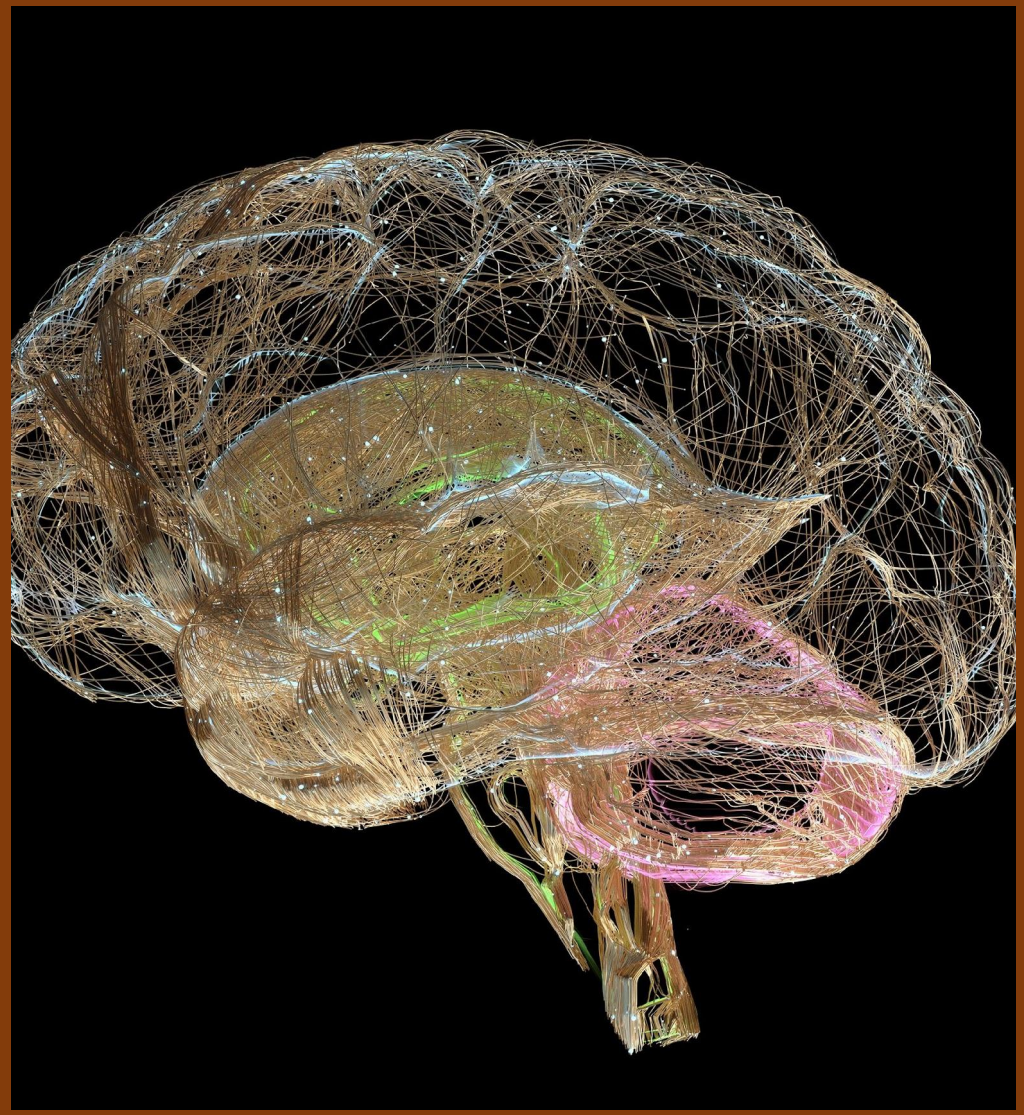
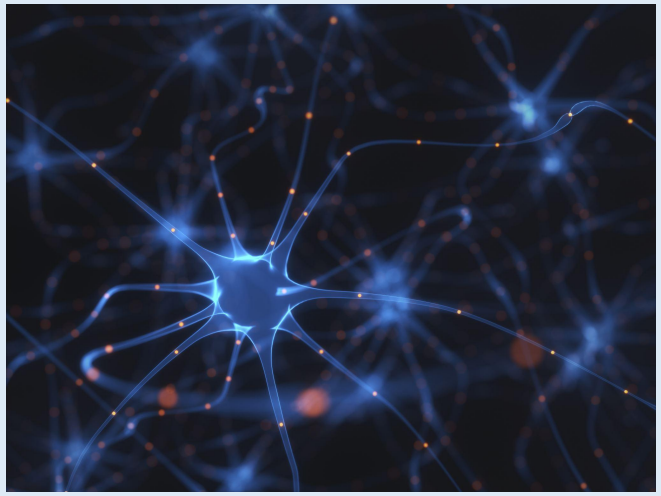
Identifying Alzheimer's Risk Factors and Prediction



Agenda

- Key Recommendations
- Data Exploration
- Variable Selection
- Models
- Results
- Summary

