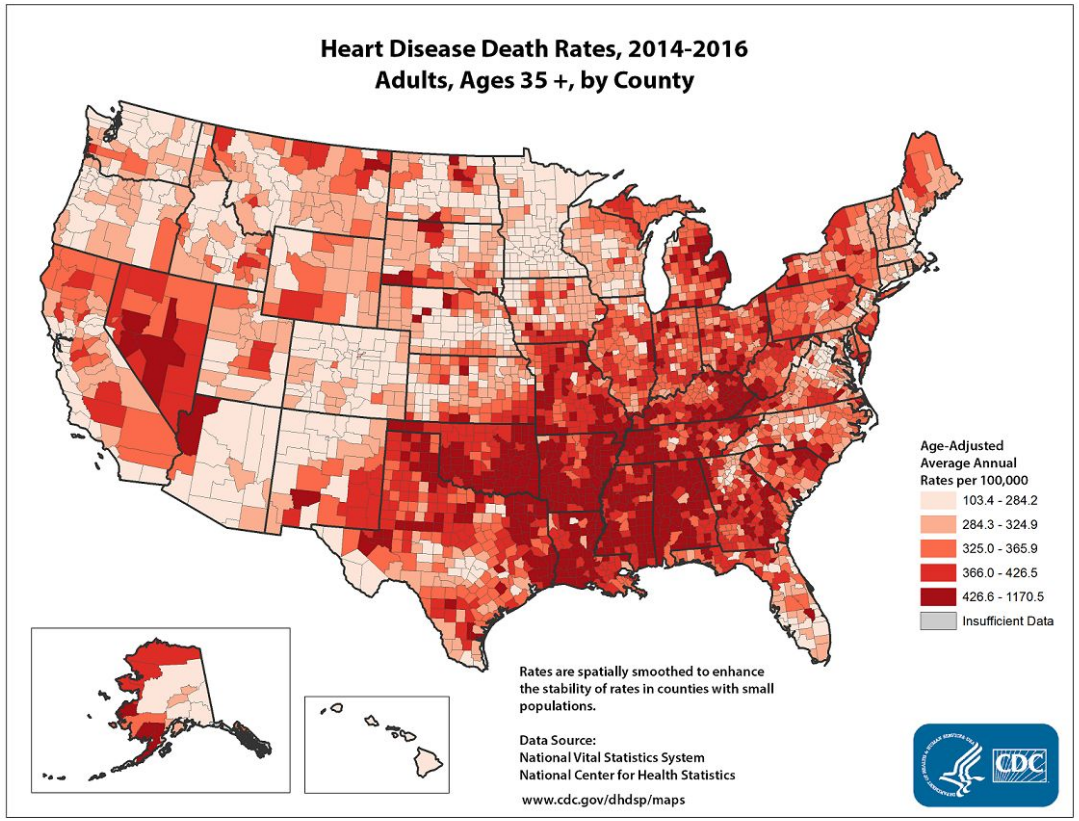
Building Predictive Models for Heart Disease

Alex Teboul

Models: Random Forest, Gradient Boosting, AdaBoost, Neural Networks

Data: 2015 BRFSS Survey

Why should you care about Heart Disease?

