Predicting Alzheimer’s Disease

Shriya Bang, Anthony Hessing and Nico Nap

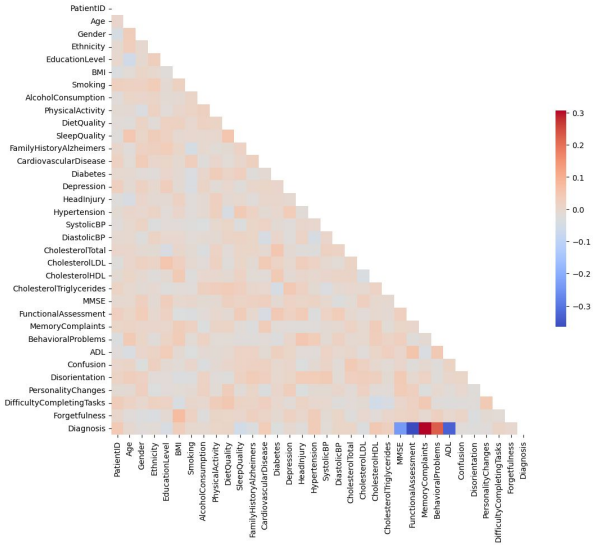
Motivation and Dataset

Motivation

- Predict someone's risk
- Earlier treatment (Gauthier, 2005)
- Preparation

Data

- Data set with 35 columns (33 usable)
- Only five columns had correlation



Correlation matrix of the dataset

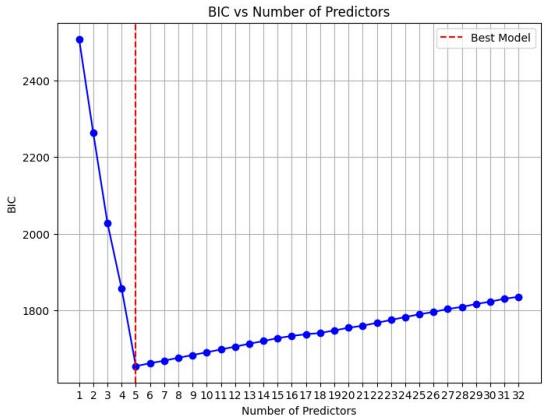
Kharoua, Rabie El (2024). Alzheimer's Disease Dataset. *Kaggle*, Kaggle, https://www.kaggle.com/dsv/8668279.

Gauthier S. G. (2005). Alzheimer's disease: the benefits of early treatment. *European journal of neurology*, *12 Suppl 3*, 11–16. https://doi.org/10.1111/j.1468-1331.2005.01322.x.

Logistic Regression

Logistic Regression - Selecting Variables

- Bayesian information criterion (BIC)
  - $BIC = k \times ln(n) - 2 \times ln(L)$
  - Used to find correlations

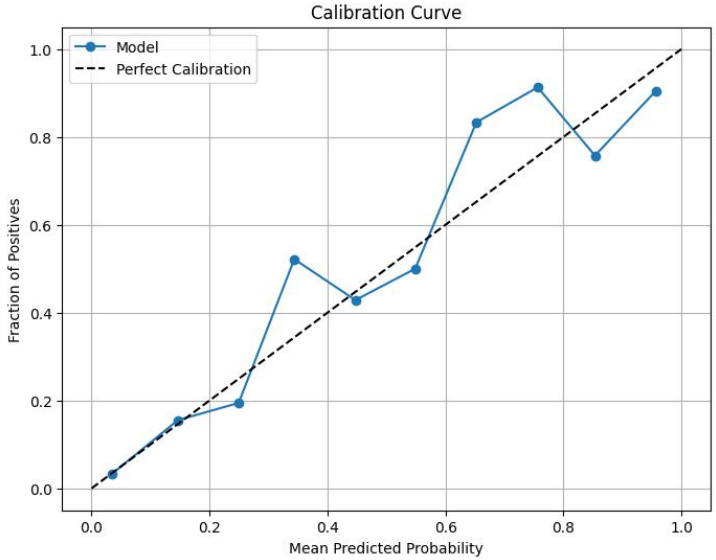


BIC with single features

-Park, S. Risk Factors of Behavioral and Psychological Symptoms in Patients with Alzheimer Disease: The Clinical Research of Dementia of South Korea Study. Korean J Fam Med. 2019;40(1):16-21. https://doi.org/10.4082/kjfm.17.0061