Key Recommendations

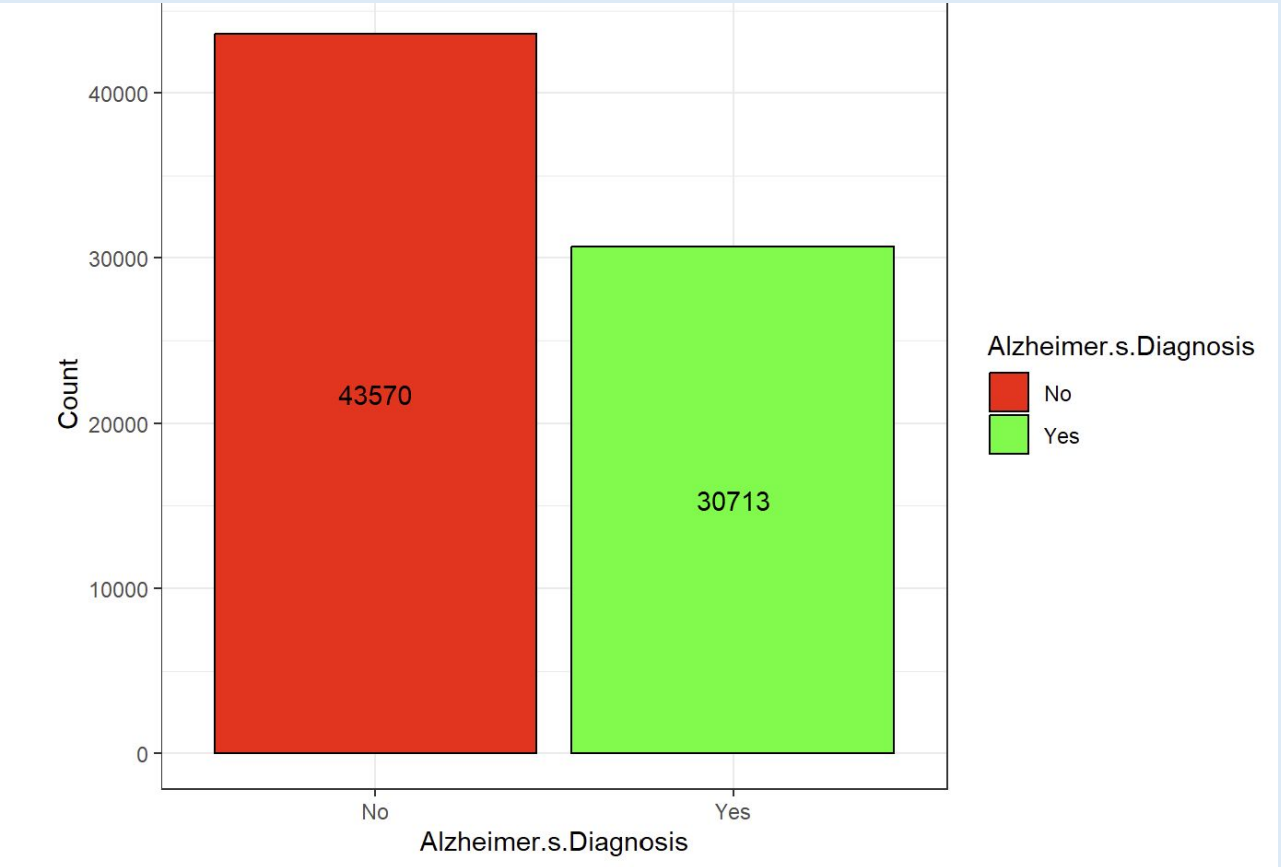


Adopt the Bayesian Network GLM Model

Use 3 Key factors to Determine Risk

Data Exploration

- **Categorization**: Countries grouped by continent to simplify the model.
- **Class Balance**: Verified equal representation of Alzheimer's and non-Alzheimer's cases.
  - Significantly more patients without Alzheimer's.
- **Class Imbalance Adjustment**: Applied weights (Alzheimer's = 1.4, Non-Alzheimer's = 1.0).
- **Data Transformation**: Converted categorical predictors to numerical values.
- **Standardization**: Centered and scaled predictor variables
  - Target variable remains unchanged.



Variable Selection

A model using no variable selection (all predictors) serves as a baseline to compare the performance of all other, reduced models.

Three categories of variable selection were used in this project:

