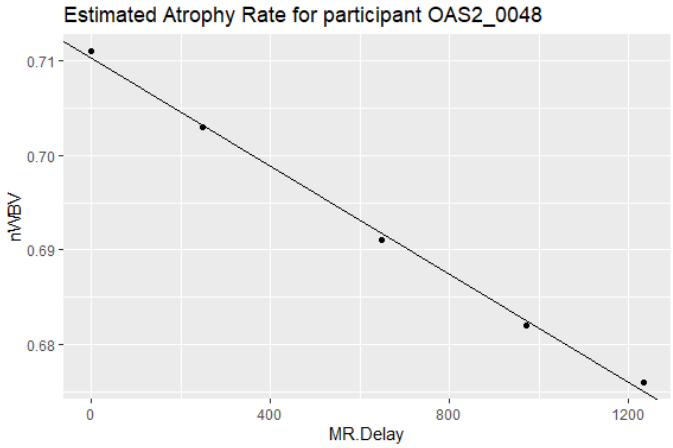



Figure 3. nWBV measurements over MR Delay (time between visits, units in days) with regression line of representative participant. Slope from linear regression line is taken to be the estimated atrophy rate for that participant.

Figure 4

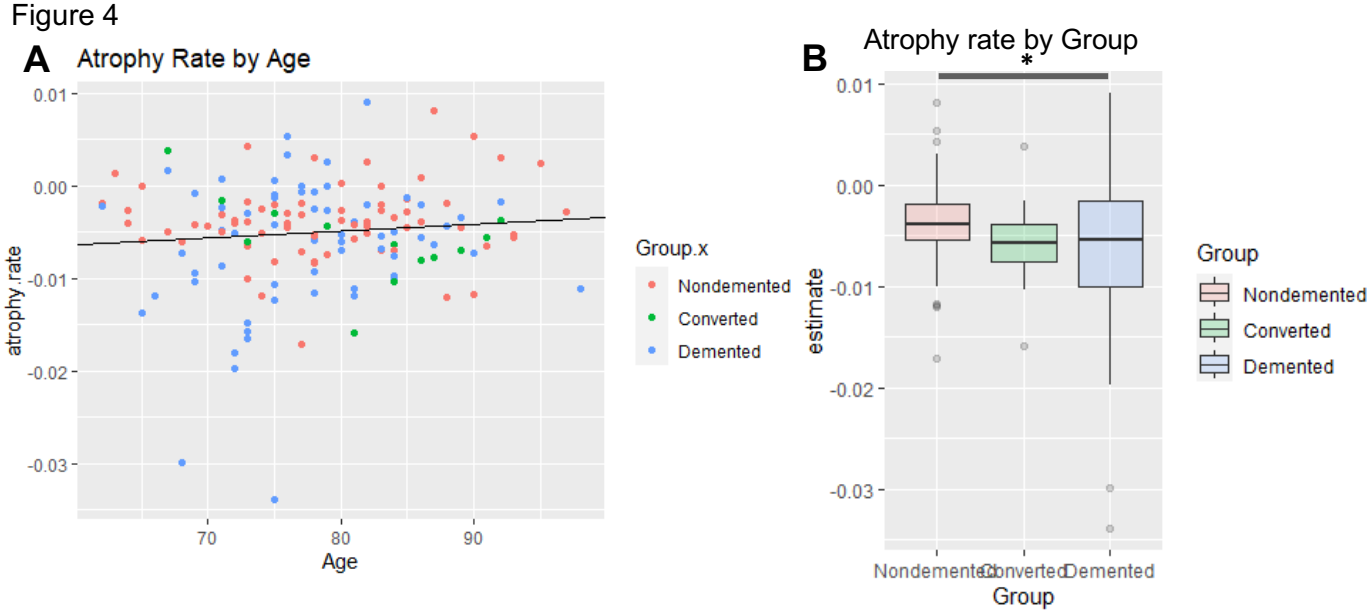
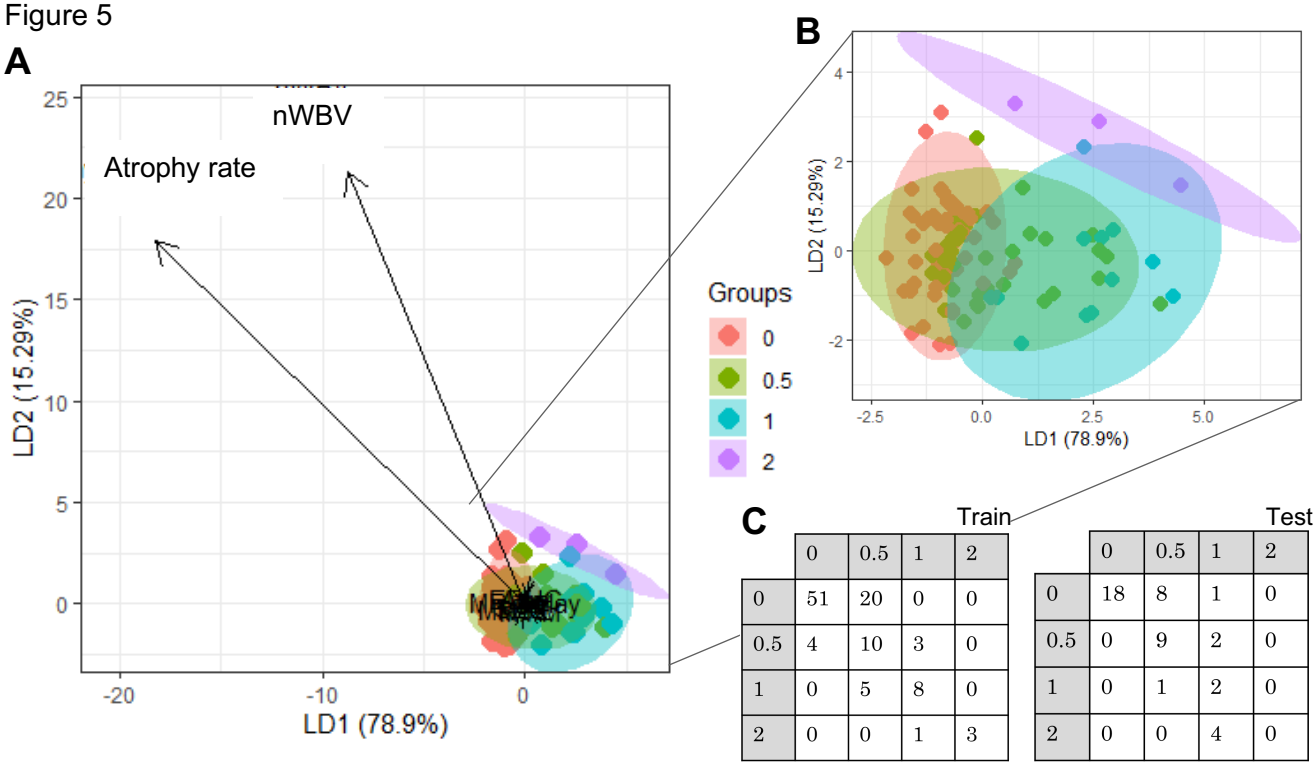



Figure 4. A) Estimated atrophy rate by Age with regression line, points color coded by Group label. Regression line intercept = -1.075e-2, p=0.0297, slope = 7.306e-5, p=0.2419. No correlation between Atrophy rate and Age, Pearson's r=0.0961, p=0.2419. B) Boxplots of estimated Atrophy rate by Group label. ANOVA p=0.024107, post-hoc Welch's t-tests: Nondemented and Converted p=0.1308, Converted and Demented p=0.6438, Demented and Nondemented p=0.009. Note Bonferroni corrected p value for chosen alpha level is p=0.01667 to account for multiple comparisons.

