Predictive Modeling and Classification for Alzheimer's Disease

Olivia Fang

March 22, 2025

Abstract

This study applies machine learning techniques to classify Alzheimer's Disease (AD) vs. Control (C) using cortical thickness measurements from 360 brain regions across 339 individuals. Given the high dimensionality and class imbalance of the dataset, we explored generalized linear models (GLM), decision trees, and random forests, incorporating LASSO regularization and feature selection to improve model performance. Initial GLM models suffered from convergence issues and overfitting, leading us to tree-based methods. Random forest emerged as the best-performing model, outperforming decision trees and LASSO-GLM.

1 Introduction

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder that affects memory and cognitive function, making early detection crucial. Machine learning provides a promising approach for identifying AD using cortical thickness measurements, but challenges such as high dimensionality, correlated predictors, and class imbalance complicate model development.

To address these issues, we began with data analysis, identifying key patterns and potential limitations. This informed our modeling approach, where we applied feature selection, regularization, and ensemble methods to improve classification accuracy and generalization. This report outlines how data exploration guided model selection, ensuring a structured and effective strategy for AD classification.

2 Data Analysis

The dataset consists of 339 observations and 360 cortical thickness measurements, with one additional column representing the outcome variable, which is binary—denoting Alzheimer's Disease (AD) or Control (C). This results in a dataset with 339 rows and 361 columns.

An important aspect of this dataset is that AD cases significantly outnumber Control cases, leading to a class imbalance. This imbalance poses a risk of training a classifier biased toward the majority class, which will be further examined in the modeling sections.

2.1 Correlation Analysis

Heatmap

The correlation heatmap (Figure 1) visualizes the relationships among the first 30 predictors. These predictors were selected as a representative subset due to the high dimensionality of the dataset (360 variables).

From the heatmap we observe that there is a moderately positive correlation among many predictors. Some variables exhibit strong correlations, indicating potential redundancy in the data.

The correlation structure suggests that cortical thickness measurements in proximal brain regions may be highly related.

Scatter Plot

A scatterplot (Figure 2) visualizes the relationship between two randomly selected predictors, colored by the outcome variable (AD vs. C).

The plot confirms the heatmap's findings. The two predictors show a moderate positive correlation, aligning with the general trend observed in the heatmap. However, there is no clear separation between AD and Control groups in this two-dimensional space, suggesting that distinguishing AD from Control cannot rely on individual predictors alone.

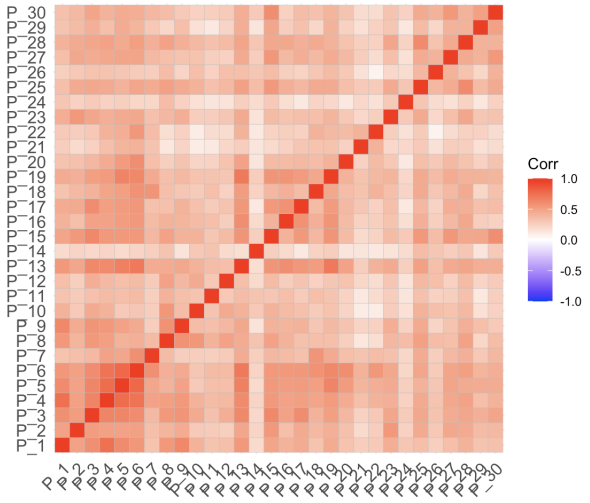

Figure 1: Correlation of First 30 Predictors

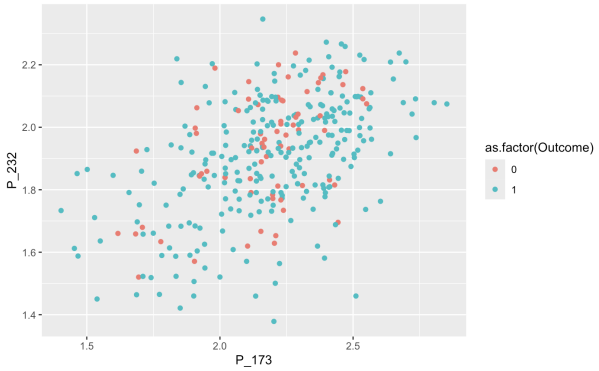

Figure 2: Scatter Plot of Two Random Predictors

2.2 Summary Statistics

Mean Comparison Table

Table 1: Mean cortical thickness values for AD and Control groups

Predictor	Control (C) Mean	AD Mean
P_1	1.655	1.679
P_2	2.112	2.081
P_3	1.794	1.801
P_4	1.857	1.864
P_5	1.937	1.925

