Revanth Rao

Predicting Alzheimer's Disease Using Statistical Modeling

1. Introduction

This project utilizes a dataset from a Kaggle competition to analyze Alzheimer's disease risk factors using information about patient demographics, medical history, lifestyle patterns, symptoms, clinical measurements, and cognitive testing. These factors are then used to predict whether patients will receive an Alzheimer's diagnosis. This Kaggle competition addresses the important real-world issue of diagnosing Alzheimer's, a disease that affects over 7 million Americans today and around 1 out of every 9 people over the age of 65. Alzheimer's can be a devastating neurodegenerative disorder, but an early diagnosis can provide significant benefits for patients, as they are better able to access medication, arrange for care and other services, and make future plans while they have the mental capabilities to do so. As a result, it is critical to diagnose Alzheimer's early to ensure patients can take advantage of these benefits.

The purpose of this project is to use statistical modeling to answer the questions, "What is the ideal method to predict Alzheimer's disease in elderly patients, and are there certain medical characteristics that may be more indicative of a positive Alzheimer's diagnosis?" This project utilizes 11 statistical models to generate predictions for Alzheimer's disease and uses prediction accuracy rate and other evaluation metrics as well as 10-fold cross-validation to estimate predictive performance across various models. Finally, the models are compared to identify a model of choice, with model performance and interpretability as the key factors in deciding on a final model.

2. Data Overview and Exploratory Data Analysis (EDA)

The dataset used for this project comes from a Kaggle competition called “Alzheimer's Disease Risk Prediction - EU Business”, and the link to the competition can be found in the References section. This Kaggle competition included a train and a test set; however, as the test set did not include the response variable, the train set was used as the dataset for this project. The train set initially included 1719 observations and 35 variables, but after further examination of the data, the variables PatientID and DoctorInCharge were dropped as they did not provide any meaningful information for this project. Additionally, there were no null values or missing data in the dataset, allowing for cleaner data visualization and processing.

Of the 33 remaining variables, 32 were predictor variables and the remaining variable was the response variable Diagnosis, which took on a value of 1 if a given patient was diagnosed with Alzheimer’s and 0 if not. In addition, according to the dataset description from Kaggle, the 32 predictor variables were grouped into the following categories: demographic details, lifestyle factors, medical history, clinical measurements, cognitive and functional assessments, and symptoms. Demographic details are information about age, gender, and ethnicity. For this dataset, patient ages ranged from 60 to 90, there was almost an equal split of male and female patients, and patients were grouped based on whether they were Caucasian, African American, Asian, or other ethnicities. Lifestyle factors included information about diet, exercise, smoking, BMI, and alcohol consumption. Medical history included five variables that indicate whether a patient was diagnosed with a certain illness, as well as whether they had a family history of Alzheimer’s. Clinical measurements provide information about six different cholesterol and blood pressure measurements. Cognitive and functional assessments include test scores for cognitive ability and indicators for whether a patient had memory and behavioral issues. Finally, symptoms included five variables that indicate whether a patient had certain mental impairments that can be seen in people with Alzheimer's disease.

To begin EDA, I first examined the data using summary statistics to understand patterns. The summary statistics for the continuous predictor variables indicated that essentially all variables were not skewed, and this was confirmed by the boxplots of these variables, shown in Figure 1. These plots appear symmetric, as the area of the boxes is roughly consistent above and below the median. Certain variables such as AlcoholConsumption did appear to show some slight skewness, as shown in Figure 2, which displays a histogram of alcohol consumption by patient. This histogram seems to show that more patients consume less than 10 units of alcohol per week, a fairly logical conclusion that indicates a very slight right skew of the data. However, this skew is not large and does not indicate the presence of outliers. In general, while slightly surprising, it is somewhat logical that there is not a clear pattern of skewness in the data. This is because many of the variables have predefined ranges (for example, BMI in this study ranges from 15 to 40), so there is less room for outliers and heavy tails in the data, making it less prone to skewness.

In addition, I also examined the predictor variables to determine whether a transformation of these continuous predictor variables was ideal. However, largely due to the relative symmetry of the distributions of these variables, both a Box-Cox transformation and normalization of these predictor variables did not improve model performance. Additionally, checking for outliers using both interquartile range and z-scores showed no outliers among predictor variables according to these metrics. This is also displayed in the boxplots, which do not show points past the whiskers. As such, it was determined that using the raw, untransformed data was appropriate for this task and that it was appropriate to keep every data point in the dataset.