Logistic Regression - Training and Testing

- Use the features found with BIC
- Trained on 80%
- Cross validation on 20%
  - Hosmer-Lemeshow test
    - 10 groups
    - Chi-square statistic
    - H_0: the model is a good fit
  - Calibration curve
    - Similar to HL test
  - Precision of 79.29%



Calibration curve

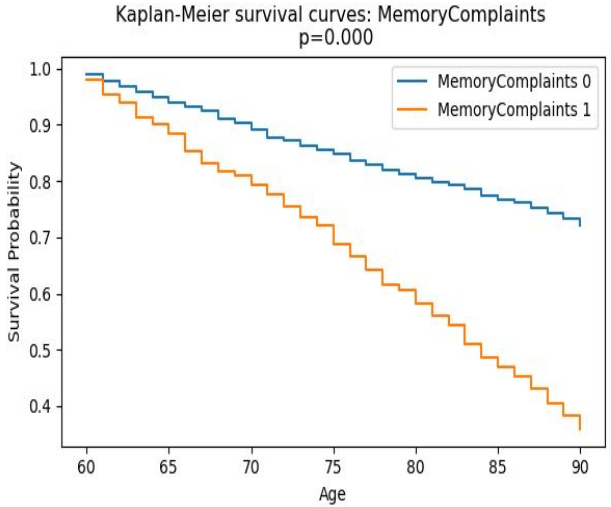
Survival Analysis

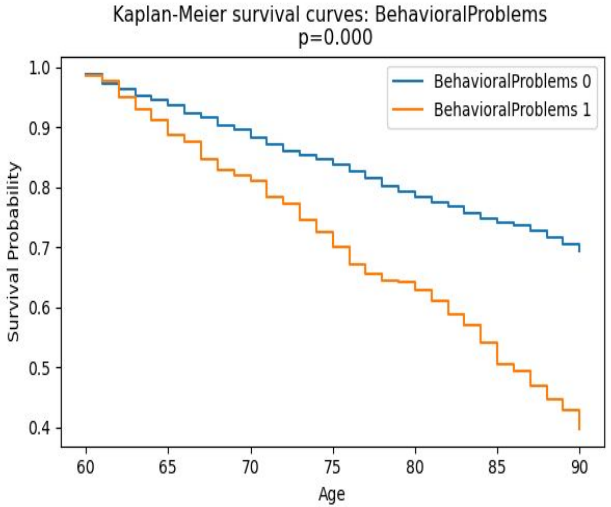
Risk Factors for Earlier Onset

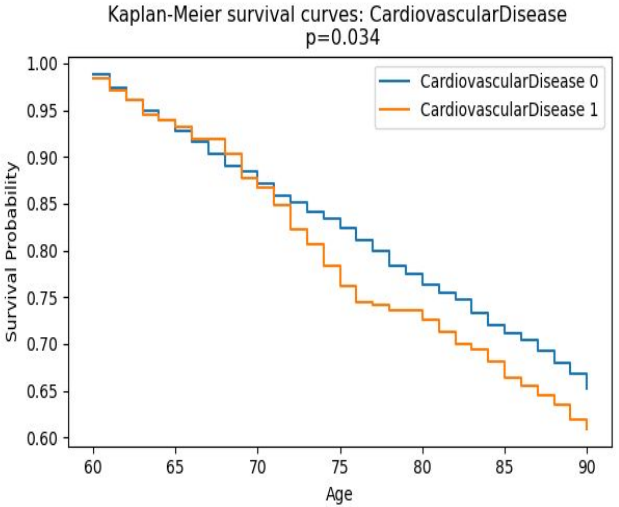
Survival Analysis - Kaplan-Meier Curves

- The survival probability at a given age is the proportion of people who have avoided Alzheimer's upto that age.

