Predicting Alcoholic Status using Various Classification Methods

Fall 2023 STATS 101C Group 13
Ahyoung Ju, **Ashton Chung**, **Emily Pham**,
Johan Chua, and **Nathan Lim**

Table of contents

01 Introduction

Background

- Excessive alcohol use is linked to over 200 health conditions, ranging from liver disease to mental disorders.
- The early and accurate prediction of alcoholic status becomes crucial. For earlier intervention and potentially reducing the harm caused, machine learning and data analysis are the key.

References: World Health Organization

Project Motivation

This study aims to **find the best model** that can predict an individual's risk of alcohol status **by trying several classification methods** and analyzing various factors and patterns in data.

Data Analysis 02

Summary Statistics for Numerical Variables

```
       ID              age            height          weight         waistline       sight_left     sight_right         SBP       
 Min.   :    1   Min.   :20.00   Min.   :135.0   Min.   : 30.00   Min.   : 35.00   Min.   :0.100   Min.   :0.100   Min.   : 80.0  
 1st Qu.:17501   1st Qu.:35.00   1st Qu.:155.0   1st Qu.: 55.00   1st Qu.: 74.50   1st Qu.:0.700   1st Qu.:0.700   1st Qu.:112.0  
 Median :35001   Median :50.00   Median :160.0   Median : 60.00   Median : 81.00   Median :1.000   Median :1.000   Median :120.0  
 Mean   :35001   Mean   :47.68   Mean   :162.2   Mean   : 63.23   Mean   : 81.27   Mean   :0.981   Mean   :0.977   Mean   :122.5  
 3rd Qu.:52500   3rd Qu.:60.00   3rd Qu.:170.0   3rd Qu.: 70.00   3rd Qu.: 87.60   3rd Qu.:1.200   3rd Qu.:1.200   3rd Qu.:131.0  
 Max.   :70000   Max.   :85.00   Max.   :190.0   Max.   :135.00   Max.   :999.00   Max.   :9.900   Max.   :9.900   Max.   :253.0  
                 NA's   :4877    NA's   :4941    NA's   :4972     NA's   :4940     NA's   :4877    NA's   :4900    NA's   :4919   
      DBP              BLDS          tot_chole       HDL_chole       LDL_chole      triglyceride     hemoglobin    serum_creatinine
 Min.   : 41.00   Min.   : 34.0   Min.   :  64.0   Min.   :  1.00   Min.   :   1.0   Min.   :   7   Min.   : 2.80   Min.   : 0.100  
 1st Qu.: 70.00   1st Qu.: 88.0   1st Qu.: 170.0   1st Qu.: 46.00   1st Qu.:  89.0   1st Qu.:  74   1st Qu.:13.20   1st Qu.: 0.700  
 Median : 76.00   Median : 96.0   Median : 194.0   Median : 55.00   Median : 111.0   Median : 107   Median :14.30   Median : 0.800  
 Mean   : 76.02   Mean   :100.5   Mean   : 195.7   Mean   : 56.88   Mean   : 113.3   Mean   : 132   Mean   :14.23   Mean   : 0.862  
 3rd Qu.: 81.00   3rd Qu.:105.0   3rd Qu.: 219.0   3rd Qu.: 66.00   3rd Qu.: 135.0   3rd Qu.: 159   3rd Qu.:15.40   3rd Qu.: 1.000  
 Max.   :145.00   Max.   :852.0   Max.   :2033.0   Max.   :192.00   Max.   :1933.0   Max.   :2737   Max.   :21.30   Max.   :81.000  
 NA's   :4895     NA's   :4821    NA's   :4864     NA's   :4816     NA's   :4914     NA's   :4877   NA's   :4961    NA's   :4847    
    SGOT_AST         SGOT_ALT        gamma_GTP           BMI        
 Min.   :   1.00   Min.   :   2.00   Min.   :  1.00   Min.   :13.33  
 1st Qu.:  19.00   1st Qu.:  15.00   1st Qu.: 16.00   1st Qu.:21.48  
 Median :  23.00   Median :  20.00   Median : 23.00   Median :23.88  
 Mean   :  25.96   Mean   :  25.67   Mean   : 36.76   Mean   :23.91  
 3rd Qu.:  28.00   3rd Qu.:  29.00   3rd Qu.: 39.00   3rd Qu.:25.95  
 Max.   :2670.00   Max.   :2530.00   Max.   :999.00   Max.   :42.45  
 NA's   :4887      NA's   :4893      NA's   :4961     NA's   :4967   
```