The next step in EDA was examining the correlations between variables. I began by identifying the five variables with the largest positive correlation and the five variables with the largest negative correlation with Diagnosis; these correlations are shown in Table 1. Overall, it appears that none of the variables show a particularly strong positive correlation with Diagnosis. However, the variables FunctionalAssessment, ADL, and MMSE show fairly large negative correlations with Diagnosis, indicating that as the values of these variables increase, a patient is more likely to be diagnosed with Alzheimer's. Moreover, I also looked at the correlations between the predictor variables, and found that each continuous predictor had a very weak correlation with the other predictors, meaning that there was a low chance of multicollinearity in any potential model.

After examining the distributions of the continuous predictor variables, I then explored the discrete variables in the dataset. The first variable explored was Diagnosis, which showed that in the dataset, 1112 patients were not diagnosed with Alzheimer's and 607 were diagnosed with Alzheimer's. This indicates that there was a slight class imbalance in the response variable. As such, it is important to ensure that statistical models accurately capture patients who were diagnosed with Alzheimer's and accuracy metrics not only capture overall prediction accuracy but also accuracy by group. Additionally, one variable that stood out was Ethnicity, which indicated that 1020 of the 1719 people in the dataset were Caucasian. This is another instance of class imbalance in a variable, as there were only 366 African American patients and 165 and 168 patients of Asian and other races, respectively. The class imbalance may also lead to potential model inaccuracies as well, as Alzheimer's prevalence does vary by ethnicity; according to the Alzheimer's Association, African Americans are about twice as likely and Hispanics are about one and a half times as likely to have Alzheimer's or other dementias than Caucasians. Thus, potential models should be able to address this imbalance as well. Finally, one more interesting variable was FamilyHistoryAlzheimers, as 1290 patients did not have a family history of Alzheimer's. However, when looking at family history versus diagnosis, about 36.12% of people without a family history were diagnosed with Alzheimer's and about 32.87% of people with a family history were diagnosed, so there does not appear to be a connection between family history and diagnosis of Alzheimer's.

The next step of EDA was identifying how Diagnosis differed based on the values of the predictor variables. Many variables did not show a clear difference in Diagnosis based on the predictor's value, but three stood out. As shown in the first two plots in Figure 3, patients who experienced behavioral problems or memory complaints were far more likely to be diagnosed with Alzheimer's than those who did not experience these health issues. This certainly makes sense, as these are two common symptoms in patients with Alzheimer's. Additionally, the third plot in Figure 3 displays the proportions of diagnoses based on ethnicity and shows that a lower proportion of African Americans in this dataset were diagnosed with Alzheimer's than any other ethnicity while a higher proportion of Asian Americans were diagnosed with Alzheimer's. This is interesting in that it goes against general research findings, but it is important to note that these proportions are all fairly close to each other and the sample size for each ethnicity is somewhat small, so this information should not seriously affect the statistical modeling process.

The final piece of EDA was exploring how the distributions of the continuous predictors differed based on diagnosis. To understand this, I generated violin plots which produced fairly interesting results. Figure 4 displays three violin plots that appear to show notable differences in the metrics MMSE, Functional Assessment, and ADL between patients who were not diagnosed with Alzheimer's and patients who were diagnosed with Alzheimer's. The violin plots all show that patients who were diagnosed with Alzheimer's tended to have lower values for these three metrics, which makes sense given that lower values indicate greater mental impairment. On the other hand, one violin plot that produced somewhat unexpected results was the plot of DietQuality grouped by diagnosis, shown in Figure 5. This plot shows that patients who were diagnosed with Alzheimer's had a slightly higher median diet quality score. While people may assume that better diet quality decreases Alzheimer's risk, an article from the National Institute on Aging notes that current research does not show a clear connection between diet and cognitive decline, meaning that the violin plot does not necessarily display an unusual trend.

3. Methodology

For this project, I utilized 11 statistical methods in total, but I will discuss five methods in detail for clarity. The four methods that easily offered the best prediction performance were the decision tree, boosting, Random Forest, and XGBoost methods, all of which are decision tree-based models. The model that easily offered the worst prediction performance was the K-Nearest Neighbors (KNN) method. To evaluate each model, 10-fold cross-validation was used to estimate prediction performance on unseen data, and accuracy rate, precision, recall, and F1 score were used to quantify model performance.

As noted above, the four highest-performing models employed a decision tree-based framework. A decision tree is a model that splits the data into regions, or branches, based on the values of the predictor variables. The regions are created by choosing a certain predictor, selecting a threshold value (denoted as "s"), and splitting the data based on if a given data point had a value greater than "s" or less than "s" for the chosen predictor. The data splitting process is repeated until there are only a small number of observations within each branch or when splitting does not meaningfully improve model performance. Finally, after all the splits are completed, each branch is assigned a value of the response variable based on a majority vote of the data points in the branch. For example, if a branch has three data points with a value of 1 for Diagnosis and two points with a value of 0, then the diagnosis prediction for the branch would be 1. The decision tree is then used to predict on unseen data. Decision tree models are highly useful methods for this project as they are fairly intuitive to understand, provide greater interpretability than other more complex models, and generally have high performance for prediction tasks.