- Log rank test:
  observed vs **expected differences under the null** (i.e. if the variable did not affect onset age)



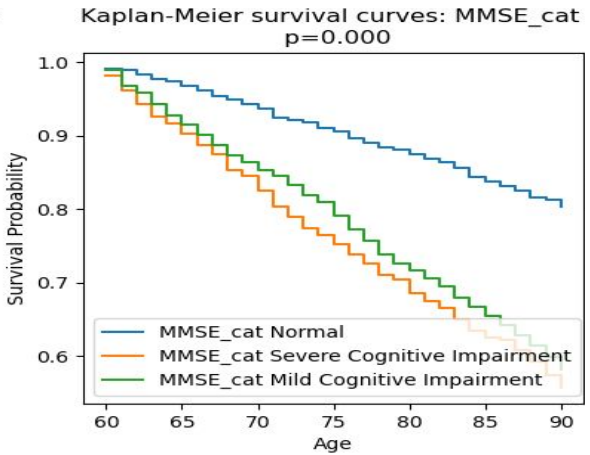


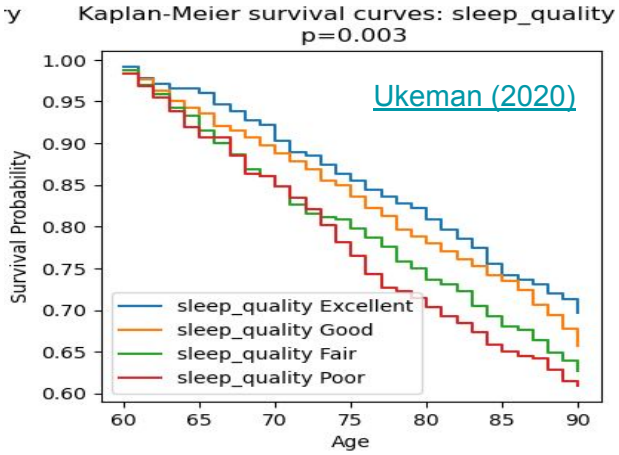


Park (2018)

Luchsinger (2004)

Incremental vs phasic risk

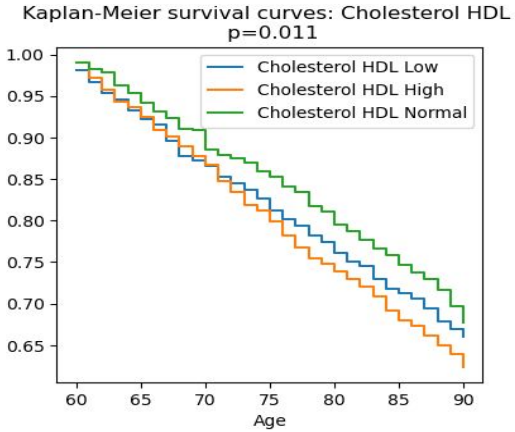




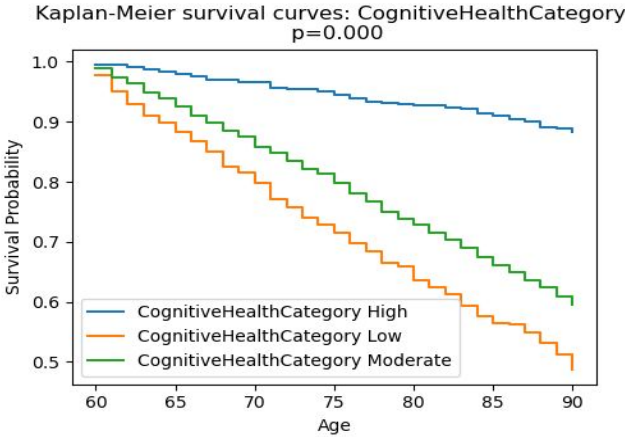
Ukeman (2020)

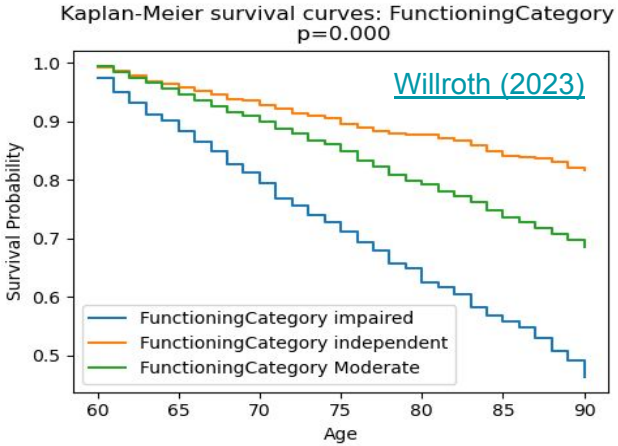
Interesting findings!