No Variable Selection

Forward-Backwards Stepwise Variable Selection

Bayesian Network Variable Selection

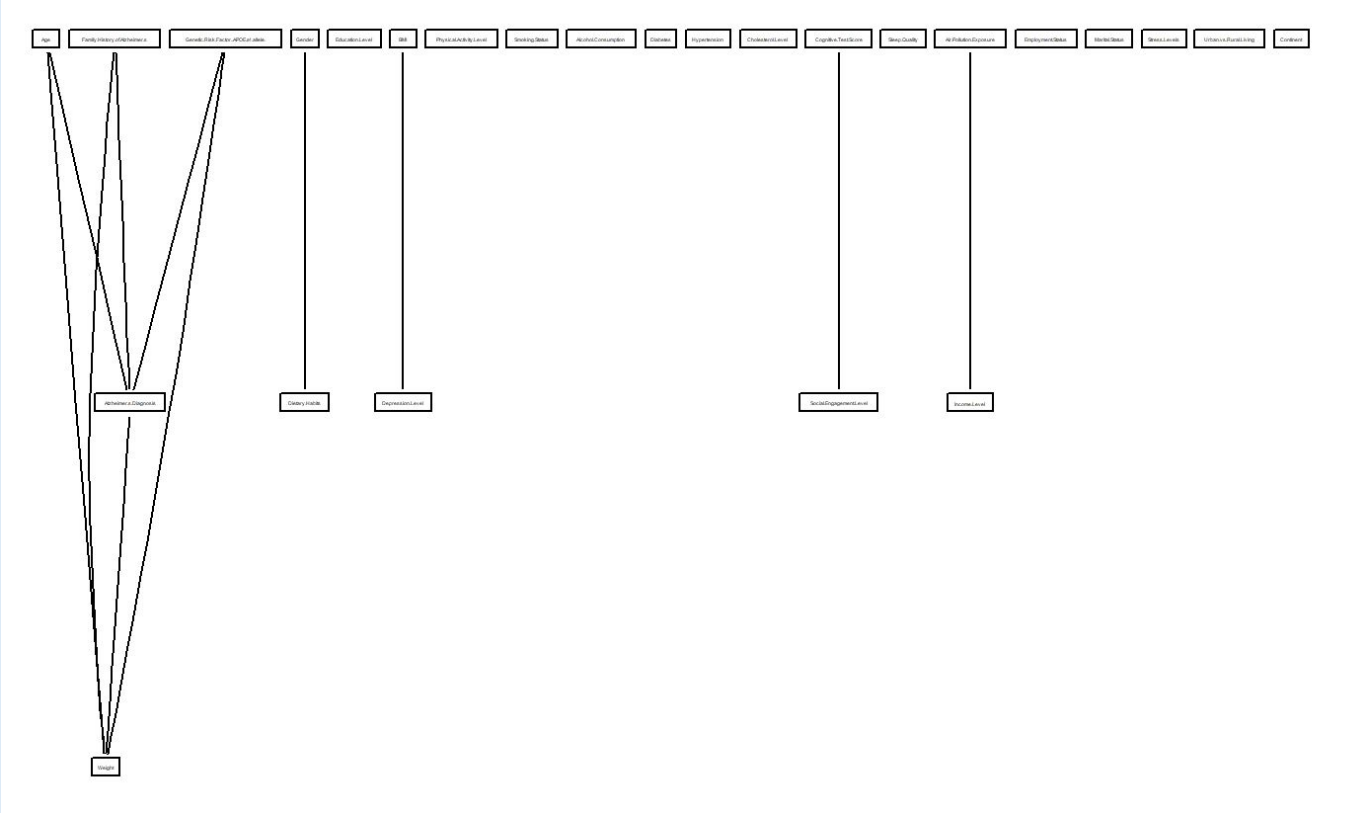
Stepwise Selection: Forward-Backward

- Stepwise selection (forward-backward) based on AIC.
  - Iteratively adds/removes variables to find the most relevant predictors.
- Reduced variables from 24 to 5:
  - Age
  - Cholesterol Level
  - Family History of Alzheimer’s
  - Genetic Risk Factor (APOE ε4 allele)
  - Stress Levels
  - Urban vs Rural Living

Stepwise Selection: Forward-Backward

- Stepwise selection (forward-backward) based on AIC.
  - Iteratively adds/removes variables to find the most relevant predictors.
- Reduced variables from 24 to 5:
  - Age
  - Cholesterol Level
  - Family History of Alzheimer’s
  - Genetic Risk Factor (APOE ε4 allele)
  - Stress Levels
  - Urban vs Rural Living