This project utilizes a simple decision tree as one model along with boosting, XGBoost, and Random Forest, which are all variations of the decision tree discussed above. Boosting is a method that starts by creating a small decision tree and making predictions with this small tree. Then, after examining the prediction errors made by the small tree, another tree is generated to help correct errors from the small tree. The new tree is then used to make predictions, and this process is repeated several times, after which the predictions are combined into one final model that is used to make predictions on the unseen data. Boosting has an advantage over a singular decision tree model because building trees with the intention of improving upon errors reduces bias for the model and improves prediction accuracy. XGBoost is a special type of boosting which builds on traditional gradient boosting by using techniques such as regularization and decision tree pruning to improve model performance and accuracy. Random Forest uses a slightly different approach to build on the decision tree framework, as it begins by creating many bootstrapped samples, or samples drawn with replacement from the original data that also have the same sample size as the original data. Then, a decision tree is fit for each bootstrapped sample, with a key distinguishing feature being that only a few randomly selected predictors are considered each time the data is split. Finally, each decision tree produces predictions, and these predictions are combined to get the final prediction. Random Forest models generally produce more accurate predictions than a single decision tree because they reduce variance by using several decision trees and control overfitting by limiting the number of predictors available when splitting the data.

Finally, the last model to be discussed is KNN, a nonparametric method. When analyzing a test data point to generate a prediction, KNN finds the K closest training data points to the test data point using the predictor variables. Then, KNN identifies the value of the response variable for each nearest neighbor, and labels the test point based on the majority of response variable values from the neighbors. KNN is a very intuitive method and is easy to explain and understand. However, it is a simplistic method and does not always provide highly accurate results.

To evaluate the models and compare model performance, the metrics accuracy rate, precision, recall, and F1 score were used. Accuracy rate, the simplest metric of the four, measures the percentage of correct predictions. Precision provides the percentage of predicted positives that were actually positive; for this project, precision identifies the percentage of people predicted to have Alzheimer's who actually do have Alzheimer's. Recall measures the percentage of actual positives that were captured by the model's predictions; for this project, recall identifies the percentage of people who have Alzheimer's who were also predicted to have Alzheimer's. Finally, F1 Score gives the harmonic mean of recall and precision, allowing for one statistic that captures both metrics together. These four metrics are appropriate for this project because they present measures of prediction accuracy in terms of percentages, which is necessary in a classification setting where the response variable Diagnosis is discrete.

To ensure that model performance across methods could be properly and fairly compared, 10-fold cross-validation was used to calculate the four accuracy metrics for each model. 10-fold cross-validation works by splitting the data into 10 even folds, then using nine folds to train a model and the held-out fold to make predictions and calculate the accuracy metrics. This process is repeated 10 times so each fold is used for predictions, and the 10 sets of accuracy metrics are averaged to produce the final metrics for each model. Additionally, a random seed was used when implementing cross-validation so that the folds were identical for each method, allowing for consistent comparisons between models. Cross-validation is appropriate for this project because it uses the given dataset to estimate prediction performance on unseen data without needing a dedicated test set.

4. Results

As noted in Methodology, the decision tree, gradient boosting, XGBoost, and Random Forest methods produced the most accurate predictions for Alzheimer's diagnosis, while KNN produced the least accurate predictions. When using 10-fold cross-validation, the decision tree model produced an accuracy rate of about 94.7%, while the gradient boosting, XGBoost, and Random Forest models all produced accuracies above 95%, meaning all four models correctly predicted the diagnosis for a given patient at least 94% of the time. Additionally, as expected, the more simplistic decision tree model had a slightly lower accuracy than the more sophisticated models which used many decision trees to create predictions. The decision tree also had a lower precision at around 92.8% while the other methods had a precision between 94% and 95%. This means that for these three methods, between 94 and 95% of people predicted to have Alzheimer's did have Alzheimer's. While the decision tree had the lowest precision, it did have a recall of about 92%, higher than both Random Forest and gradient boosting and slightly lower than XGBoost at 92.4%. In the context of this project, around 92% of people who have Alzheimer's were also predicted to have Alzheimer's by the decision tree and XGBoost methods. Based on the results for precision and recall, the decision tree method had the lowest F1 score, while XGBoost had the highest F1 score. A detailed breakdown of the results for all four models is available in Table 2.