Curvilinearity



Protective factors





Willroth (2023)

Decision Tree

Decision Tree - How It Works

The Algorithm

Classification is a two-step process: the model learns on training data and predicts the response to given data.

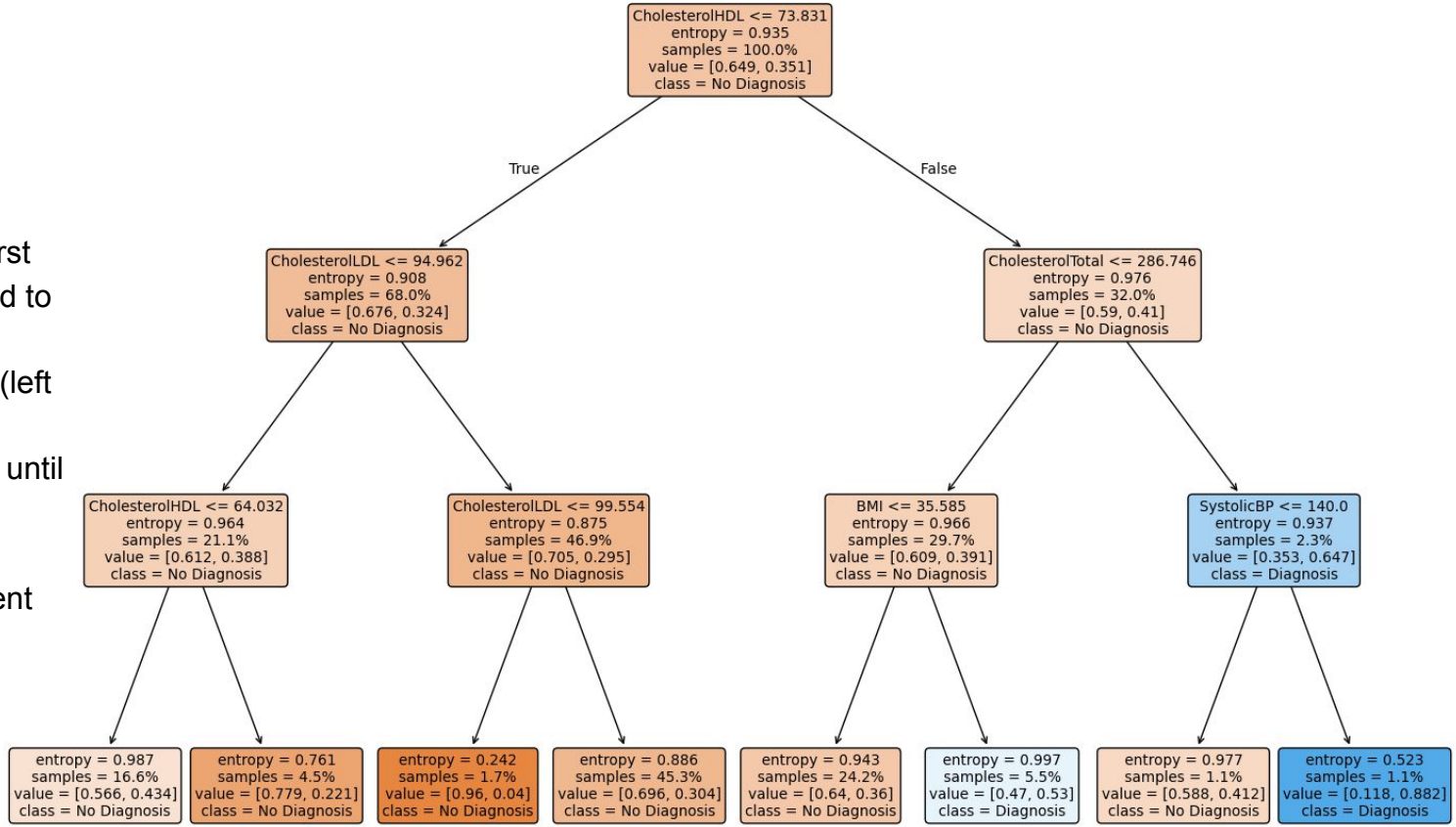
Select the best variable (the one which reduces entropy the most) and use it to split the data into smaller subsets.