Figure 5




C

Train

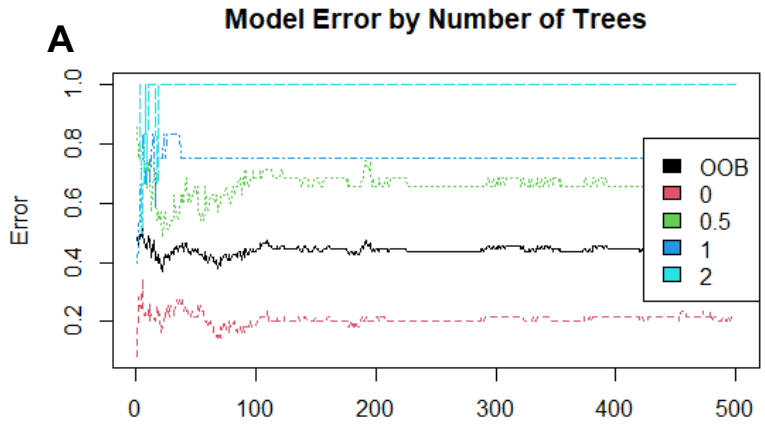
| | 0 | 0.5 | 1 | 2 |
|---|---|---|---|---|
| 0 | 51 | 20 | 0 | 0 |
| 0.5 | 4 | 10 | 3 | 0 |
| 1 | 0 | 5 | 8 | 0 |
| 2 | 0 | 0 | 1 | 3 |

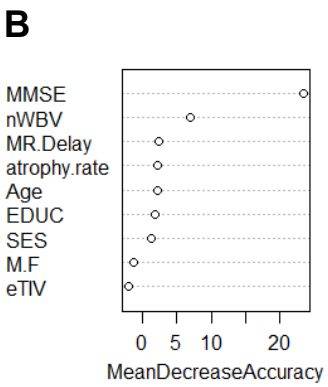
Test

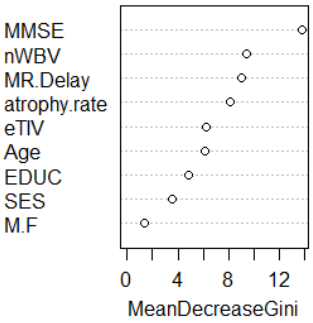
| | 0 | 0.5 | 1 | 2 |
|---|---|---|---|---|
| 0 | 18 | 8 | 1 | 0 |
| 0.5 | 0 | 9 | 2 | 0 |
| 1 | 0 | 1 | 2 | 0 |
| 2 | 0 | 0 | 4 | 0 |

Figure 5. A) Ordination plot showing classes separated by LDA plotted against the first two LD components. Axes indicate that nWBV and Atrophy rate were the most important in separating classes. B) Zoomed in view of ordination plot to view overlap. C) Confusion matrices for training and testing (70/30 split).

Figure 6




C

Train

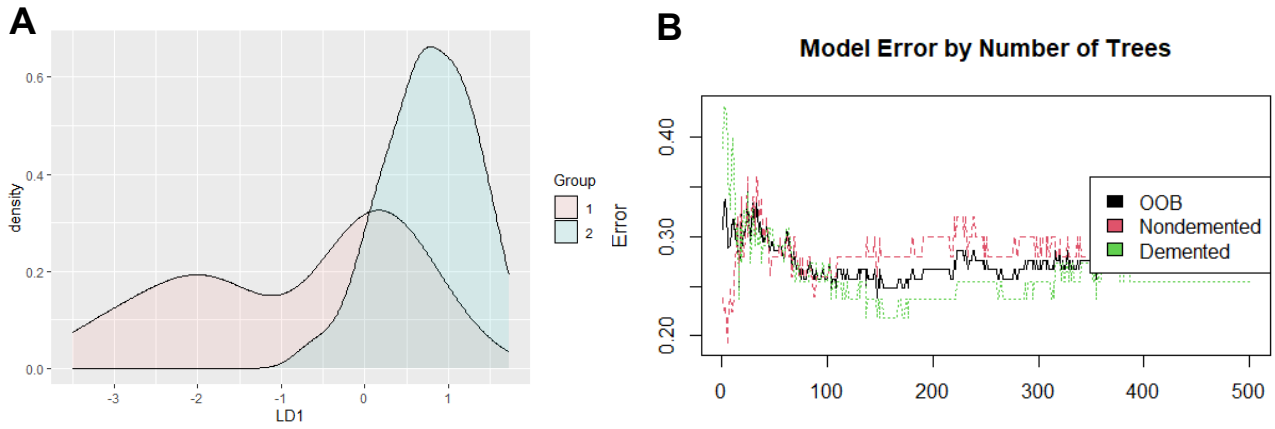
| | 0 | 0.5 | 1 | 2 |
|---|---|---|---|---|
| 0 | 51 | 20 | 0 | 0 |
| 0.5 | 4 | 10 | 3 | 0 |
| 1 | 0 | 5 | 8 | 0 |
| 2 | 0 | 0 | 1 | 3 |

