

Alzheimer's disease is often undiagnosed until symptoms appear. Therefore, creating an accurate model is crucial for identifying those most at risk of developing the disease. This project compares the performance of multiple logistic regression models to determine the factors and the best model using data from patients with and without Alzheimer's [1].

Key Recommendations

This project evaluates key risk factors to determine a patient's likelihood of developing Alzheimer's disease; specifically, age, family history, and APOE genotype. By analyzing these variables, identifying individuals at higher risk and providing insights into early detection. Additionally, a GLM model is incorporated to predict the presence of Alzheimer's, leveraging its ability to handle complex interactions between risk factors. This approach enhances predictive capabilities and offers an interpretable model.

Data Preparation and Exploration

The first step in data preparation is to categorize the countries in the dataset by continent. This reduces the number of factors that predictors must account for, resulting in a simpler model. Next, we will examine the predictor variables to determine whether the dataset contains a roughly equal number of cases for both classes: individuals with Alzheimer's and those without. Figure 1 displays the counts for each class in the dataset.

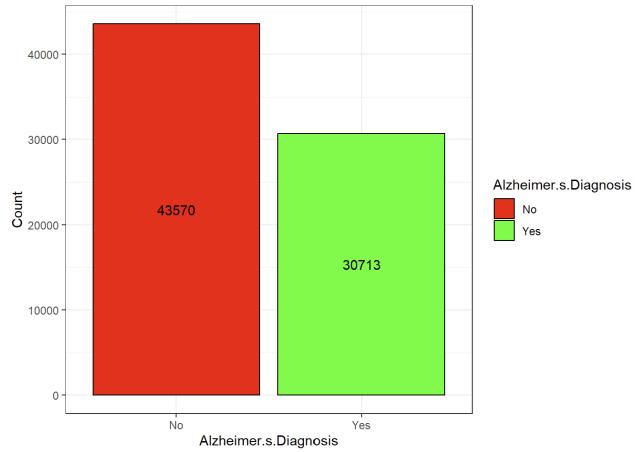


Figure 1. Class Counts for Diagnosis

Since there are significantly more cases of individuals without Alzheimer's, we will adjust the class weights to account for this imbalance. An Alzheimer's diagnosis will be assigned a weight of 1.4, while a non-diagnosis will have a weight of 1.0.

The categorical predictor variables were then converted to numerical values for data analysis. Finally, all predictor variables were centered and scaled to ensure they were on the same scale, while the target variable, Alzheimer's diagnosis, remained unchanged.

Variable Selection

Baseline

Three methods were selected to determine the key variables that influence the prediction of Alzheimer's. The first was simply to have no variable selection at all, thus allowing for a baseline in each model to compare the performance.

Stepwise

The second uses stepwise variable selection with a forward-backward direction. This method will add and remove a variable at a time until the AIC is optimized, giving the most relevant

variables based on the full model. The process reduced the total number of variables from 24 to 5, including “Age”, “Cholesterol Level”, “Family History of Alzheimer’s”, “Genetic Risk Factor APOE ε4 allele”, “Stress Levels”, “Urban vs Rural Living”.

Bayesian Network

The final method of variable selection utilizes Bayesian Networks, which use conditional probabilities to determine relationships between variables and represent them graphically. The first Bayesian Network tested, hpc, employs a constraint-based structure to identify connections between predictors; however, this approach produced an undirected graph and will not be used for variable selection. Figure 2 displays the resulting diagram.