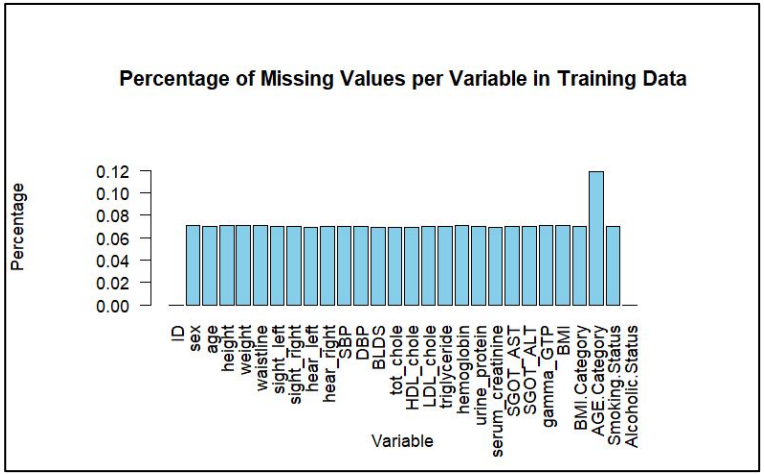
Levels for Categorical Variables

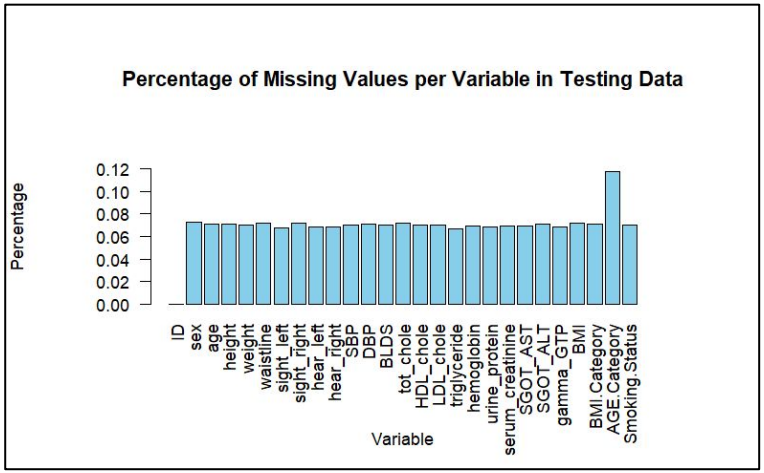
Variable Name	Level Names	Total Number of Levels
sex	“Female”, “Male”	2
hear_left	“Abnormal”, “Normal”	2
hear_right	“Abnormal”, “Normal”	2
urine_protein	1, 2, 3, 4, 5, 6	6
BMI.Category	“Healthy”, “Obese”, “Overweight”, “Underweight”	4
AGE.Category	“Mid-aged”, “Old”, “Very Old”, “Young”	4
Smoking.Status	“Never Smoked”, “Still Smoking”, “Used to Smoke”	3
Alcoholic.Status	“N”, “Y”	2