By quickly scanning over the table, we can tell that the differences in mean values are minimal across predictors. No single predictor stands out as a strong differentiator between AD and Control groups.

This suggests that a single-variable threshold approach would be ineffective in distinguishing AD from Control cases, necessitating the use of multi-variable models.

Boxplot Comparison

Boxplots (Figure 3) illustrate the distribution of cortical thickness values for selected predictors across AD and Control groups. One can easily observe that there are significant overlaps between AD and Control distributions. Some predictors exhibit higher variability, but overall, differences between groups remain subtle. A few outliers exist, but they do not provide a strong distinguishing factor between AD and Control.

2.3 Guide to Modeling

The findings from this exploratory data analysis directly inform the next steps in modeling:

- Cortical thickness measurements are correlated, which suggests that feature selection or dimensionality reduction techniques are necessary to mitigate redundancy and improve model interpretability.
- AD and Control groups do not exhibit clear separation in individual predictors, reinforcing the need for multivariate modeling rather than relying on a single predictor.
- The class imbalance issue must be addressed to prevent a model from defaulting to predicting the majority class (AD).

Given these insights, the next step is to implement machine learning models that incorporate feature selection and handle multicollinearity effectively.

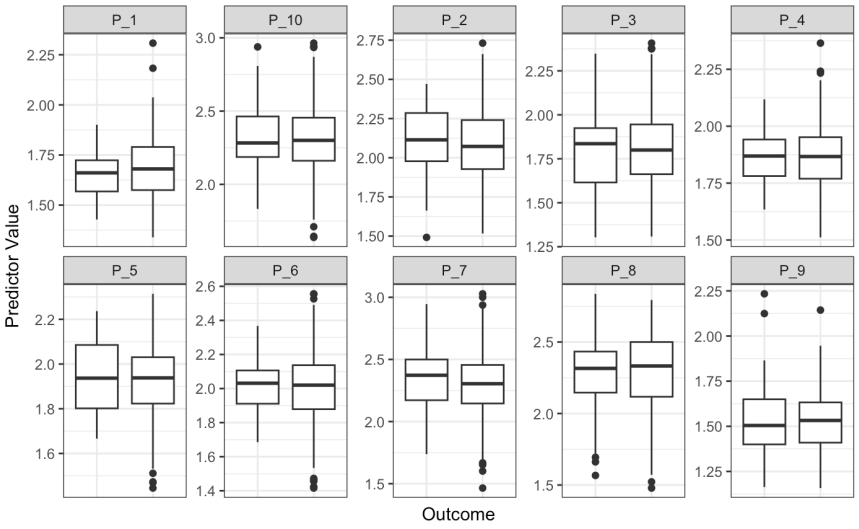

Figure 3: Boxplots of C (left) and AD (right) For First 10 Predictors

3 Generalized Linear Model (GLM)

The dataset was randomly split into training (250 observations) and testing (the remaining observations) sets. This ensures that the model is evaluated on unseen data to assess generalizability.

3.1 GLM Model without Regularization

Initially, a GLM model was fitted using all 360 predictors. However, the algorithm did not converge, likely due to the small sample size compared to the high-dimensional predictor space, as well as multicollinearity.

To examine the model's performance, the confusion matrices for the training and testing sets are provided in Table 2.

This significant drop in performance from training to testing indicates that the model memorized the training data rather than learning generalizable patterns. Thus, feature selection or regularization is necessary.

Table 2: Confusion Matrices for Unregularized GLM Model

Training Set	Actual		**Test Set**	Actual	
Predicted	0	1	**Predicted**	0	1
0	54	0	0	10	29
1	0	196	1	6	44

Training Accuracy: 1.0 **Testing Accuracy:** 0.607

3.2 LASSO-GLM

To address overfitting and high dimensionality, a LASSO logistic regression model was applied. This technique shrinks coefficients to zero, effectively performing feature selection.