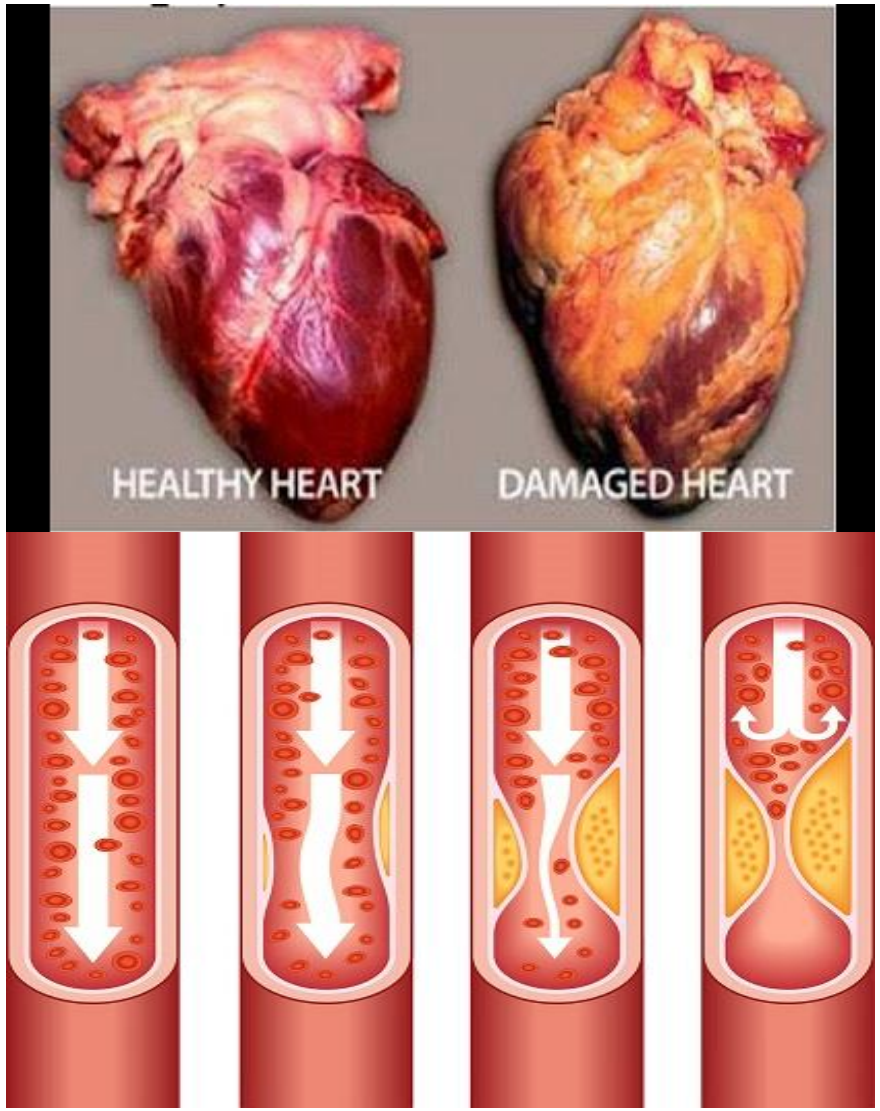


- Heart disease is the leading cause of death in the U.S.
- 1 in 4 deaths
- $219 Billion
- All across the country

How do you get Heart Disease?

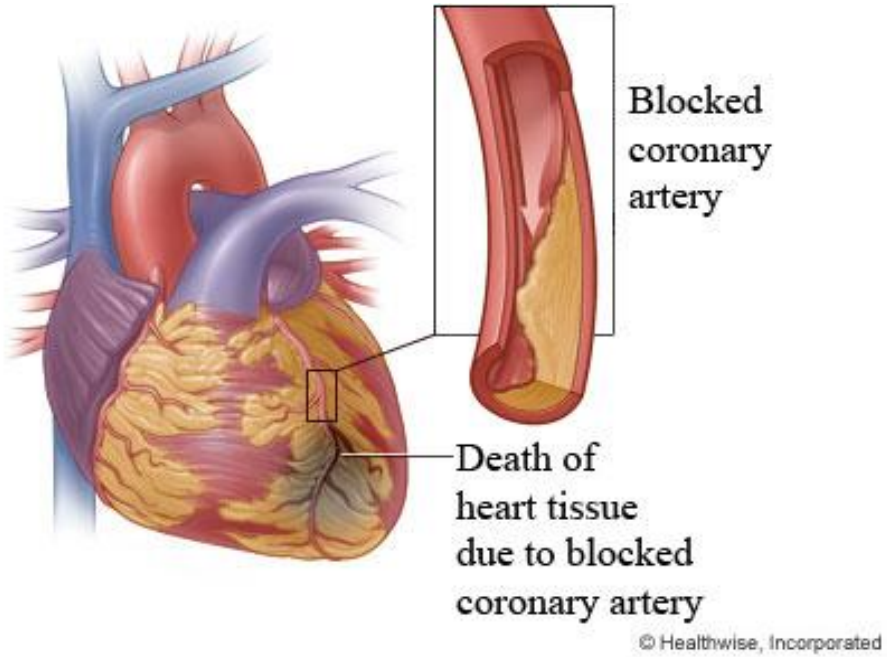
Multiple Risk Factors

- Lifestyle + Genetics
- Unhealthy Diet
- Physical Inactivity
- Alcohol
- Smoking
- Obesity
- Diabetes