Bayesian Network: Constraint-Based



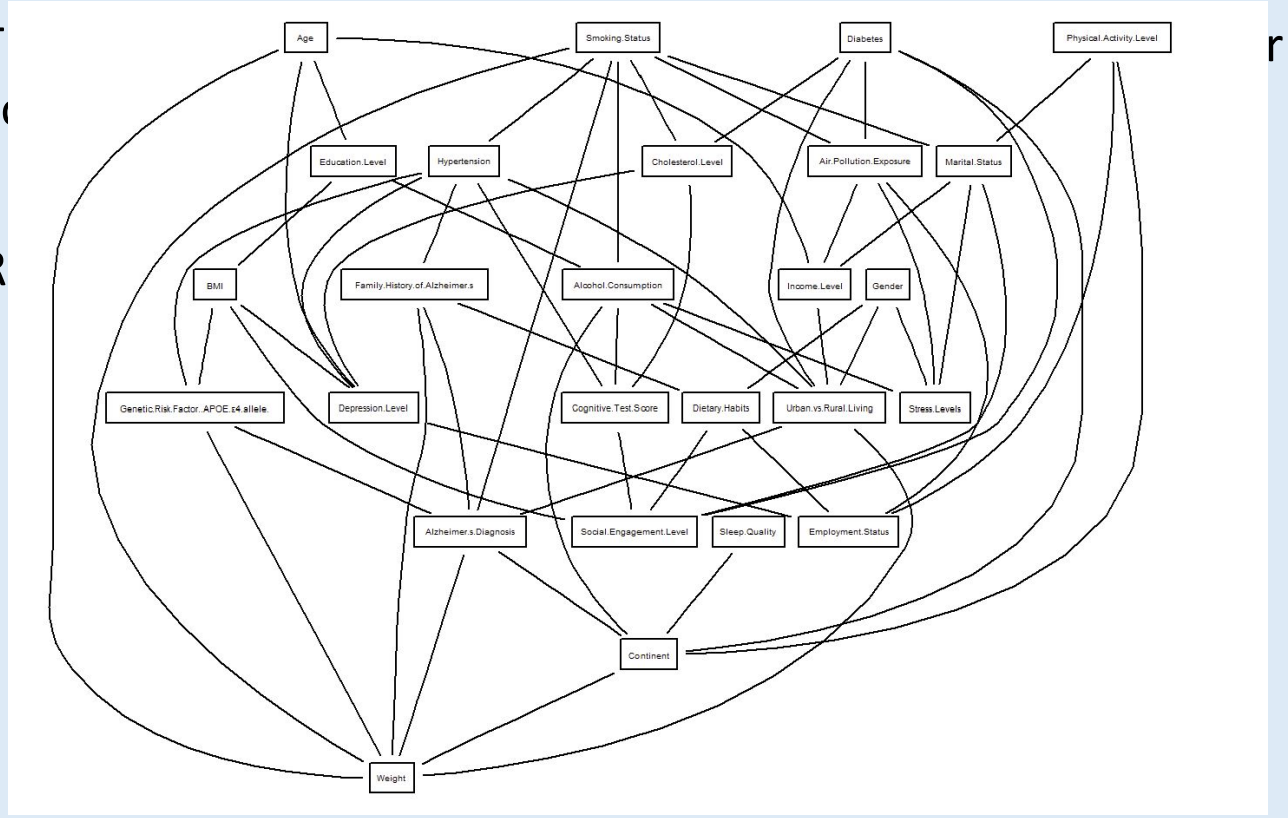
Bayesian Network: Score-Based

Bayesian Network: Hybrid

- The HC model utilizes a hybrid-based algorithm for predictor

Bayesian Network: Local Discovery

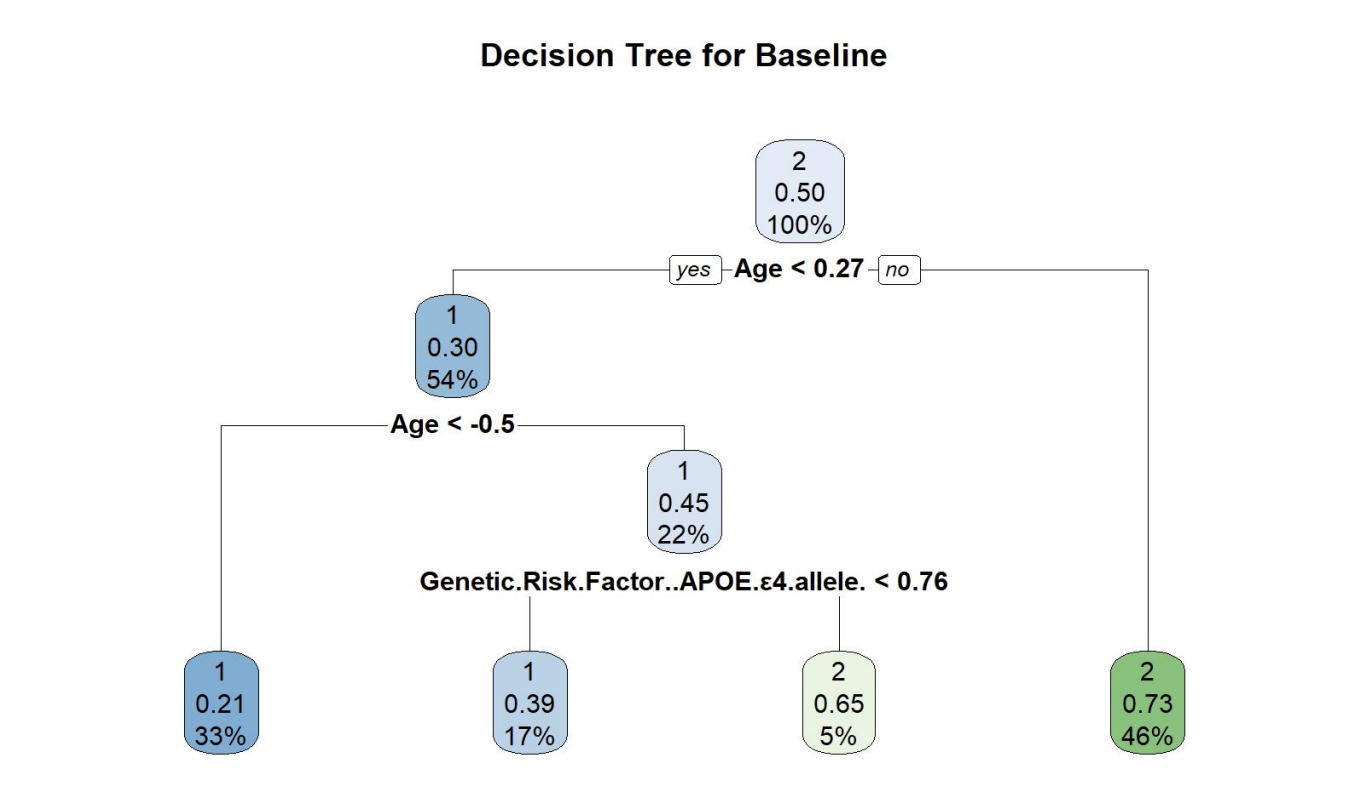
- T
  c
- R



Models: Baseline

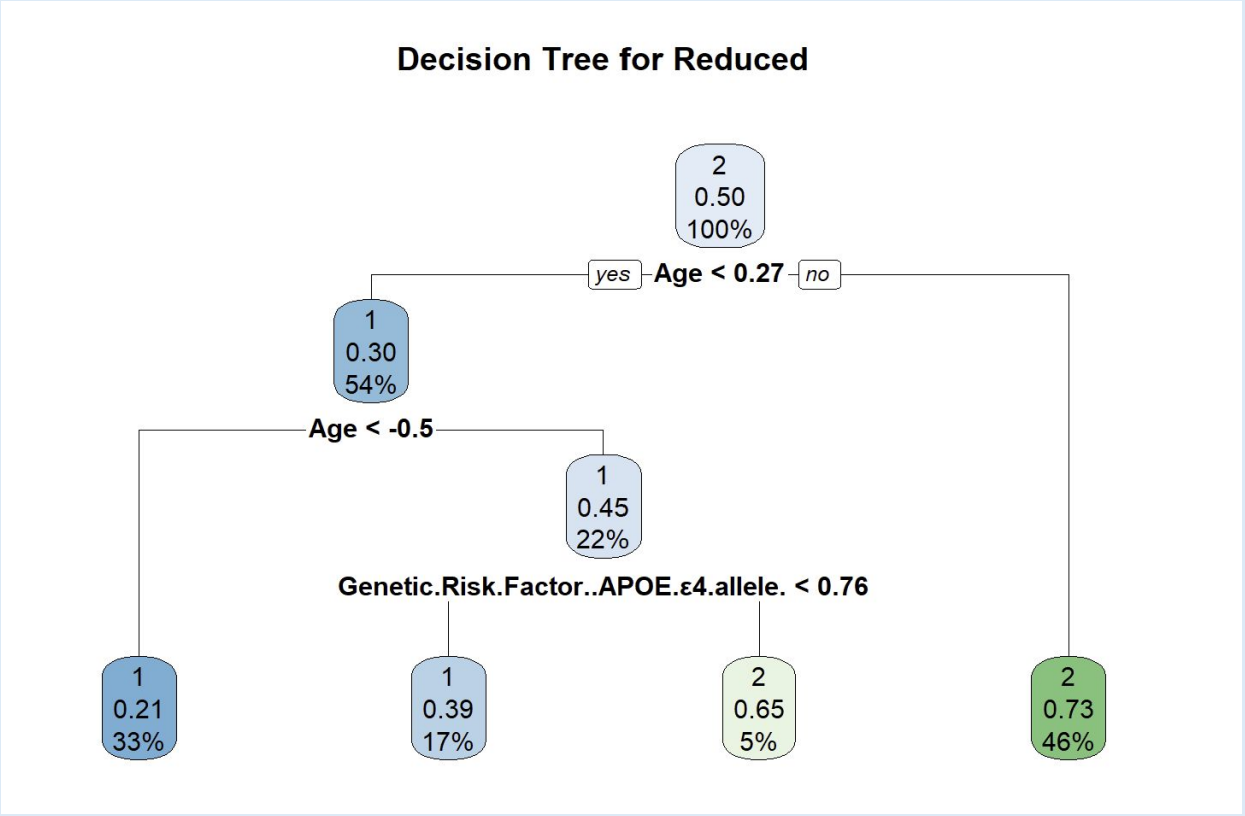
- Four models were developed for this project:
  - Generalized Linear Model (GLM)
  - Decision Tree
  - Neural Networks
  - Random Forest
- Each model includes:
  - Baseline model (all predictor variables)
  - Stepwise selection model (forward-backward selection)
  - Bayesian Network-based model
- GLM Approach:
  - Uses logistic regression to estimate probability.

