Alzheimer’s disease often remains undiagnosed until symptoms appear, making early predictions crucial for identifying at-risk individuals. This project seeks to evaluate multiple logistic regression models to determine the most effective approach for predicting Alzheimer’s disease.

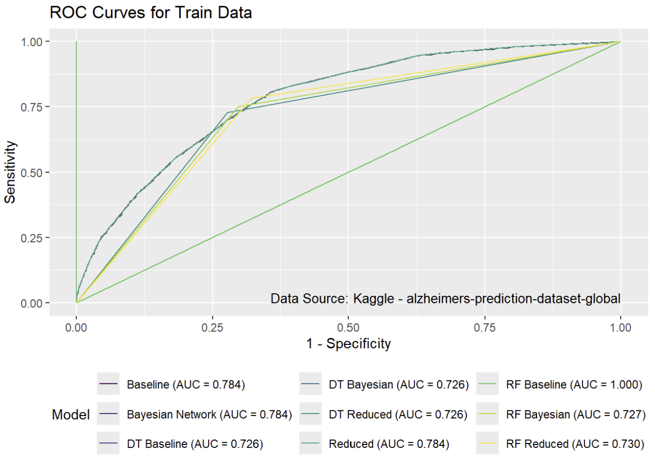
Different variable selection techniques were applied, including stepwise forward-backward selection and Bayesian Networks. Several models were tested, including GLM, Decision Trees, Neural Networks, and Random Forests, with performance assessed through Receiver Operating Characteristic (ROC) curves and Area Under the Curve (AUC) scores.

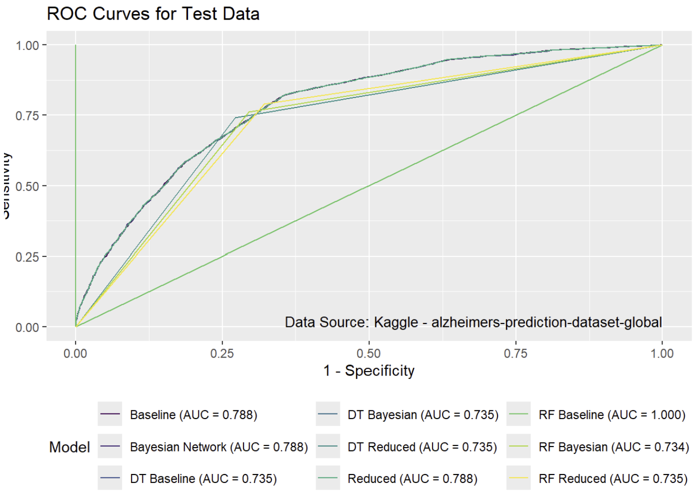
Key Recommendations

1. **Adopt the Bayesian Network GLM Model** – This model demonstrated the best balance between accuracy and simplicity, using only a few key predictors based on the results of a Bayesian Network.
2. **Use 3 Key factors to Determine Risk** – To identify if a patient is at higher risk of developing the disease, there exists key factors that can give an indication: “Age”, “Family History of Alzheimer’s”, and “Genetic Risk Factor (APOE ε4 allele)”.

Key Results

This project aimed to identify the simplest model that maintained strong predictive performance. Both the Baseline GLM and Bayesian Network GLM demonstrated high accuracy on training and test data. Although the Random Forest Baseline model achieved a perfect AUC score, its high false positive rate rendered it impractical. The Reduced GLM model matched the Baseline models in AUC performance while using fewer variables. However, the Bayesian Network model proved to be the most efficient, achieving the same AUC score with even fewer predictors. The ROC and AUC values for the training and test data are shown in Figure 1.


ROC Curves for Train Data
Sensitivity
1 - Specificity
Data Source: Kaggle - alzheimers-prediction-dataset-global
Model
Baseline (AUC = 0.784)
Bayesian Network (AUC = 0.784)
DT Baseline (AUC = 0.726)
DT Bayesian (AUC = 0.726)
DT Reduced (AUC = 0.726)
Reduced (AUC = 0.784)
RF Baseline (AUC = 1.000)
RF Bayesian (AUC = 0.727)
RF Reduced (AUC = 0.730)


ROC Curves for Test Data
Sensitivity
1 - Specificity
Data Source: Kaggle - alzheimers-prediction-dataset-global
Model
Baseline (AUC = 0.788)
Bayesian Network (AUC = 0.788)
DT Baseline (AUC = 0.735)
DT Bayesian (AUC = 0.735)
DT Reduced (AUC = 0.735)
Reduced (AUC = 0.788)
RF Baseline (AUC = 1.000)
RF Bayesian (AUC = 0.734)
RF Reduced (AUC = 0.735)