On the other hand, the KNN model with K = 3 easily had the lowest value for each of the four performance metrics among all models. KNN achieved the lowest accuracy at 72.1%, while it had a precision of 64.5% and a recall of 48.4%, meaning less than half the patients who have Alzheimer's were actually predicted to have Alzheimer's by KNN. This resulted in an F1 score of only 54.8%. The full table of results for KNN can be found in Table 3.

The results for the remaining six methods can be found in Table 4. These methods all displayed metrics which were substantially better than the KNN method, but worse than the four decision tree-based methods. As such, the results from the decision tree-based methods and KNN will be discussed further.

In addition to the four metrics discussed above, I also created a decision tree, Random Forest, and XGBoost model using the full dataset to understand which predictor variables were important in building the models. For each model, it was clear that the variables FunctionalAssessment, MMSE, ADL, BehavioralProblems, and MemoryComplaints were the most critical predictors. The decision tree split the data using only these five variables while Random Forest and XGBoost identified these five as the most important variables. Table 5 presents the top five most important variables according to the XGBoost model, with FunctionalAssessment, the most important predictor, having an importance of 100 and the importance of other variables scaled based on FunctionalAssessment. As shown, MMSE and ADL are slightly less important than FunctionalAssessment, while MemoryComplaints and BehavioralProblems were less important than the first three predictors, but far more important than any of the remaining predictors. The first three predictors mentioned also clearly match the results from the EDA, as FunctionalAssessment, MMSE, and ADL all displayed clear differences when comparing patients who were diagnosed with Alzheimer's and patients who were not diagnosed. It is also not surprising that the variable MemoryComplaints was identified as important because Alzheimer's is a disease that causes memory loss over time, so patients with Alzheimer's will be more likely to have memory problems. Moreover, behavioral issues like anxiety, irritability, depression, violence, and other unusual behaviors are frequently observed in people with Alzheimer's, so it is understandable that elderly patients with behavioral problems are more likely to have Alzheimer's.

When comparing model performance, it certainly makes sense that decision tree-based methods produce the most accurate predictions. Decision trees are highly flexible and versatile, and can easily adapt to handle nonlinear patterns in data. Random Forest and boosting methods increase prediction accuracy by using multiple trees to limit overfitting the training data, and XGBoost builds on a typical boosting model by adding complexity and sophistication to create a model that can better capture trends in the data. As a result, these four models are generally highly accurate for classification tasks, and XGBoost in particular is used regularly to generate winning submissions for Kaggle competitions, which is relevant given that this dataset comes from such a competition.

Conversely, it also makes sense that KNN has the worst prediction performance. KNN relies on the nearest neighbors to a test data point to make predictions; however, given that this data is very high dimensional because it uses over 30 predictors, these neighbors are often far away. This leads to lower prediction accuracy as the neighbors are not as closely related to a test point in higher dimensions, a concept known as the curse of dimensionality. Additionally, KNN is a nonparametric method that only uses nearby data points to make predictions for test data rather than building a model using patterns identified in the training data, which can lead to inaccuracies in prediction.

Ultimately, the four decision tree-based models are clearly superior to every other method tested in this project. However, it is not immediately clear which one would be considered the best model. XGBoost has the highest prediction accuracy, recall, and F1 scores, but it is the least interpretable of the four options. Random Forest and gradient boosting have accuracies which are fairly close to the XGBoost accuracy and also have the two highest values for precision, but have the two lowest values for recall. The decision tree has the lowest accuracy, precision, and F1 score, but does have the second-highest recall and is the easiest model to interpret. In the end, however, XGBoost is the model of choice for this project. XGBoost was generally the most accurate model of the four and was especially appealing because it had the best recall. Recall is particularly important when predicting medical conditions because it measures the percentage of patients with a disease who were also predicted to have the disease. As mentioned earlier, properly diagnosing patients with Alzheimer's is crucial because it allows patients to understand their health status as well as set up care and make plans for the future. Thus, it is vitally important to ensure that false negative diagnoses are avoided, and prioritizing a model with high recall allows for this. As a result, despite its lesser interpretability, XGBoost is the optimal model for this task given its extremely high recall value.

5. Conclusion

This project presents a set of statistical methods used to predict Alzheimer's diagnosis in a group of patients. Utilizing multiple accuracy metrics and 10-fold cross-validation, I tested 11 statistical methods to understand the models that performed best in predicting Alzheimer's, as well as the most critical variables for predicting Alzheimer's. Ultimately, four decision tree-based models stood out, with XGBoost being the model of choice for prediction given its high recall, an important metric in the context of medical diagnoses.