Imputing

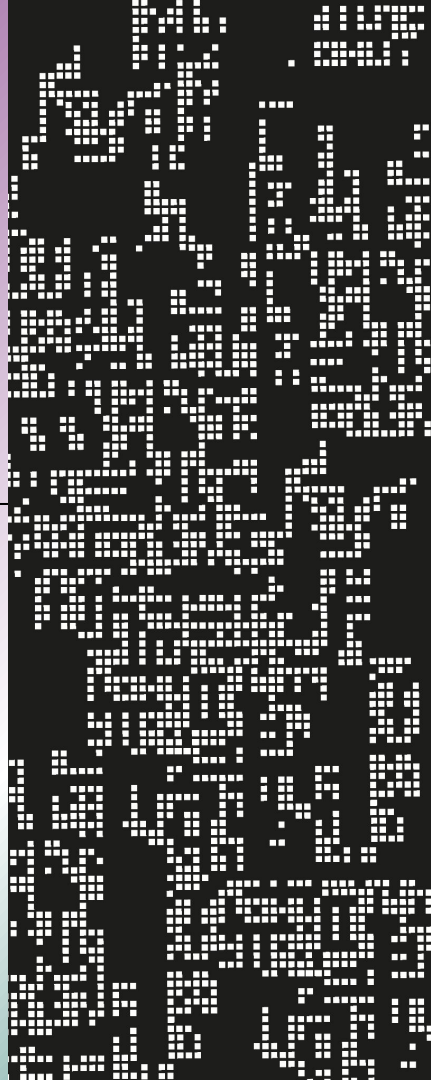
- ~8% of values were missing across training and testing data
- Data deletion unpractical
  - Deprives us of training data
  - Affects nearly all variables

- Used HMISC and MICE to imputing missing values
- Found that HMISC outperforms MICE





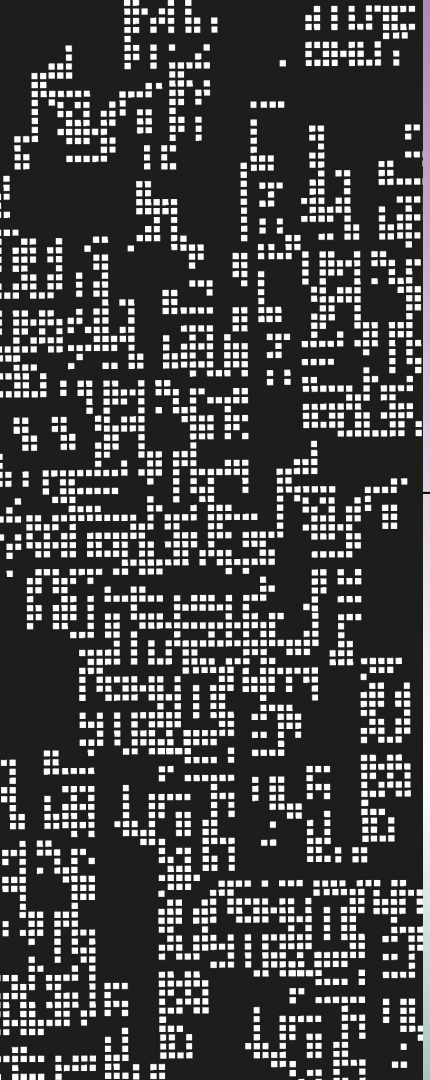
03 Variable Selection



Optimal Subset of Variables

- Some variables overlapped / had high collinearity (eg. BMI numerical & BMI categorical)
- Including overlapping variables would severely hinder our model performance

- Used varImp() function from the Caret library—which calculates variable importance for regression and classification models
- **FINAL SUBSET:** sex, age, height, waistline, SBP, BLDS, tot_chole, HDL_chole, LDL_chole, triglyceride, hemoglobin, SGOT_AST, SGOT_ALT, gamma_GTP, BMI.Category, AGE.Category, Smoking.Status



Methods and Models

04

K-Nearest Neighbors (KNN)

- Finds K nearest neighbors to training points to make predictions.
- Used class library's knn() function
- Scaled non-categorical variables
- Best tested k-value was 100, could have led to over-generalization
- Highest accuracy
    - On our test data: 0.7022143
    - On Kaggle test data: 0.71083

	N	Y
N	23945	9720
Y	11125	25210

Logistic Regression

- Predicts to model a binary outcome
- Used caret library's glm() function
- Efficient as number of predictors are outnumbered by observations
- Highest accuracy
    - On our test data: 0.7252143
    - On Kaggle test data: 0.7211

	N	Y
N	25610	9460
Y	9775	25155

Support Vector Machine (SVM)