Repeat recursively until the whole subset belongs to a single class, or no variables/data remain.

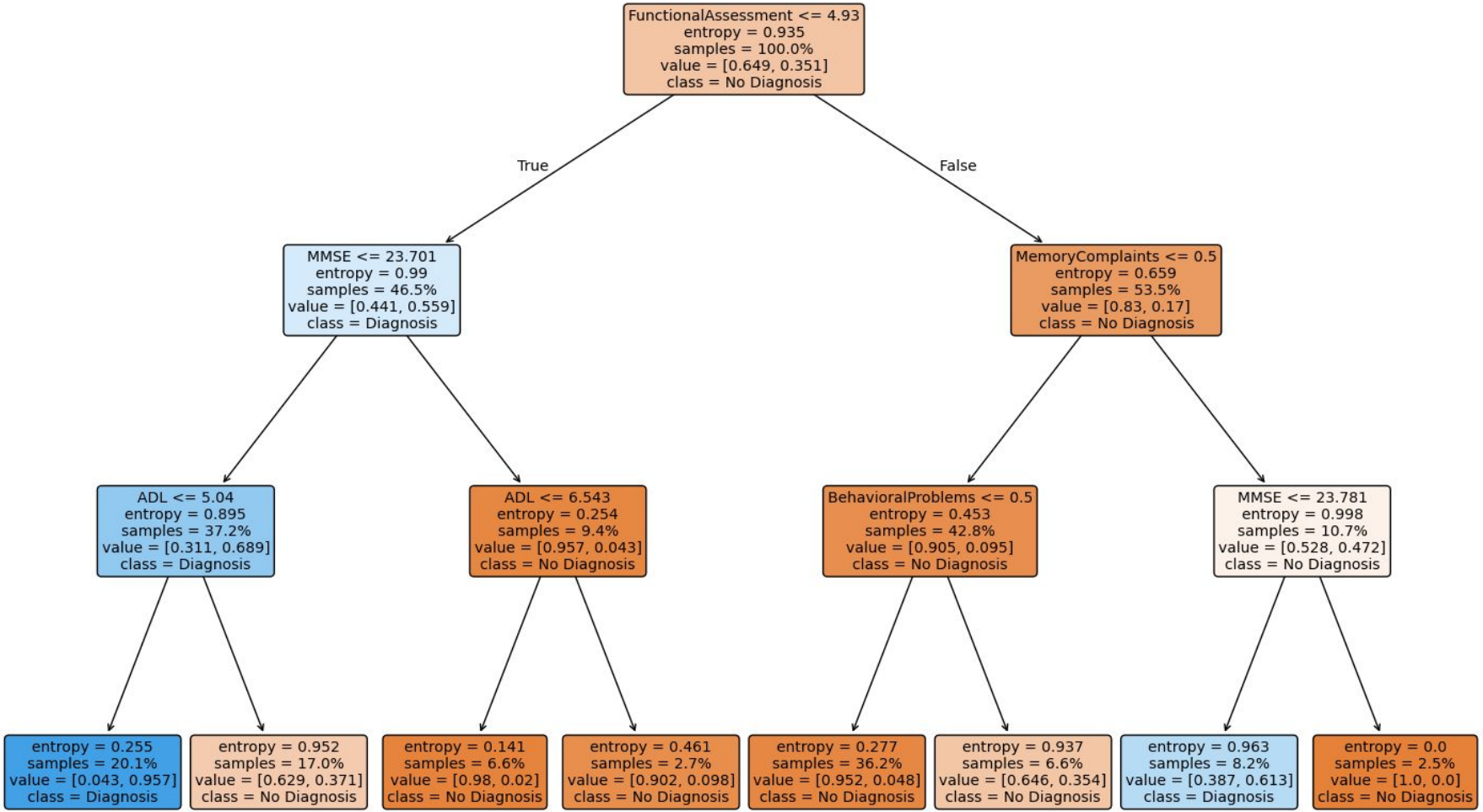
Decision Tree: Biological Variables. Model accuracy = 61%