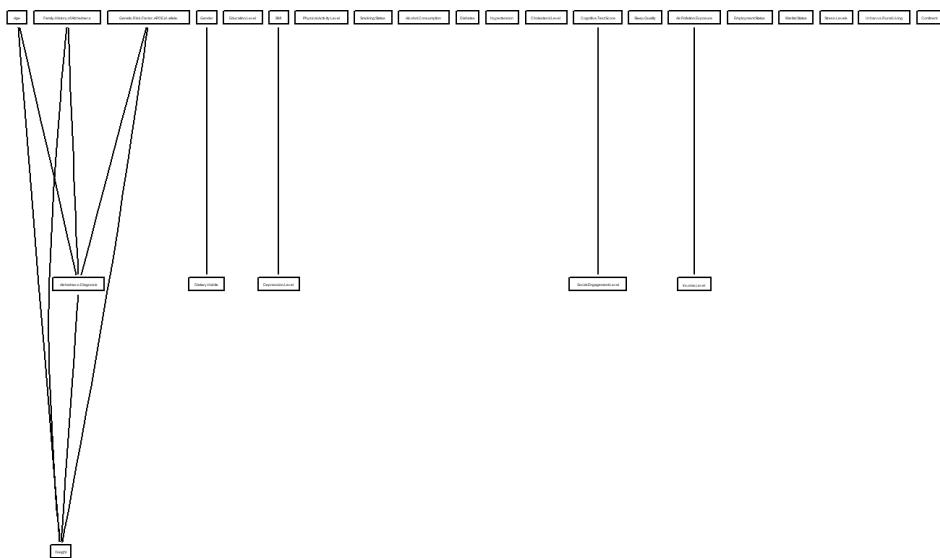


Figure 2. Constraint-Based Bayesian Network

The second Bayesian Network tested, h2pc, utilizes a score-based algorithm to determine connections between predictors. The resulting network linked Alzheimer’s Diagnosis with “Age”, “Family History of Alzheimer’s”, and “Genetic Risk Factor (APOE ε4 allele)”. Figure 3

illustrates the diagram generated by this algorithm.

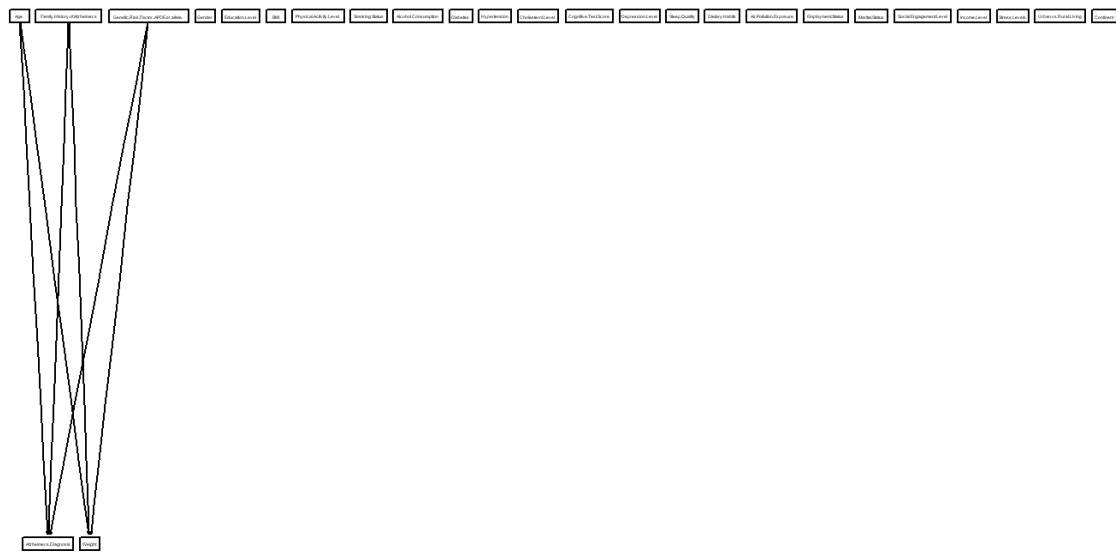


Figure 3. Score-Based Bayesian Network

The third Bayesian Network tested, hc, employs a hybrid-based algorithm to determine predictor connections. This approach yielded an identical model to the score-based algorithm. Figure 4 presents the corresponding diagram.