A cross-validation process was performed to determine the optimal lambda value (Figure 4)

The new confusion matrices and accuracy scores for training and testing sets are shown in Table 3.

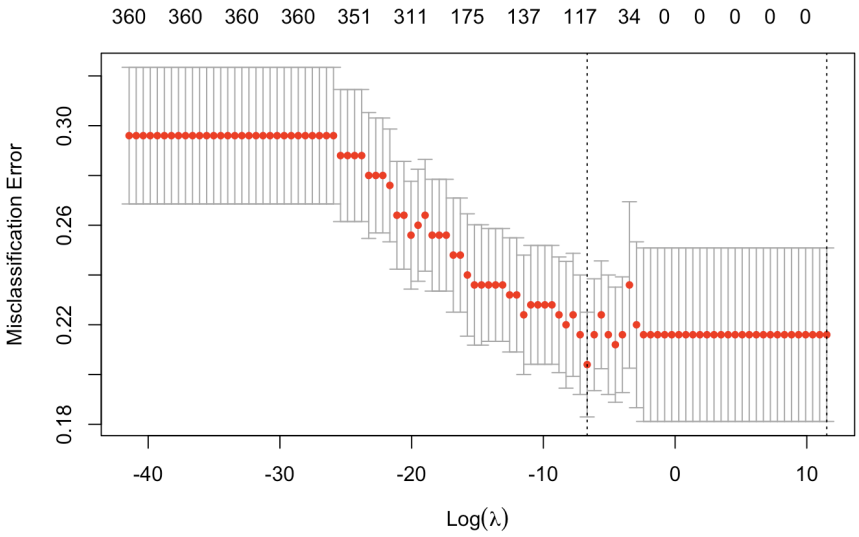

Figure 4: Boxplots of C (left) and AD (right) For First 10 Predictors

Table 3: Confusion Matrices for LASSO Model

Training Set	Actual	
Predicted	0	1
0	54	0
1	0	196

Test Set	Actual	
Predicted	0	1
0	9	16
1	7	57

Training Accuracy: 1.0 **Testing Accuracy:** 0.742

Compared to the original GLM model, LASSO significantly improves test accuracy, suggesting that removing redundant predictors helps generalization.

4 Tree-Based Models for Classification

We now examine tree-based models, decision tree and random forest. We will see that decision tree tend to either overfit or oversimplify, depending on hyperparameter choices. This is why to address these limitations, we explore random forest, which improves generalization by averaging multiple decision trees.

4.1 Decision Tree

A decision tree classifier was first fitted without hyperparameter tuning. Then, hyperparameters such as max_depth and min_samples_leaf were adjusted. The goal was to examine their impact on validation accuracy.

We observed changes in these hyperparameters led to slight variations in validation accuracy, but no clear pattern emerged. The choice of the validation set significantly affected accuracy, likely due to the small dataset size. This made it difficult to identify optimal hyperparameters.

To address this, a grid search cross-validation approach was used.

Grid Search: First Attempt

```
'min_samples_leaf': [1, 10, 20, 30, 40, 50, 60, 70, 80, 90]
'max_depth': [1, 3, 4, 5, 7, 10, 15, 20]
```

A 3D plot (Figure 5) visualizes how training and validation accuracy vary across different hyperparameter values.

We can observe that he best validation accuracy occurs when:

max_depth $= 1$ and min_samples_leaf ≥ 60

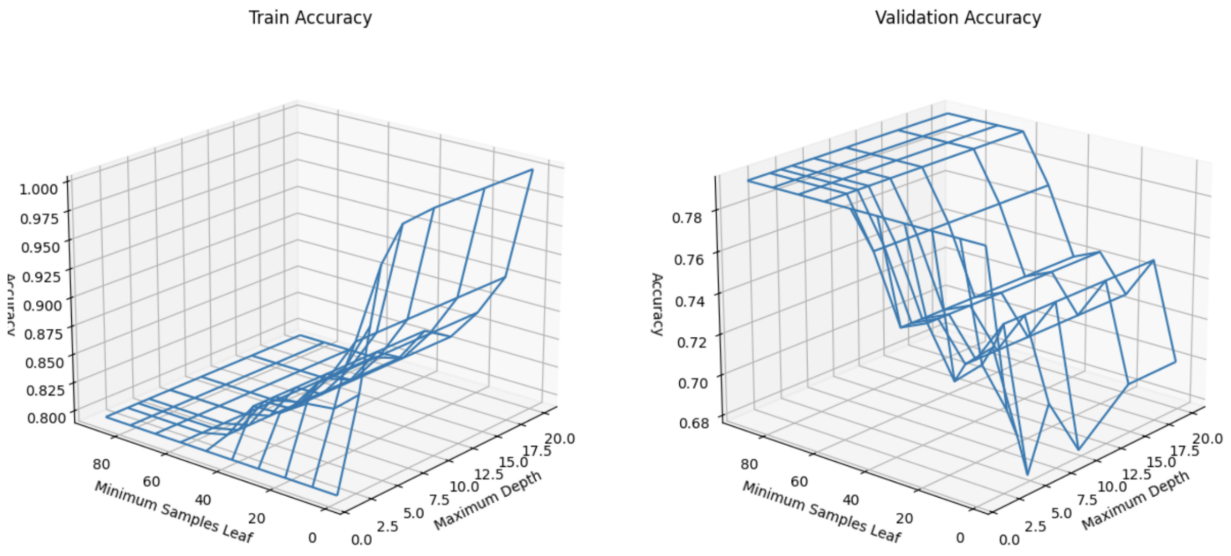

Figure 5: Gridsearch CV Results

Table 4: Confusion Matrices for Gridsearch Tree Model

Training Set	Actual		**Validation Set**	Actual	
Predicted	0	1	**Predicted**	0	1
0	211	0	0	58	0
1	60	0	1	10	0

Training Accuracy: 0.79 **Testing Accuracy:** 0.853

Note: The accuracy here really depends on how the data is split

Why did this happen? We can observe the confusion matrices in Table 4. The tree classified everything as the majority class (AD). With max_depth=1, the model can only make one split, leading it to default to the majority class. With min_samples_leaf ≥ 60, the tree struggles to make meaningful splits due to the small dataset.

From the grid search plot, validation accuracy does not improve with increasing max depth, indicating that the model cannot learn more flexible patterns from the data.

Grid Search: Second Attempt

The hyperparameter search was refined

```
'min_samples_leaf': [1, 10, 20, 30, 40, 50]
'max_depth': [3, 4, 5, 7, 10, 15, 20]
```

The problems with the previous parameters were resolved. Though the validation accuracy decreased, and the tree no longer defaulted to predicting only the majority class. This, in the long run, will improve the predicting performance, especially when the majority class is not AD anymore for actual data. The best accuracy achieved by a decision tree model using optimal hyperparameters was 0.76, slightly better than LASSO-GLM.

4.2 Random Forest

Random forest is particularly useful here because it naturally performs feature selection by training each tree on a subset of predictors, reducing the impact of multicollinearity. Additionally, random forest helps stabilize performance by averaging multiple trees, making it more robust to small dataset variations.

Instead of showing individual results, we directly compare the performance of decision trees vs. random forest across different max_depth values, as shown in figure 6.