Test

| | 0 | 0.5 | 1 | 2 |
|---|---|---|---|---|
| 0 | 18 | 8 | 1 | 0 |
| 0.5 | 0 | 9 | 2 | 0 |
| 1 | 0 | 1 | 2 | 0 |
| 2 | 0 | 0 | 4 | 0 |

Figure 6. A) Model error by number of trees from Random Forest model classifying CDR as 0, 0.5, 1 or 2; 500 trees, two variables tried at each split. B) Importance of variables. Both mean decrease in accuracy and mean decrease in GINI indicate that MMSE is the most important feature, followed by nWBV, MR Delay and atrophy rate. C) Confusion matrices for training and testing (70/30 split)

Figure 7




C

| | Train | | Test | |
|---|---|---|---|---|
| | 1 | 2 | 1 | 2 |
| 1 | 30 | 3 | 20 | 0 |
| 2 | 20 | 52 | 7 | 18 |

| | Train | Test |
|---|---|---|
| Sensitivity | 60.00% | 74.07% |
| Specificity | 94.55% | 100.0% |
| Accuracy | 78.1% | 84.44% |

D

| | Train | | Test | |
|---|---|---|---|---|
| | 1 | 2 | 1 | 2 |
| 1 | 36 | 14 | 22 | 5 |
| 2 | 14 | 41 | 5 | 13 |

| | Train | Test |
|---|---|---|
| Sensitivity | 60.00% | 81.48% |
| Specificity | 74.55% | 72.22% |
| Accuracy | 73.33% | 77.78% |

Figure 7 A) Density plot of points cast onto the first LD component from the binary LDA model where Class 2 is treated as nondemented and Class 1 is treated as demented; Class 1 is the positive class. B) Model error by number of trees from the binary Random Forest model only trying to classify demented (class 1) vs nondemented (class 2); 500 trees, 2 variables tried at each split. C,D) Confusion matrices from training and testing (70/30 split) along with summary table describing sensitivity, specificity and accuracy for the binary LDA model (C) and the binary Random Forest model (D).

Figure 8

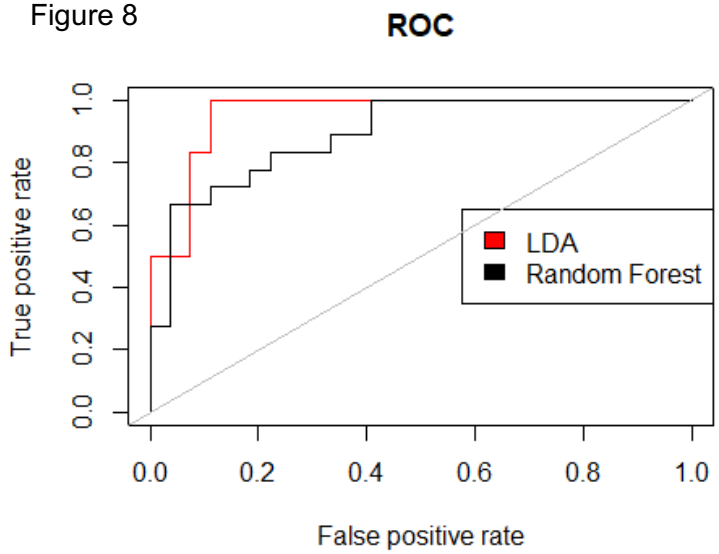



Figure 8. Response operator characteristic curves for the binary LDA model and the binary Random Forest model. Area under the curve for LDA: 0.957. Area under the curve for Random Forest: 0.893.