Models: Decision Tree

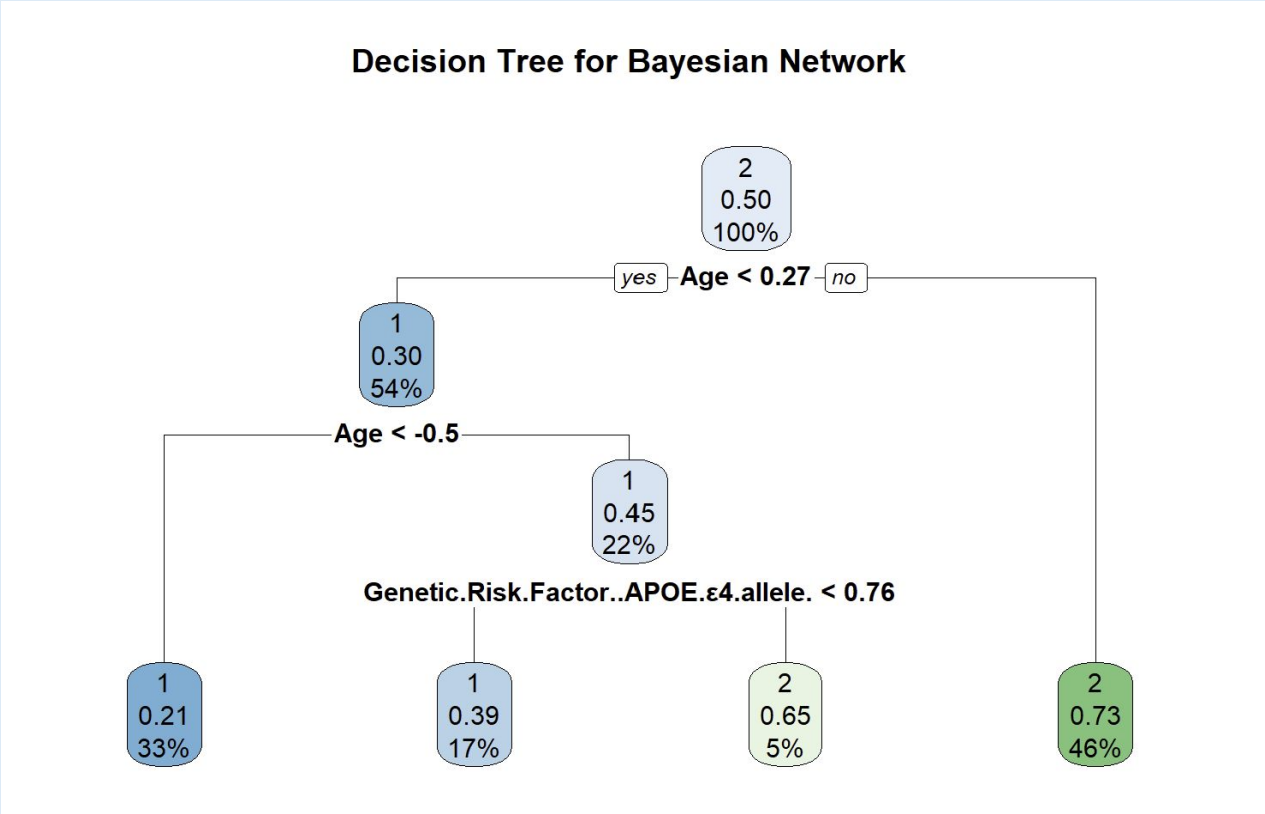


- **Decision Trees** split data into **subsets** based on training data.
  - Creates a **classification system** to predict the target variable.
  - **Pruning** is applied using **cross-validated error** to prevent overfitting.
  - The following figures show the Decision Tree Models