Ultimately, this project displays strengths and limitations in its analysis of the Kaggle dataset. One strength is the use of numerous statistical methods, which gave a comprehensive understanding of prediction accuracy for different types of models. Additionally, using multiple accuracy metrics to measure the performance of each model increased the depth of the analysis and enabled better comparisons between models. Finally, employing cross-validation allowed me to use the entire dataset when measuring accuracy for each model and it allowed for fair comparisons of accuracy between each model. These strengths increase the credibility of the results and make them more generalizable to different datasets.

On the other hand, this project did demonstrate certain limitations in its analysis. The dataset from Kaggle contained less than 2,000 observations, so the sample size of patients was fairly small. As a result, an analysis of this dataset could produce results which may not hold when applied to a larger dataset. The dataset was also fairly clean and required minimal preprocessing, meaning that the models created may not perform as well when predicting real-world data that includes outliers and other unusual data. Finally, the XGBoost model chosen as the model of choice is a highly accurate model for many purposes, but it sacrifices some interpretability to achieve higher accuracy. As such, it is harder to explain XGBoost to someone without a statistical background than it is to explain a simpler model like a decision tree. These limitations certainly present some challenges in properly generalizing the findings of this project. One way to address these limitations and extend the project would be to use a larger dataset with more outliers and noise, which would help improve the quality of the analysis. In addition, having a designated test dataset with labels for the response variables would eliminate the need for cross-validation, and provide cleaner results that are easier to understand.

Overall, this project presents a robust analysis of predicting the diagnosis of Alzheimer’s for patients using a Kaggle dataset. The analysis presented is an excellent starting point in utilizing machine learning models to predict Alzheimer’s, and can be extended to datasets of increasing complexity to create more comprehensive and predictive models which can ultimately make a difference in accurately diagnosing this devastating disease in elderly individuals.

Figures and Tables

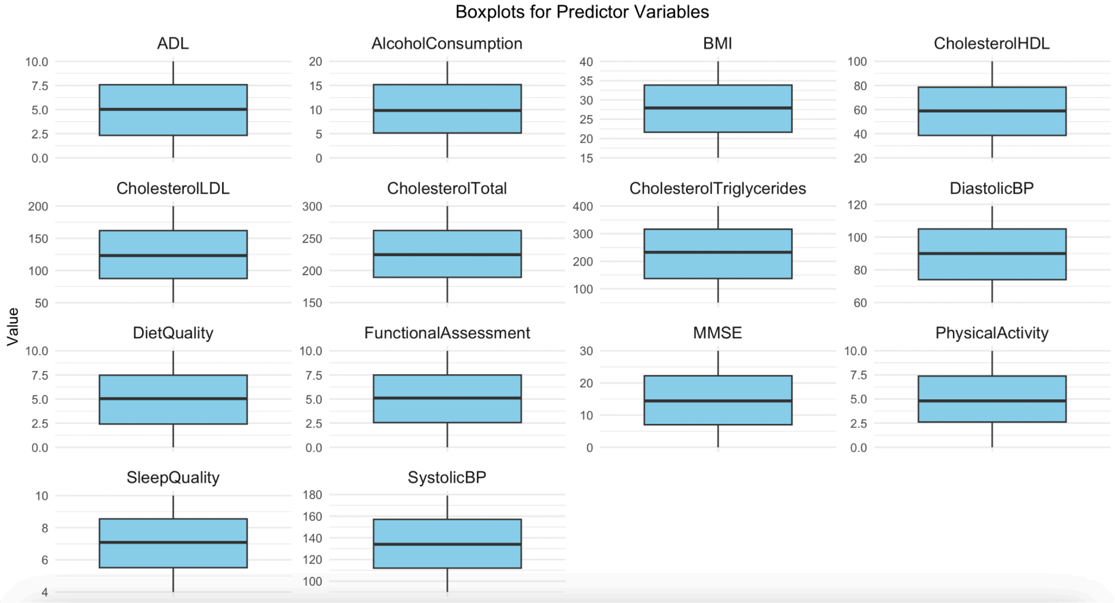

Figure 1: A boxplot of every continuous predictor variable. These boxplots show that these predictors are not clearly skewed.