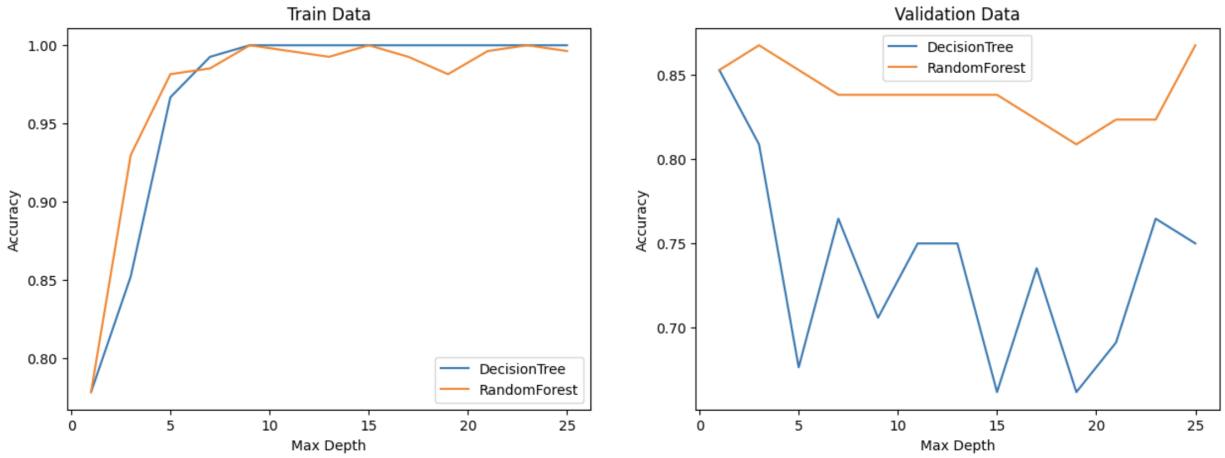

Figure 6: DT And RF Model Accuracy Comparison

Note: The validation accuracy for DT is highest at max_depth = 1 again, then it starts to decrease.

Training accuracy is similar for both models. While validation accuracy shows a stark difference. Decision trees fluctuate significantly, showing instability in validation performance. Random forest maintains high validation accuracy, reaching 0.863, without defaulting to predicting only the majority class.

This improvement suggests that random forest effectively generalizes to unseen data, making it the best-performing model among the models explored in this report.

5 Feature Selection

5.1 Comparing LASSO and Random Forest Feature Selection

To further refine our models and reduce overfitting, we explored feature selection by comparing the most important predictors identified by LASSO-GLM and random forest. LASSO, by design, shrinks some coefficients to zero, leaving 112 nonzero features. Meanwhile, random forest ranks features by importance, so we extracted the top 112 features for comparison. Interestingly, only 40 features overlapped between the two methods, suggesting that each model prioritizes different aspects of the data.

5.2 Evaluating Decision Tree and Random Forest with Selected Features

Using these 40 common features, we re-trained both decision tree and random forest models to assess whether feature selection improved generalization.

For decision tree, we conducted grid search to optimize hyperparameters again. The best validation accuracy achieved was 0.794, an improvement over the previous 0.76. This suggests that feature selection helped reduce overfitting, but the improvement was modest.

For random forest, the accuracy remained unchanged compared to using all features. This aligns with expectations, as random forest already performs built-in feature selection, meaning that reducing the feature set did not provide additional benefits.

6 Conclusions and Future Work

This study explored various machine learning approaches to classify Alzheimer's Disease (AD) using cortical thickness measurements. We began with generalized linear models (GLM) but encountered convergence issues due to high dimensionality. Applying LASSO-GLM improved performance by selecting relevant features, though overfitting remained a challenge.

To address this, we examined tree-based models, starting with decision trees, which struggled with stability and overfitting. We then implemented random forest, which significantly improved generalization, achieving the highest validation accuracy of 0.863, demonstrating the effectiveness of ensemble methods.

Further refinement through feature selection compared LASSO-identified features with random forest feature importance. Training models on the 40 overlapping features showed a modest improvement for decision trees (0.794 validation accuracy) but no significant change for random forest, reinforcing its built-in feature selection capabilities.

Despite these improvements, challenges remain. The class imbalance may have affected model performance, suggesting that techniques like oversampling, undersampling, or synthetic data generation (e.g., SMOTE) could improve results. Additionally, exploring alternative feature selection methods such as recursive feature elimination (RFE) or Principal Component Analysis (PCA) could refine model efficiency. Fine-tuning random forest hyperparameters and testing advanced models like gradient boosting (XGBoost, LightGBM) or neural networks may further enhance accuracy.

Overall, random forest emerged as the best-performing model, effectively balancing complexity, feature selection, and generalization. Future work should focus on refining feature selection, addressing data imbalance, and testing advanced models to improve predictive accuracy for Alzheimer's Disease classification.