Binary Classification: Heart Disease and Heart Attack

1 vs. 0

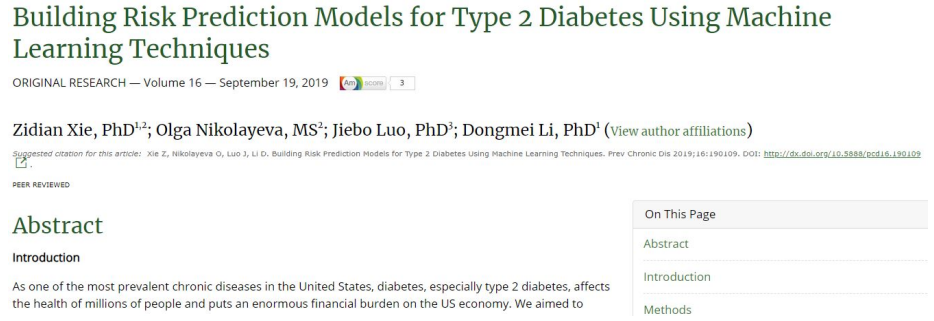


Part 1: Getting and Cleaning the Data

- **Dataset:** Behavioral Risk Factor Surveillance System Survey (BRFSS 2015)
- Initial Feature Selection
- Cleaning
- Addressing Class Imbalance

Dataset: Behavioral Risk Factor Surveillance System Survey (BRFSS 2015)

- U.S. Health Survey by Telephone
- 330 Features
- 441,456 Responses
- Health-Related Risk Behaviors
- Chronic Health Conditions
- ML Techniques in Literature



Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques

ORIGINAL RESEARCH — Volume 16 — September 19, 2019

Zidian Xie, PhD[1,2]; Olga Nikolayeva, MS[2]; Jiebo Luo, PhD[3]; Dongmei Li, PhD[1] (View author affiliations)

Suggested citation for this article: Xie Z, Nikolayeva O, Luo J, Li D. Building Risk Prediction Models for Type 2 Diabetes Using Machine Learning Techniques. Prev Chronic Dis 2019;16:190109. DOI: http://dx.doi.org/10.5888/pcd16.190109

PEER REVIEWED

Abstract

Introduction

As one of the most prevalent chronic diseases in the United States, diabetes, especially type 2 diabetes, affects the health of millions of people and puts an enormous financial burden on the US economy. We aimed to

On This Page

Abstract

Introduction

Methods

Initial Feature Selection - Response Variable

- **Response Variable / Dependent Variable: (1)**
- Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)
  - '_MICHD'
  - Renamed as 'HeartDiseaseorAttack'

Ever had CHD or MI

CalculatedVariables: 6.1 Calculated Variables

Type: Num

Column: 1899

SAS Variable Name: _MICHD

Prologue:

Description: Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Reported having MI or CHD Notes: CVDINFR4=1 OR CVDCRHD4=1	38,633	8.83	6.42
2	Did not report having MI or CHD Notes: CVDINFR4=2 AND CVDCRHD4=2	398,881	91.17	93.58
BLANK	Not asked or Missing Notes: CVDINFR4=7, 9 OR MISSING OR CVDCRHD4=7, 9, OR MISSING	3,942		

Initial Feature Selection - Features

- **Renamed all features for clarity**

- <u>Features / Independent Variables: (22)</u>

- **High Blood Pressure** - HighBP
- **High Cholesterol** - HighChol, CholCheck
- **BMI -** BMI
- **Smoking -** Smoker
- **Other Chronic Health Conditions -** Stroke, Diabetes
- **Physical Activity -** PhysActivity
- **Diet** - Fruits, Veggies
- **Alcohol Consumption -** HvyAlcoholConsump
- **Health Care** - AnyHealthcare, NoDocbcCost
- **General and Mental Health** - GenHlth, MentHlth, PhysHlth, DiffWalk
- **Demographics** - Sex, Age, Education, Income