- Machine learning algorithm that determines hyperplane to distinguish between classes
- Used e1071 library's svm() function
- Used linear kernel for computational efficiency
- Optimal gamma value was 0.8
- Highest accuracy
    - On our test data: 0.7202857
    - On Kaggle test data: 0.71813

	N	Y
N	25410	9775
Y	9805	25010

Linear Discriminant Analysis (LDA)

- k-dimensional projection that creates the greatest between-group separation based on Y
- Use lda() function, built into the MASS library
- Accuracy rate: 0.7220286

	N	Y
N	25518	9863
Y	9595	25024

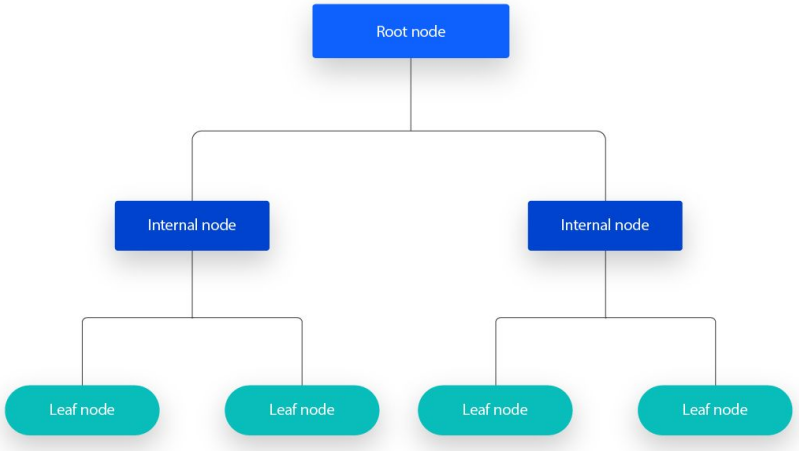
Quadratic Discriminant Analysis (QDA)

- Discriminant Analysis with multiclass classification.
- Assumes different classes have difference covariances
- Use qda() function, built into the MASS library
- Accuracy rate: 0.7024234803

	N	Y
N	25904	11626
Y	9199	23253

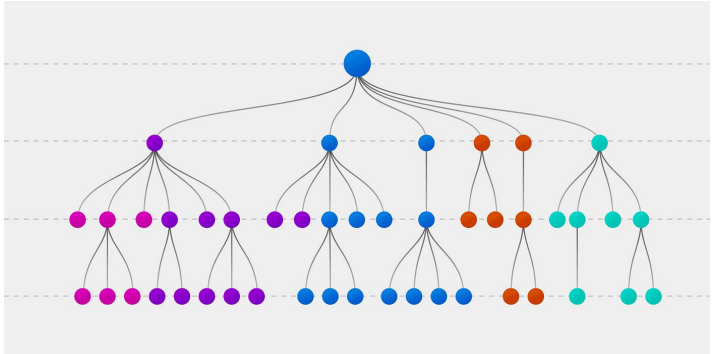
Recursive Partitioning & Decision Trees

- Decision Trees utilizes a hierarchical structure using nodes, branches, and leaves to assist with decision making
- Utilized recursive partitioning, a method associated with decision trees
    - Decision is made about which feature to split on and what threshold to use at the end of each node.
- Use rpart() function, built into the rpart library
- Accuracy rate: 0.4999429
    - Relatively poor performance



Boosting (XGBoost)

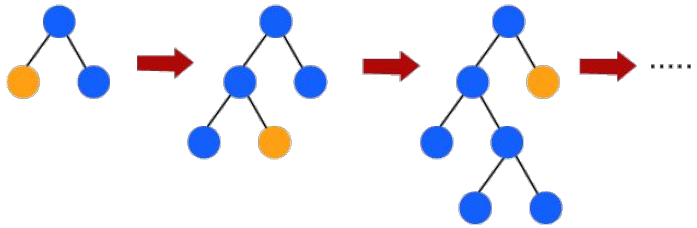
- Boosted classification tree model
- Used train() function, built into the XGBoost library
- Model Specification: xgbTree method with with 10-fold CV
- Accuracy rate: 0.7262857



Gradient Boosting Machines (GBM)