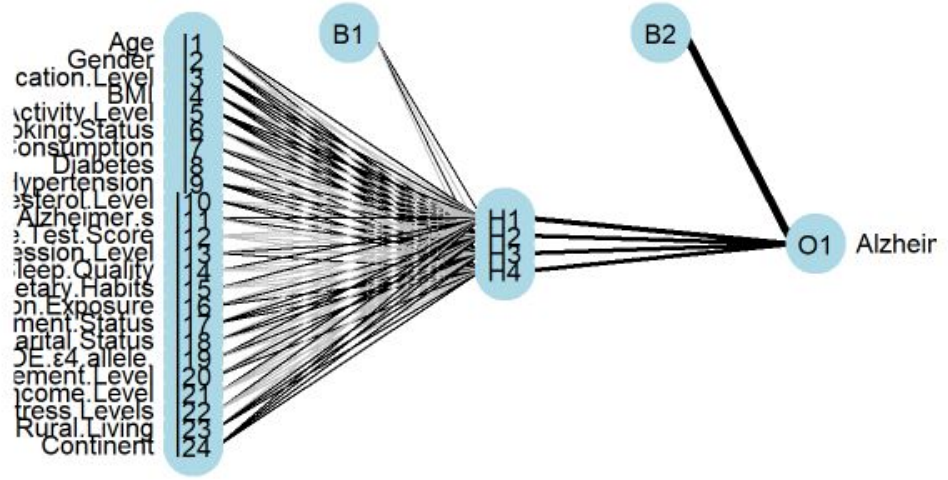
Models: Decision Tree



Models: Decision Tree

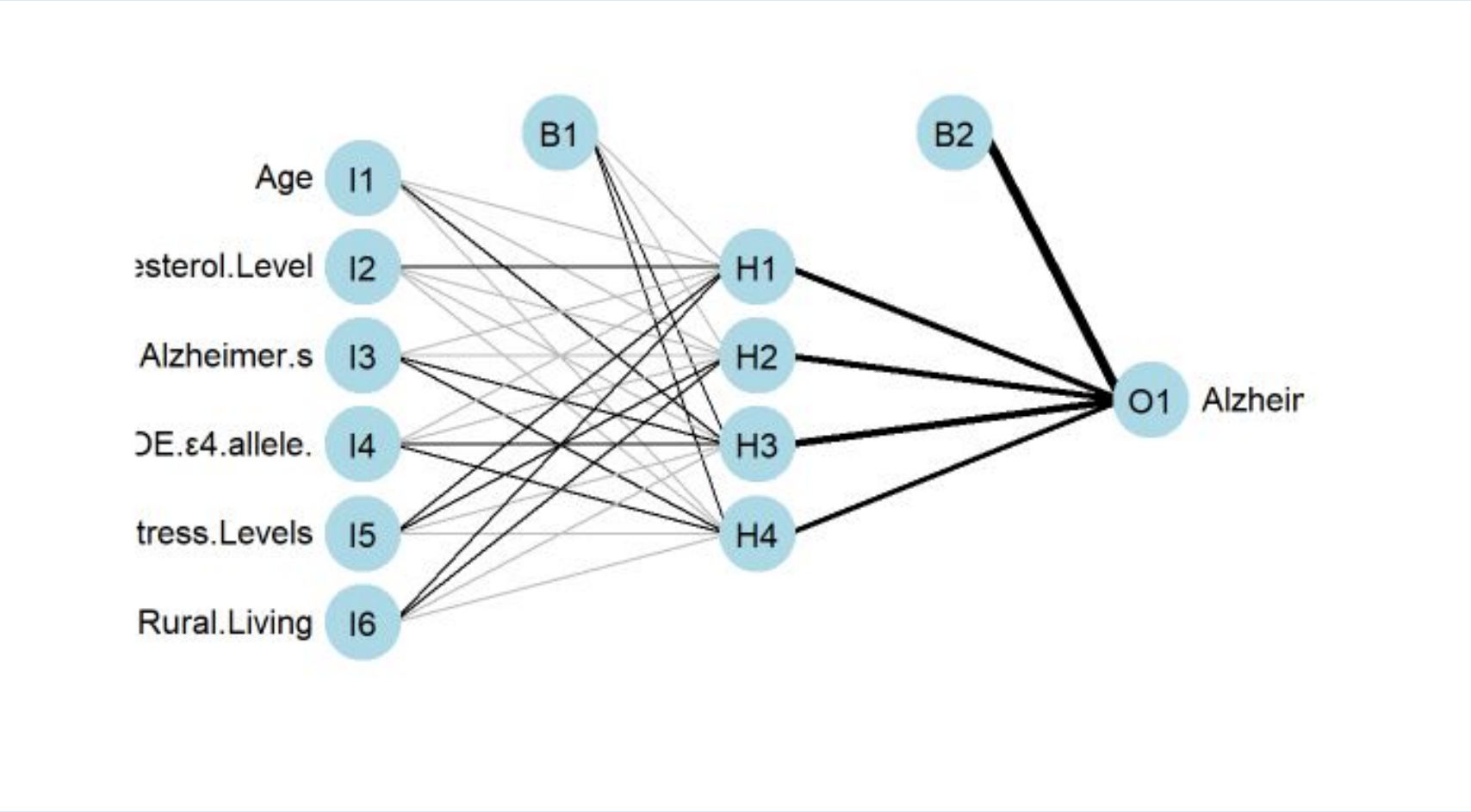


Models: Neural Networks - Baseline

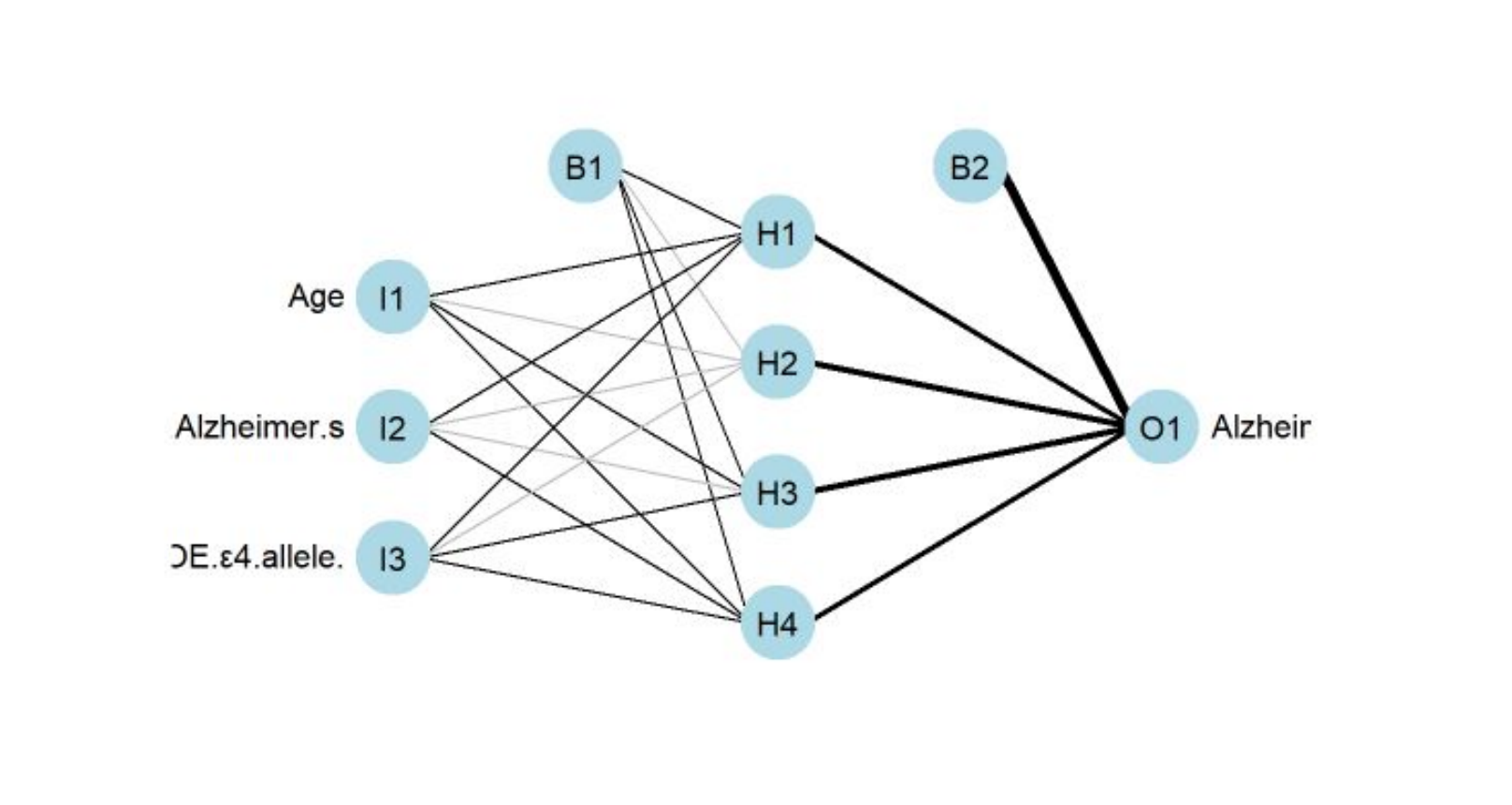


- Neural Network Structure:
  - Composed of multiple layers of interconnected “neurons”.
  - Each connection has a weight, adjusted during training.