Cleaning

- Used BRFSS Codebook:

Ever had CHD or MI

CalculatedVariables: 6.1 Calculated Variables **Type:** Num

Column: 1899 **SAS Variable Name:** _MICHD

Prologue:

Description: Respondents that have ever reported having coronary heart disease (CHD) or myocardial infarction (MI)

Value	Value Label	Frequency	Percentage	Weighted Percentage
1	Reported having MI or CHD Notes: CVDINFR4=1 OR CVDCRHD4=1	38,633	8.83	6.42
2	Did not report having MI or CHD Notes: CVDINFR4=2 AND CVDCRHD4=2	398,881	91.17	93.58
BLANK	Not asked or Missing Notes: CVDINFR4=7, 9 OR MISSING OR CVDCRHD4=7, 9, OR MISSING	3,942		

- Removed all Missing Values
- Removed all ‘Don’t know/Not Sure’ and ‘Refused to Answer’
- Variables Modified to be Ordinal (1,2,3,4…) or Binary (0,1)
- Final Dataset: 253,680 rows and 22 columns

Addressing Class Imbalance:

Full Dataset	
HeartDiseaseorAttack	
0.0	229,787
1.0	23,893

50-50	
*HeartDiseaseorAttack	
0.0	23,893
1.0	23,893

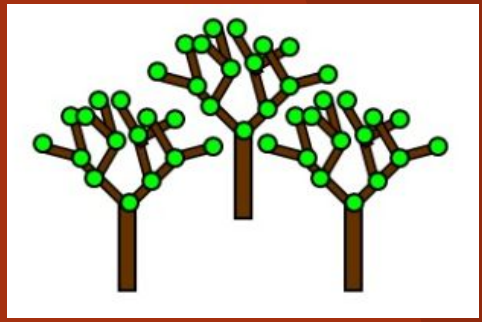
60-40	
*HeartDiseaseorAttack	
0.0	47,786
1.0	23,893

*Random Subsets from 0 (no Heart Disease) and all 1 (has Heart Disease)