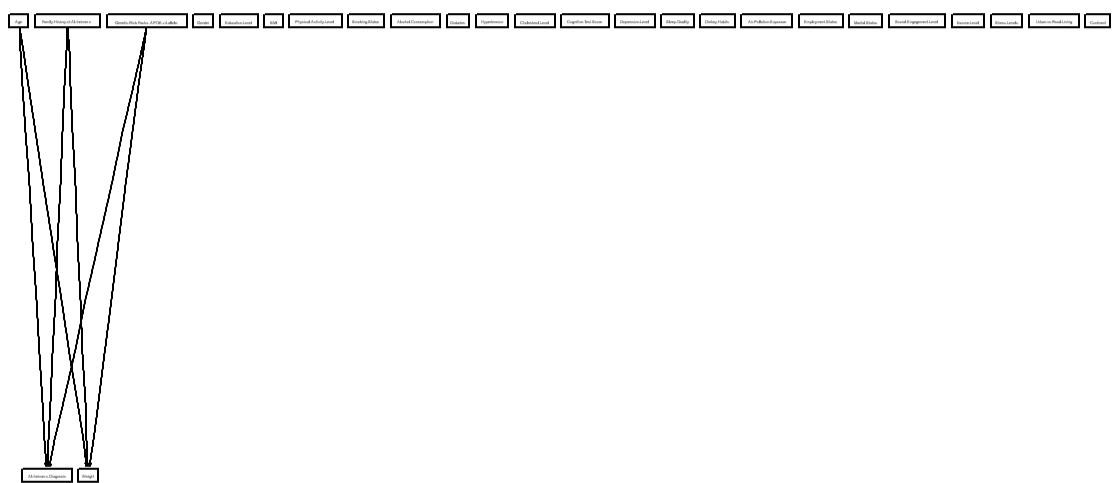


Figure 4. Hybrid-Based Bayesian Network

Finally, arcane utilizes a local-discovery-based algorithm to identify connections between predictors. However, like the constraint-based model, it produced an undirected graph and will not be used for variable selection. Figure 5 shows the resulting diagram.

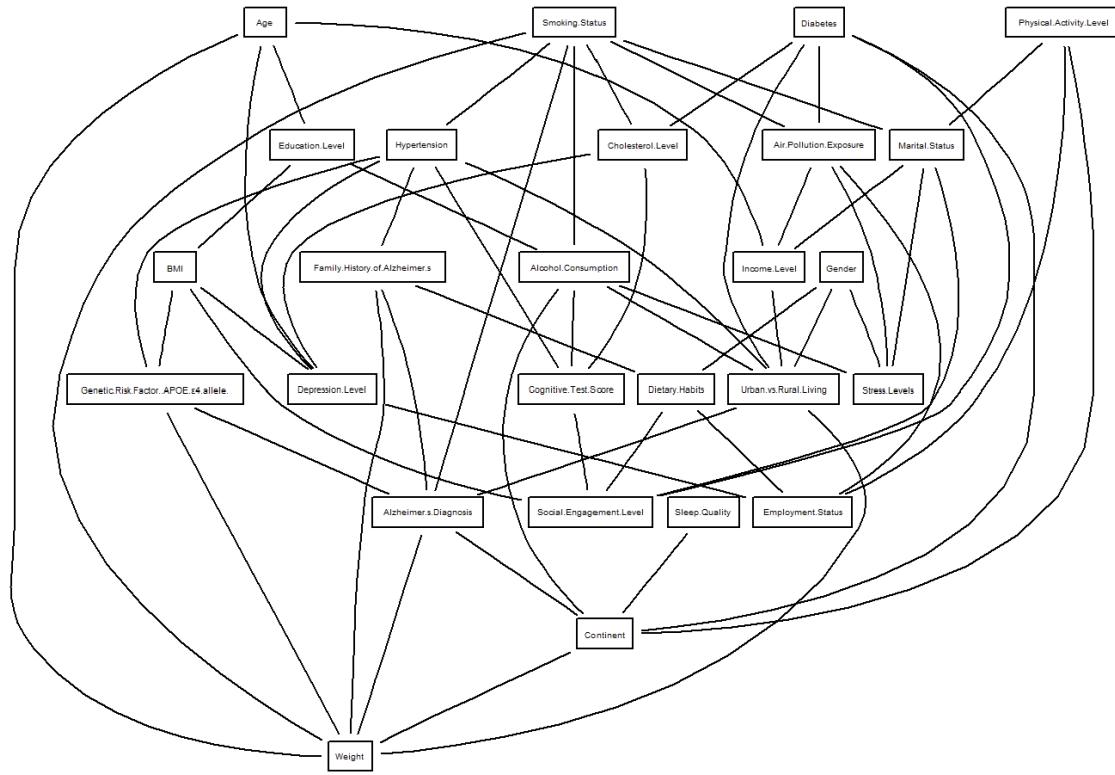


Figure 5. Local-Discovery-Based Bayesian Network

Since the two models that produced directed Bayesian Networks also generated identical structures, they consequently yielded identical AIC scores. Therefore, we will use the variables identified by these networks in our models.

Models

Four models were developed to achieve Alzheimer's Diagnosis prediction: Generalized Linear Model, Decision Tree, Neural Networks, and Random Forest. Within each model there exists a baseline model that incorporates all predictor variables, a model that uses the variables determined through forward-backward stepwise selection, and a model that uses the variables determined through the Bayesian Network.