  - Functionality:
    - Identifies patterns and makes predictions.
    - Accuracy improves over time through iterative training and weight adjustments.
  - Figures show the Neural Network models for each variable selection method.

Models: Neural Networks - Reduced



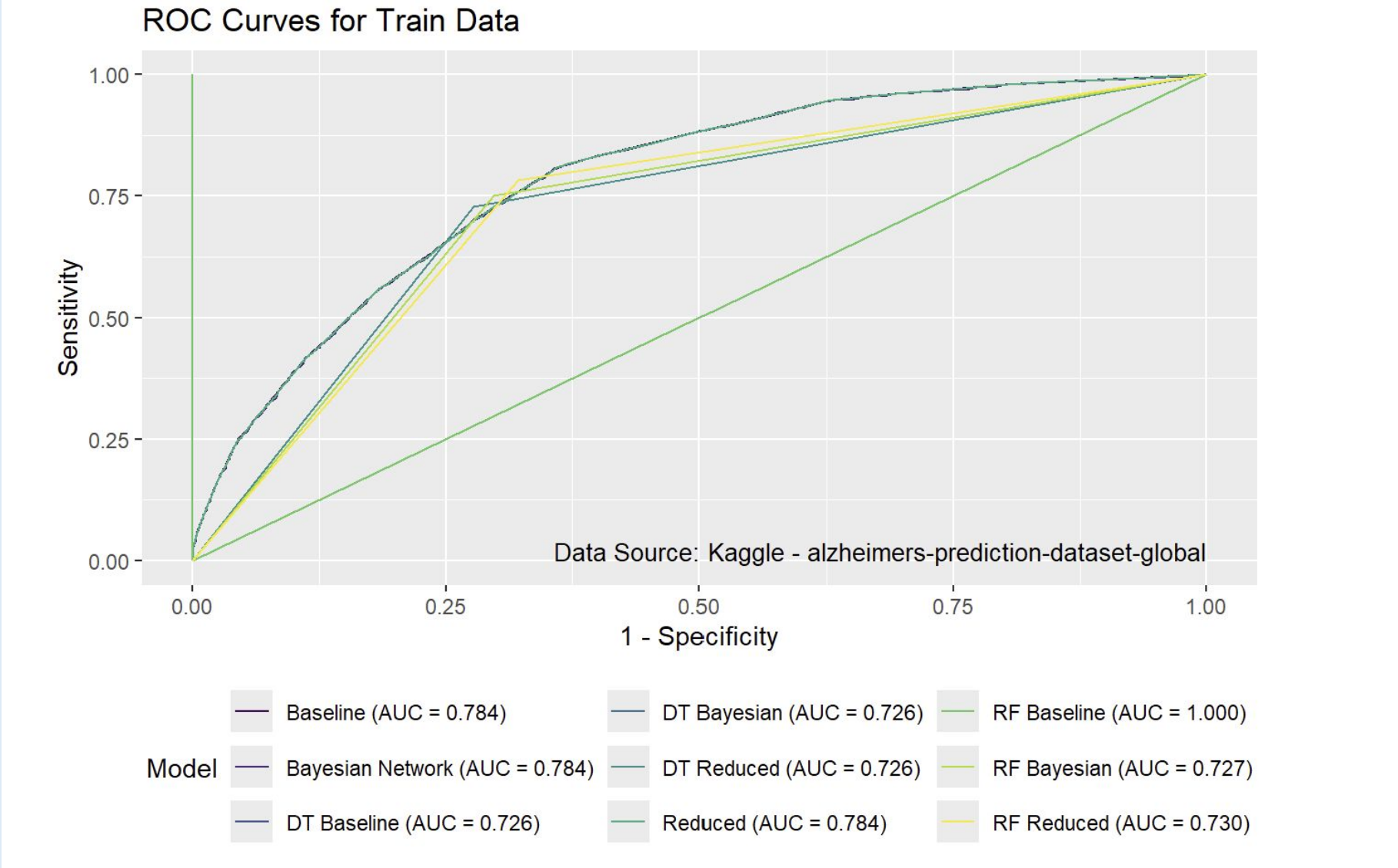
Models: Neural Networks – Bayesian Network