Visual interpretation

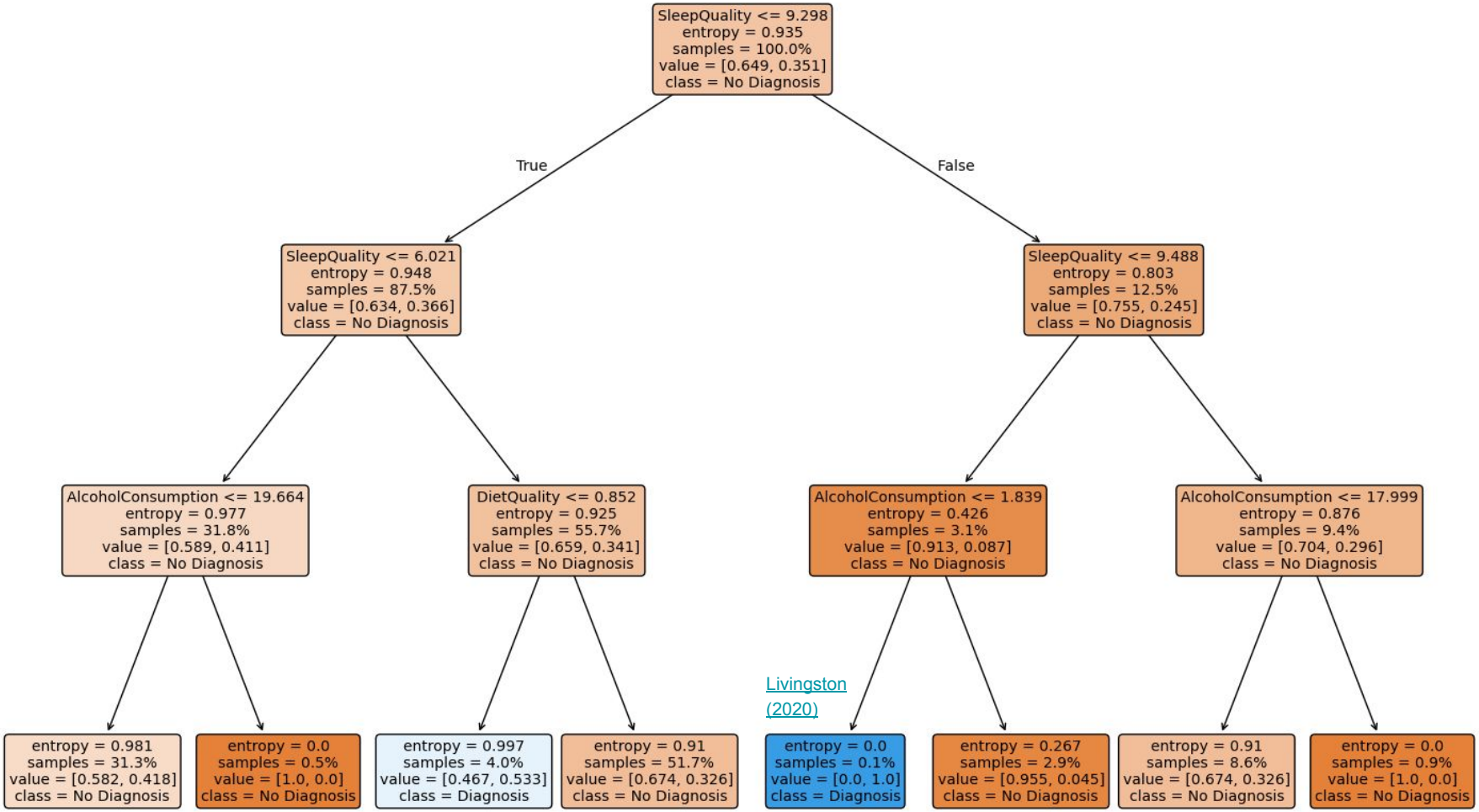
- Start at the top
  - Most important variables are first
- Compare the threshold to the patient
- Take the correct path (left or right)
- Stop at a leaf node or until the next predictor is unavailable
- Darker Colors represent low entropy.