Generalized Linear Models

The first model tested is a Generalized Linear Model (GLM). With logistic regression, the equivalent regression formula uses the logistic function to determine the probability of success [3]. Thus, the formula becomes

$$E(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}.$$

Decision Tree

Decision Trees work by splitting the data into a series of subsets based on the training data to create a classification system to predict the target variable. Each tree is then pruned based on the cross-validated error to find where the tree begins to become overfitted. The Figures below show the Decision Tree Model for each method of variable selection.

Decision Tree for Baseline

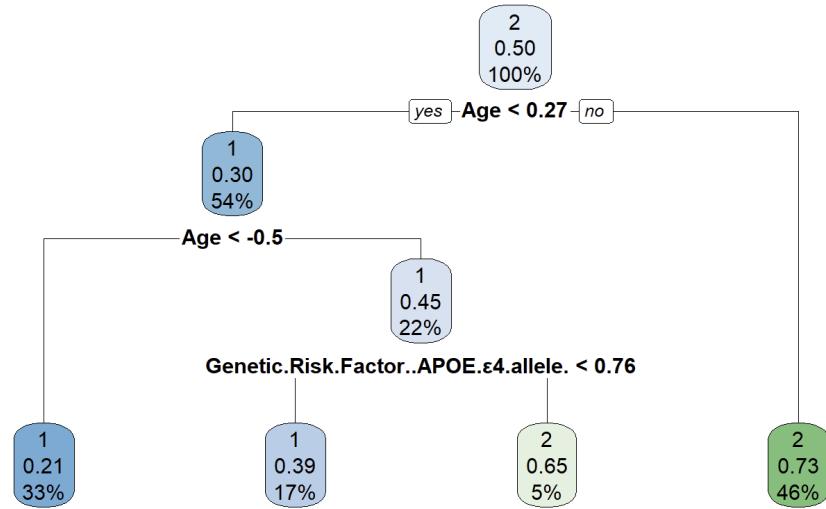


Figure 6. Decision Tree Model Using All Predictors

Decision Tree for Reduced

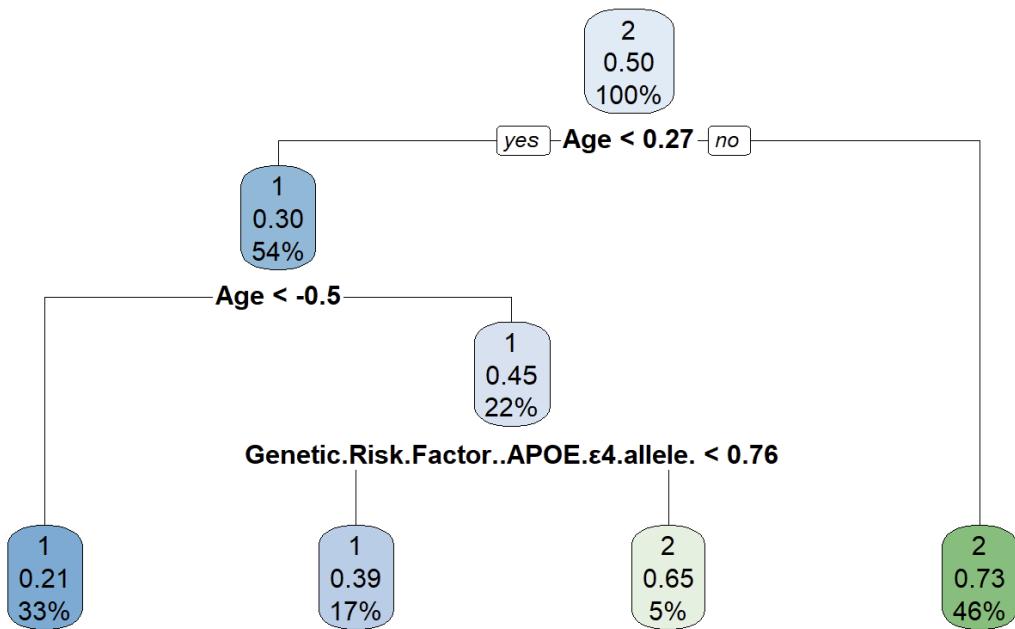


Figure 7. Decision Tree Model Using Stepwise Variables

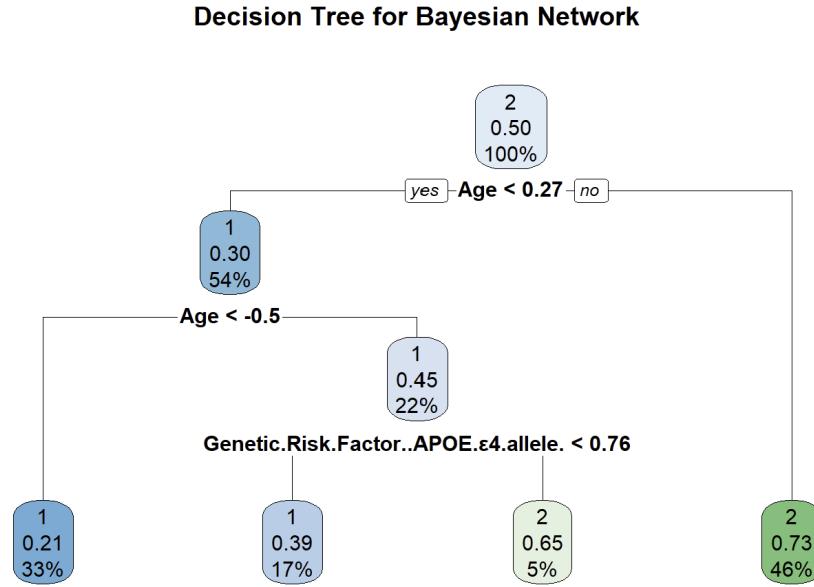


Figure 8. Decision Tree Model Using Bayesian Network Variables

Neural Network

A neural network consists of multiple layers of interconnected nodes, called "neurons," that process information. Each connection between neurons is associated with a weight, which is adjusted during the learning process to optimize the network's performance. This structure allows neural networks to identify patterns and make predictions. Through iterative training and weight adjustments, neural networks improve their accuracy over time [3]. The Figures below show the Neural Networks for each method of variable selection.