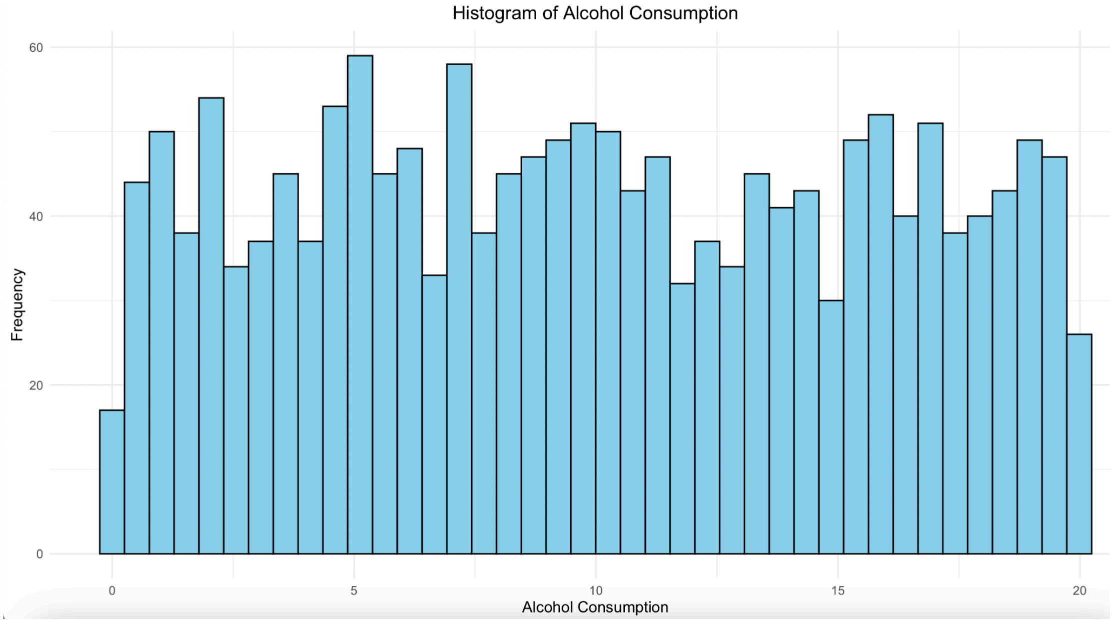

Figure 2: A histogram of AlcoholConsumption. This histogram shows that the distribution of alcohol consumption among patients was very slightly right-skewed.

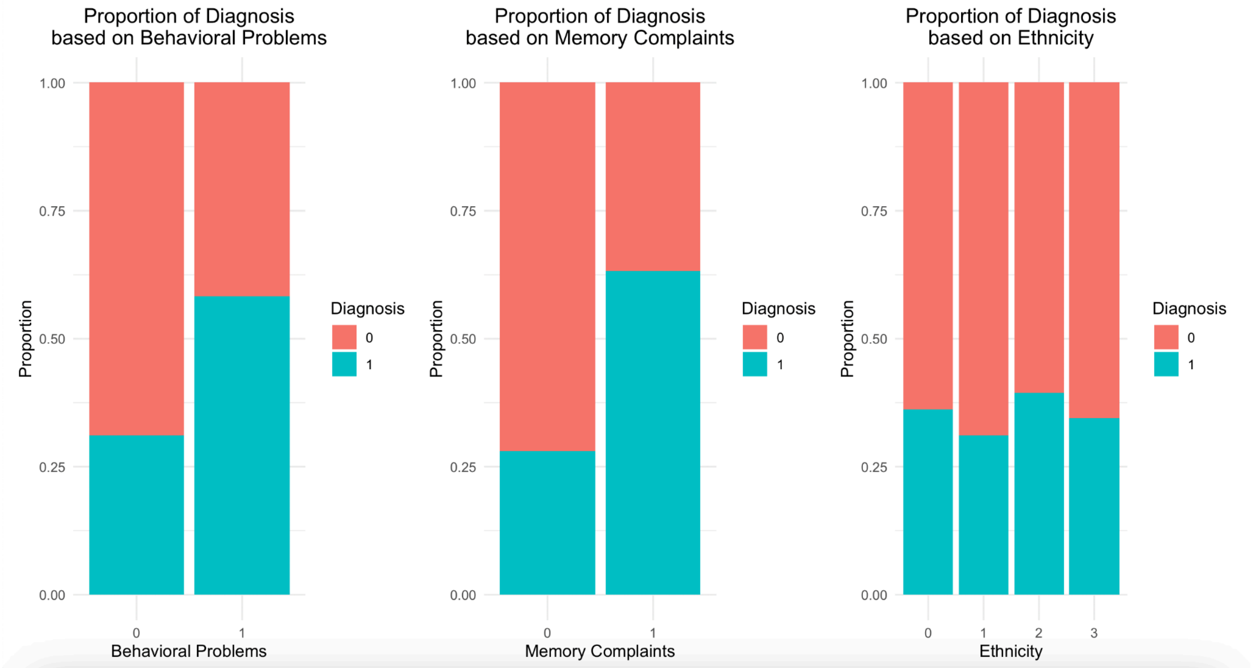

Figure 3: Bar graphs displaying the proportions of diagnosis based on Behavioral Problems, Memory Complaints, and Ethnicity. These graphs show a clear increase in diagnosis for people who experience memory complaints and behavioral problems.