Decision Tree: Cognitive Variables. Model accuracy = 85%



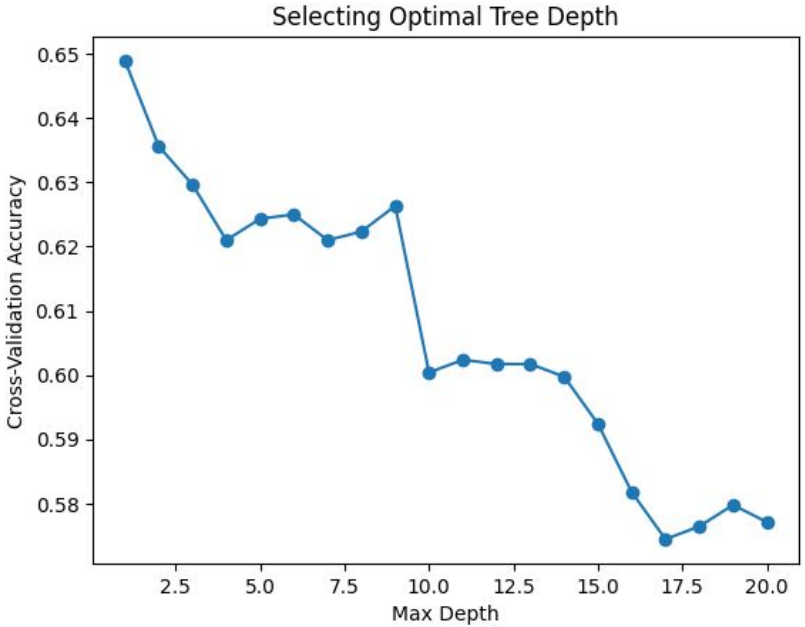
Decision Tree: Lifestyle Variables. Model accuracy = 62%



Livingston (2020)

Decision Tree - Accuracy

- Though 60% is not great accuracy, we tried different depths using 5-fold cross-validation on the training set.
- A higher depth would mean overfitting, so we stuck to `max_depth = 3`.

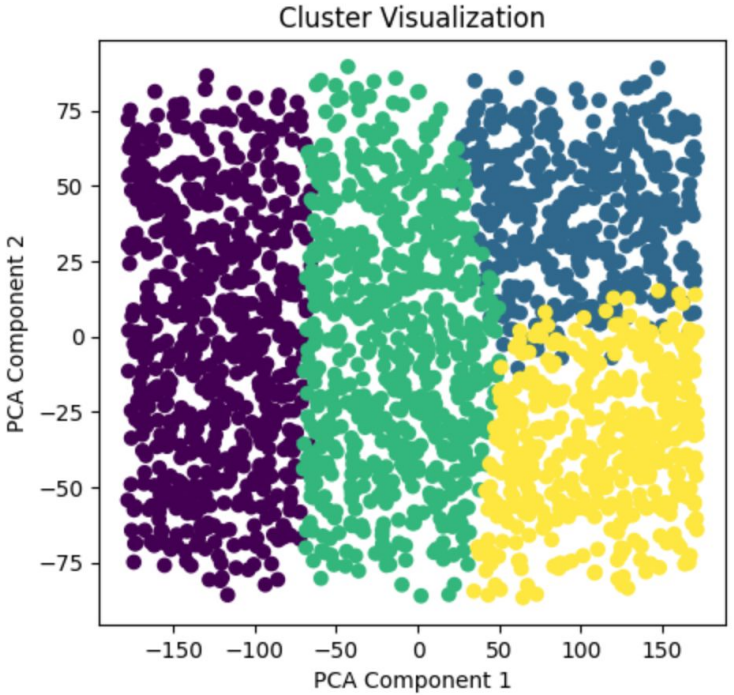


Clustering and Neural Network

(failed methods)

Clustering

- Finding correlations
- *k*-prototypes clustering
  - Clusters with 0% diagnoses
  - Clusters with 100% diagnoses
- Cholesterol variance
  - Only significant difference between clusters



One of the cluster visualizations using PCA to two dimensions. Since clustering is partially random, a different result is achieved every time.

Neural Network

- Experimented with a neural network
  - Goal was to compare this to logistic regression
  - Tested different implementations and sets
- Either everyone had Alzheimer’s or no one did
  - Not very useful…
- Lack of data (likely)

Conclusion

- It is possible to predict Alzheimer’s
- Both physical and mental variables are predictors
  - Activities of daily living
  - Sleep quality
  - Cardiovascular disease
- Regression model and decision trees

Thank you for your attention

Any questions?