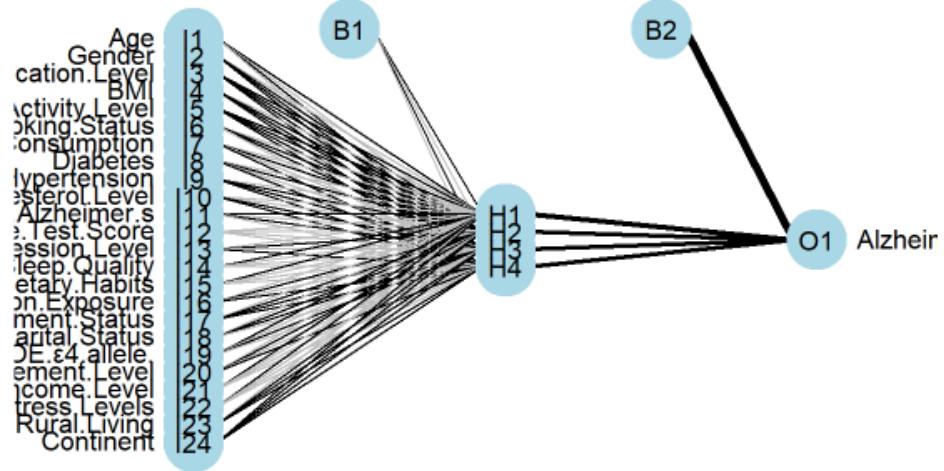


Figure 9. Neural Networks for All Predictors

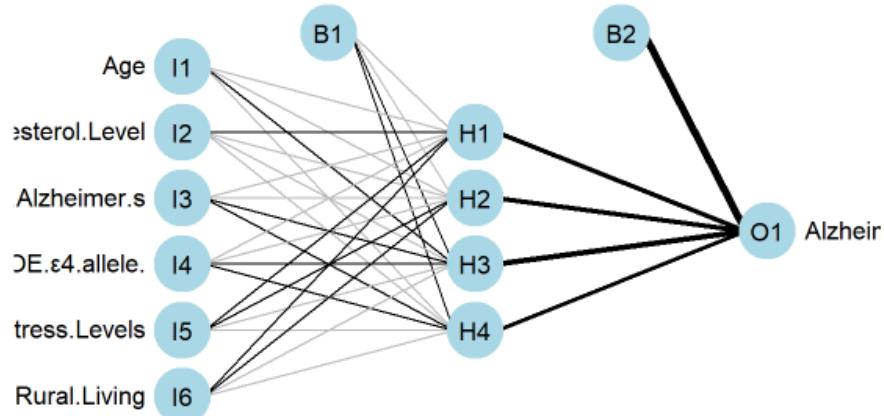


Figure 10. Neural Networks for Stepwise Variables

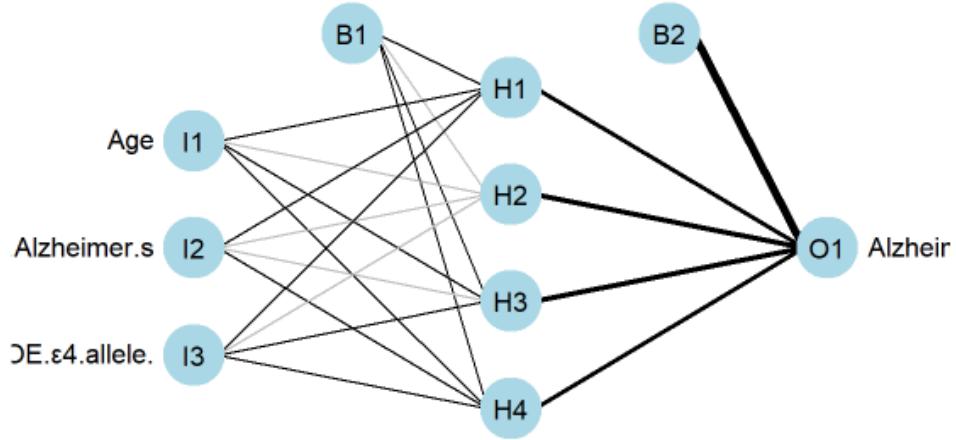


Figure 11. Neural Networks for Bayesian Network Variables

Random Forest

Random Forests are an extension of decision trees, using random subsets of the data to create multiple decision trees. The results of these trees are averaged to improve the model's accuracy and reduce overfitting.

Results

The performance of each model is evaluated using the Receiver Operating Characteristic (ROC) curve and the Area Under the Curve (AUC). The AUC score quantifies the model's ability to distinguish between classes, with values closer to 1 indicating better performance. The ROC curve visualizes the model's classification performance by plotting sensitivity (true positive rate)

against the values of 1 minus the specificity (false positive rate). Figures 12 and 13 show the ROC curve and AUC scores for the training and Test data.