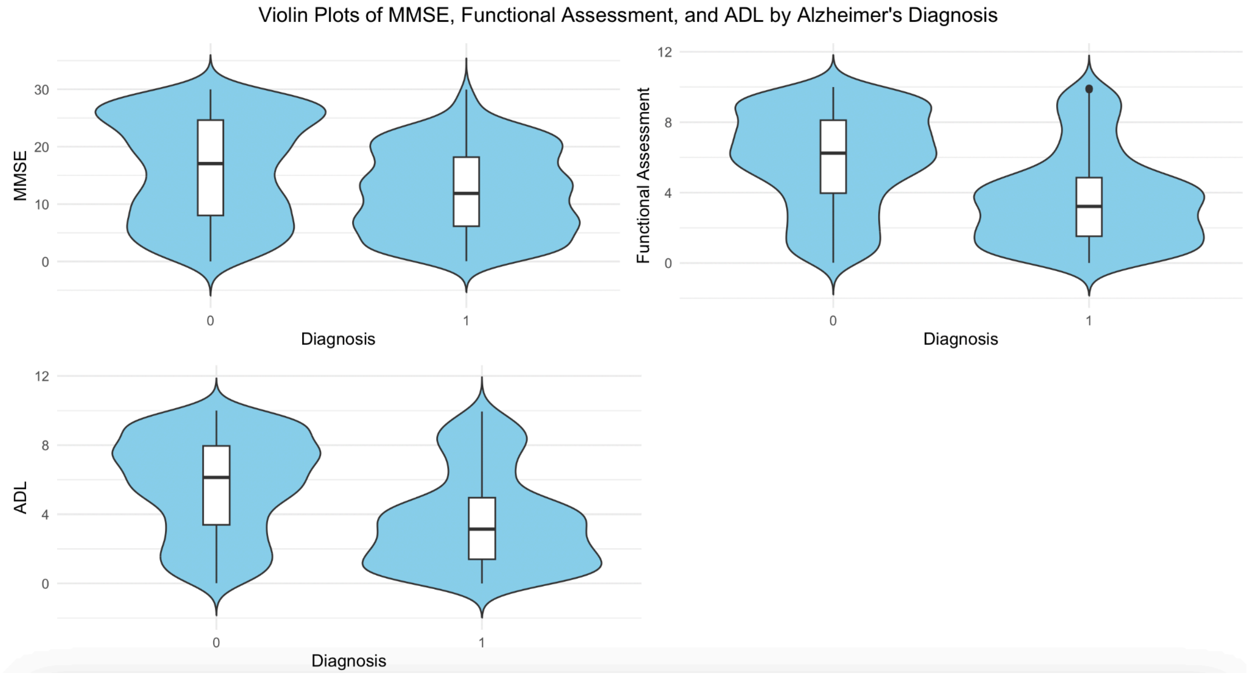

Figure 4: Violin plots of MMSE, Functional Assessment, and ADL scores grouped by Alzheimer's diagnosis. All three plots indicate that patients diagnosed with Alzheimer's produced lower values for the metrics.