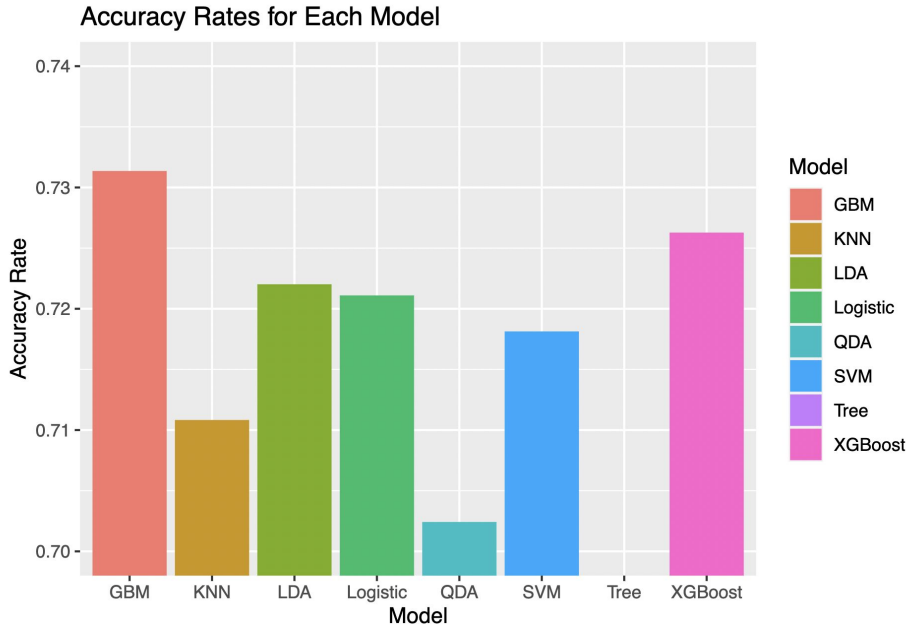
BEST MODEL

- Identifies weak learners / decision trees using gradients in the loss function
- Model Specification: gaussian SSE loss function, 5000 trees, with 10-fold CV
- Accuracy Rate: 0.7313623



05 Conclusion & Further Work

Summary of Models



Rank	Model	Accuracy.Rate
1	GBM	0.73136
2	XGBoost	0.72628
3	LDA	0.72202
4	Logistic	0.72110
5	SVM	0.71813
6	KNN	0.71083
7	QDA	0.70242
8	Tree	0.49994

Limitations & Further Work

- Find better performing k for KNN
- Perform multiple rounds of imputations with both methods
  - potentially combine imputations
- Logistic regression may have been influenced by non-linear or complex relationships in the data
- LDA and QDA are also situational on data's true relationships as well
- Optimize Tree Method because it performed extremely poorly (0.49 accuracy)

References

- IBM. "What is a Decision Tree." *IBM*, www.ibm.com/topics/decision-trees.
- Mitchell, J & Collen, Jacob & Petteys, S & Holley, Aaron. (2011). A simple reminder system improves venous thromboembolism prophylaxis rates and reduces thrombotic events for hospitalized patients. Journal of thrombosis and haemostasis : JTH. 10. 236-43. 10.1111/j.1538-7836.2011.04599.x.
- Natekin, A., & Knoll, A. (2013, October 21). Gradient Boosting Machines, a tutorial. Frontiers. https://www.frontiersin.org/articles/10.3389/fnbot.2013.00021/full
- Stenetorp, P & Pyysalo, Sampo & Topic, Goran & Ohta, Tomoko & Ananiadou, Sophia & Tsujii, Jun'ichi. (2012). brat: a Web-based Tool for NLP-Assisted Text Annotation.The 3th Conference of the European Chapter of the Association for Computational Linguistics; Avignon, France. 102-107.
- World Health Organization. (n.d.). *Harmful use of alcohol*. World Health Organization. https://www.who.int/health-topics/alcohol#tab=tab_1
- Zach. "Introduction to Quadratic Discriminant Analysis." *Statology*, 2 Nov. 2020, www.statology.org/quadratic-discriminant-analysis/.

Thank You!