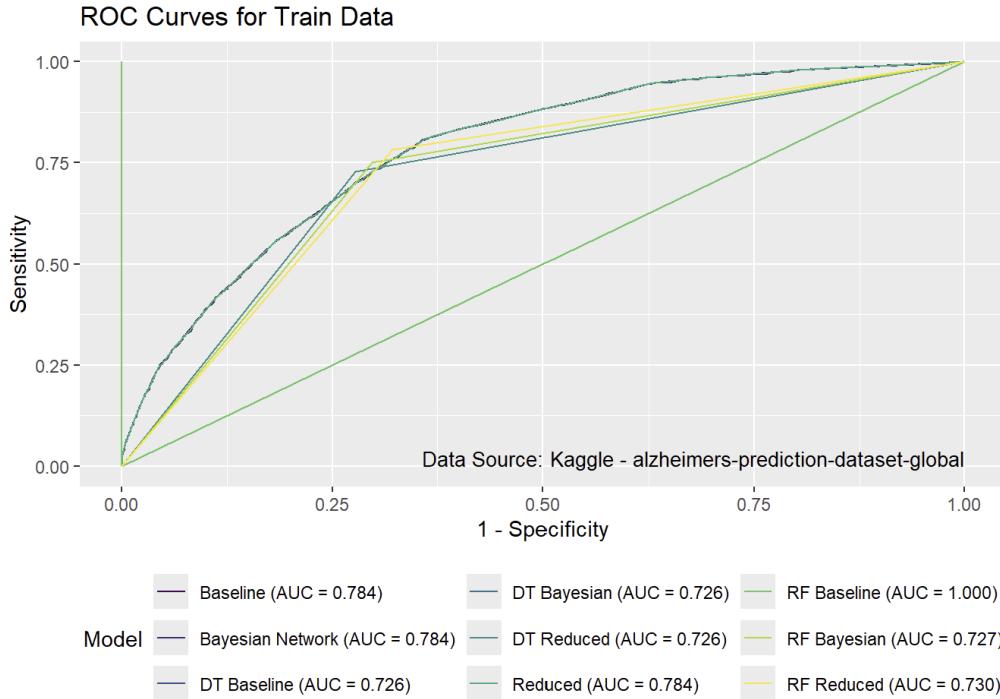


Figure 12. ROC Curve for Training Data

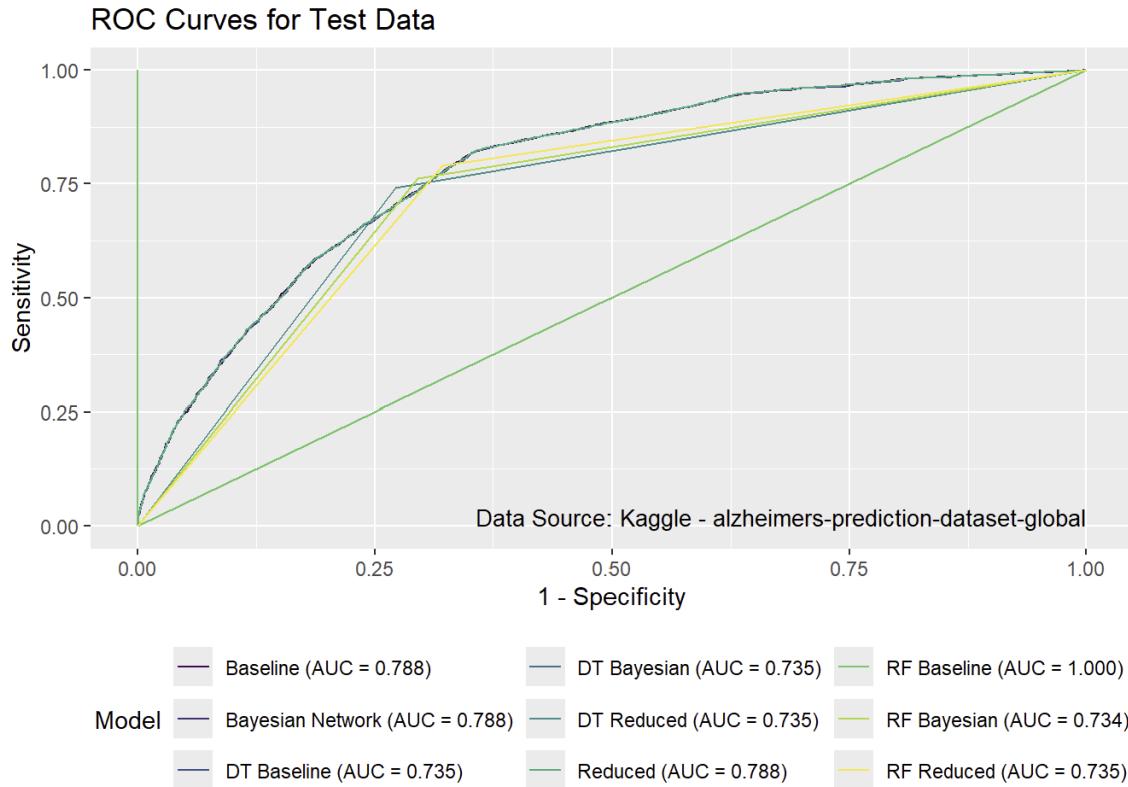


Figure 13. ROC Curves for Test Data

The goal of this project was to identify the simplest model that still delivered strong performance. Both the Baseline GLM and Bayesian Network GLM models performed well on the training and test curves. While the Random Forest Baseline model achieved a perfect AUC score, its high false positive rate made it unsuitable for consideration. The Reduced GLM model matched the Baseline models in AUC score while utilizing fewer variables. However, the Bayesian model used even fewer variables while maintaining the same AUC performance, making it the most efficient choice. By focusing only on age, family history, and APOE genotype, this model offers a streamlined yet effective approach for identifying individuals at higher risk of the disease and should be prioritized for implementation.

Works Cited

- [1] Panday, A. (2025, January 30). *Alzheimer's Prediction dataset (global)*. Kaggle.
<https://www.kaggle.com/datasets/ankushpanday1/alzheimers-prediction-dataset-global>
- [2] Bnlearn - Bayesian Network Structure Learning. (n.d.). <https://www.bnlearn.com/>
- [3] Smith, A. (2024). Logistic Regression Notes. Orlando; University of Central Florida.