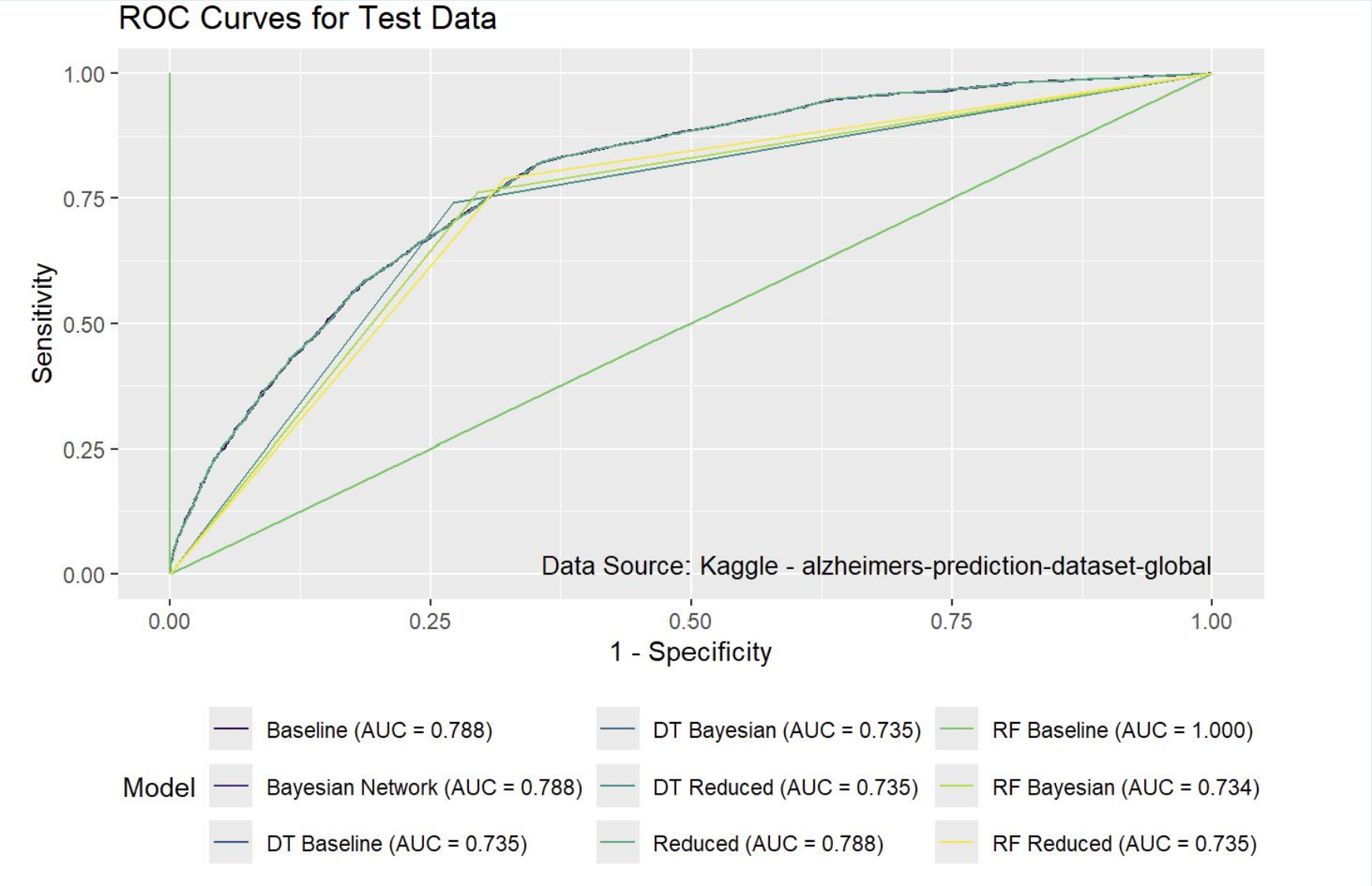
Models: Random Forest

- Extension of Decision Trees using multiple trees.
- Random subsets of data are used to build each tree.
- Averaging results improves accuracy and reduces overfitting.

Results: Training Data



Results: Test Data



Summary

- Goal: Identify the simplest model with strong performance.

- Top Performers:
  - Baseline GLM and Bayesian Network GLM performed well on both training and test curves.
  - Random Forest Model: Achieved perfect AUC, but had a high false positive rate, making it unsuitable.
  - Bayesian Network Model: Used the fewest variables (age, family history, APOE genotype) while maintaining same AUC, making it the most efficient choice 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.

Thank you

Kyle Brinker
ky601465@ucf.edu