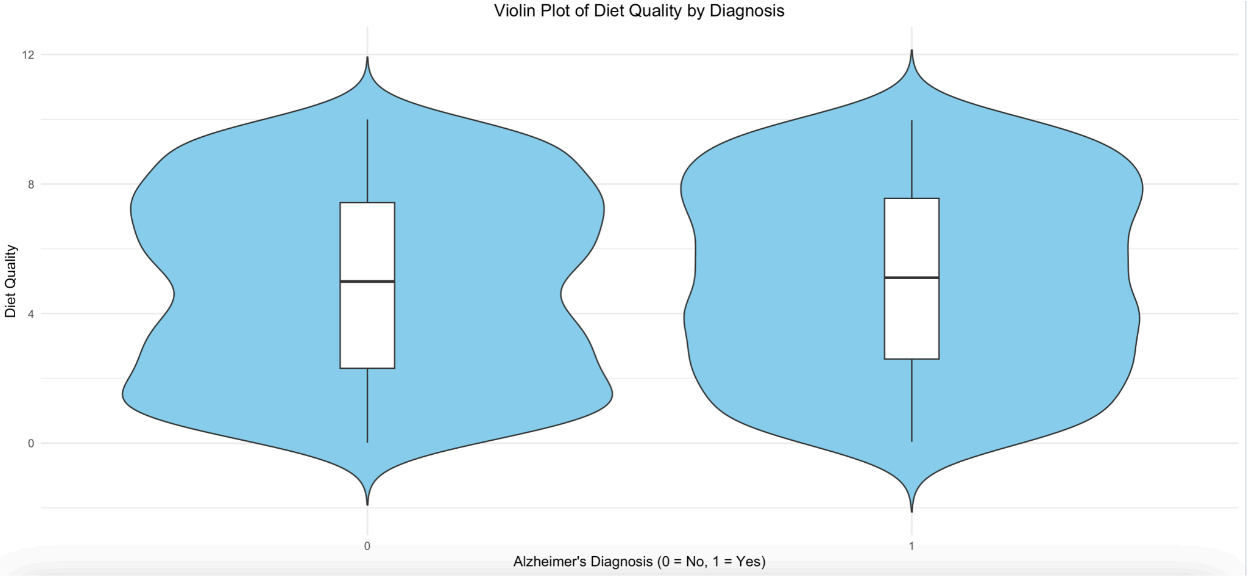

Figure 5: A violin plot of diet quality grouped by Alzheimer's diagnosis. This plot shows that patients who were diagnosed with Alzheimer's disease had a slightly higher median diet quality.

Top 5 Positive Correlations		**Top 5 Negative Correlations**	
Variable	**Correlation**	**Variable**	**Correlation**
CholesterolHDL	0.04588765	FunctionalAssessment	-0.37791309
DiastolicBP	0.02975557	ADL	-0.33981075
CholesterolTriglycerides	0.02923249	MMSE	-0.22462676
DietQuality	0.02220177	SleepQuality	-0.04542621
BMI	0.02050575	SystolicBP	-0.03163637

Table 1: The top 5 most positively and negatively correlated variables with Diagnosis. All positive correlations are fairly small, but FunctionalAssessment, ADL, and MMSE display notable negative correlations with Diagnosis.

Method	Accuracy	Precision	Recall	F1 Score
Decision Tree	0.9470556	0.9284872	0.9199427	0.9237140
Random Forest	0.9511247	0.9469676	0.9143169	0.9293707
Gradient Boosting	0.9505399	0.9495308	0.9093716	0.9283453
XGBoost	0.9534470	0.9443295	0.9241803	0.9333637

Table 2: Accuracy, Precision, Recall, and F1 Score for four different decision tree-based methods. These methods achieved the best results for each of the four metrics described above.

Method	Accuracy	Precision	Recall	F1 Score
KNN	0.7213416	0.6452139	0.4842564	0.5483939

Table 3: Accuracy, F1 Score, Precision, and Recall for K-Nearest Neighbors (KNN) with K = 3. This method produced the least accurate results for every metric among the models tested.

Method	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.8376921	0.7942684	0.7389551	0.7608921
Logistic Regression Best Subset Selection	0.8458316	0.8074641	0.7489675	0.7736542

LASSO	0.8400245	0.8037756	0.7324925	0.7624772
Ridge	0.8528118	0.8267492	0.7383922	0.7770027
LDA	0.8388481	0.7908340	0.7466384	0.7643653
QDA	0.7940500	0.7118531	0.7027692	0.7037105

Table 4: Accuracy, F1 Score, Precision, and Recall for the remaining methods. These methods produced better results than KNN, but lesser results than the decision tree methods.

Variable	**Importance**
FunctionalAssessment	100
MMSE	93.7474
ADL	88.0308
MemoryComplaints	57.8066
BehavioralProblems	38.6813

Table 5: Top 5 most important variables based on the XGBoost model. The importance metric is scaled such that the most important variable FunctionalAssessment has an importance of 100.

References

Asian Americans and Pacific Islanders and Alzheimer's. (2021). Alzheimer's Disease and Dementia. https://www.alz.org/help-support/resources/asian-americans-and-alzheimers

National Institute on Aging. (2023, November 20). *What Do We Know About Diet and Prevention of Alzheimer's Disease?* National Institute on Aging. https://www.nia.nih.gov/health/alzheimers-and-dementia/what-do-we-know-about-diet-and-prevention-alzheimers-disease

National Institute on Aging. (2024, July 11). *Alzheimer's Caregiving: Managing Personality and Behavior Changes*. National Institute on Aging. https://www.nia.nih.gov/health/alzheimers-changes-behavior-and-communication/alzheimers-caregiving-managing-personality-and

Race, Ethnicity, and Alzheimer's. (2020, March). Alzheimer's Association International Conference. https://aaic.alz.org/downloads2020/2020_Race_and_Ethnicity_Fact_Sheet.pdf

SFAR, H. (2025, February 15). *Alzheimer's Disease Risk Prediction - EU Business*. Kaggle. https://www.kaggle.com/competitions/alzheimers-disease-risk-prediction-eu-business/

Github link to project code: https://github.com/revrao/Alzheimers-Prediction-Project