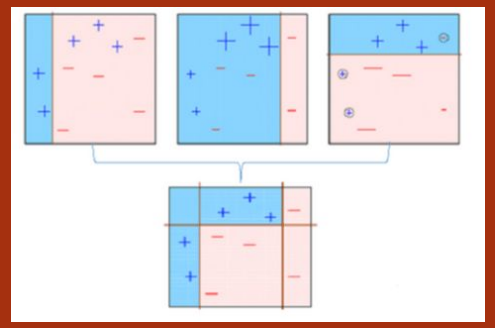
Part 2: Model Building

- Random Forests
- Gradient Boosting
- AdaBoost
- Neural Networks

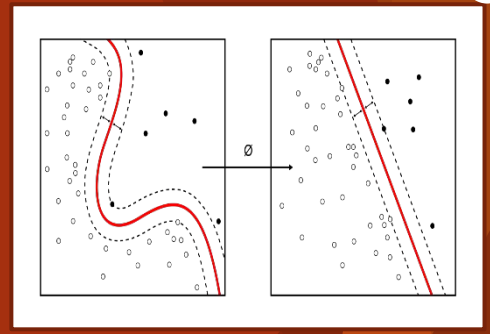
Random Forests



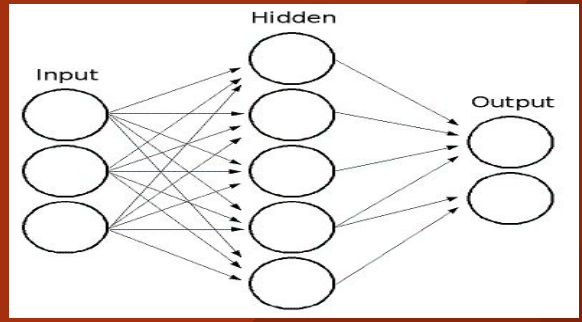
AdaBoost

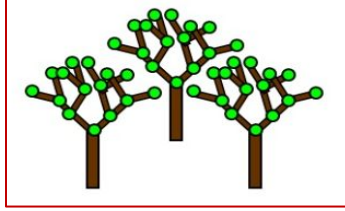


Gradient Boosting



Neural Networks



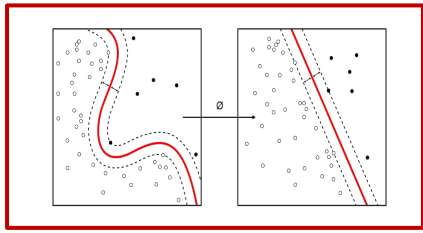


Random Forests

Dataset	Model	Accuracy	AUC	Runtime
Full Dataset	RF w/ Feature Selection	0.89 (+/- 0.00)	0.74 (+/- 0.01)	48 sec
50-50 Dataset	RF w/ Feature Selection	0.72 (+/- 0.01)	0.78 (+/- 0.01)	10 sec
60-40 Dataset	RF w/ Feature Selection	0.73 (+/- 0.01)	0.78 (+/- 0.01)	15 sec
Full Dataset	RF w/o Feature Selection	0.90 (+/- 0.00)	0.82 (+/- 0.01)	66 sec
50-50 Dataset	RF w/o Feature Selection	0.76 (+/- 0.02)	0.83 (+/- 0.01)	12 sec

- Best Parameter Setting Results Displayed: Different CV and #Trees Tested.
- 50 trees, 5-fold CV Reported

- Full Dataset Selected: ['BMI', 'GenHlth', 'MentHlth', 'PhysHlth', 'Age', 'Education', 'Income']
- Balanced Datasets Selected: ['HighBP', 'BMI', 'GenHlth', 'MentHlth', 'PhysHlth', 'Age', 'Education', 'Income']

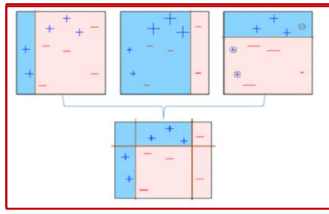


Gradient Boosting

Dataset	Model	Accuracy	AUC	Runtime
Full Dataset	GB w/ Feature Selection	0.91 (+/- 0.00)	0.85 (+/- 0.01)	54 sec
50-50 Dataset	GB w/ Feature Selection	0.76 (+/- 0.01)	0.84 (+/- 0.01)	8 sec
60-40 Dataset	GB w/ Feature Selection	0.78 (+/- 0.01)	0.84 (+/- 0.01)	13 sec
Full Dataset	GB w/o Feature Selection	0.91 (+/- 0.00)	0.85 (+/- 0.01)	153 sec

- Best Parameter Setting Results Displayed: n_estimators, loss, and max_depth
- 5-fold CV Reported, 100 estimators, loss='deviance', max_depth=3

- Full Dataset Selected: ['HighBP', 'HighChol', 'Stroke', 'GenHlth', 'DiffWalk', 'Sex', 'Age']
- 50-50 Selected: ['HighBP', 'HighChol', 'GenHlth', 'Sex', 'Age']
- 60-40 Selected: ['HighBP', 'HighChol', 'Stroke', 'GenHlth', 'Sex', 'Age']

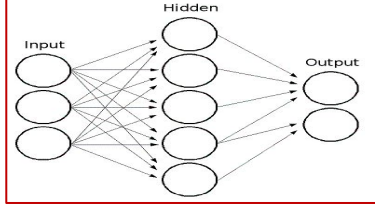


AdaBoost

Dataset	Model	Accuracy	AUC	Runtime
Full Dataset	Ada w/ Feature Selection	0.91 (+/- 0.00)	0.84 (+/- 0.01)	49 sec
50-50 Dataset	Ada w/ Feature Selection	0.76 (+/- 0.01)	0.83 (+/- 0.01)	8 sec
60-40 Dataset	Ada w/ Feature Selection	0.77 (+/- 0.01)	0.84 (+/- 0.01)	13 sec
Full Dataset	Ada w/o Feature Selection	0.91 (+/- 0.00)	0.84 (+/- 0.01)	92 sec

- Best Parameter Setting Results Displayed: n_estimators, learning_rate
- 5-fold CV Reported, 100 estimators, learning_rate=0.1

- Full Dataset Selected: ['HighBP', 'HighChol', 'Stroke', 'GenHlth', 'DiffWalk', 'Sex', 'Age']
- 50-50 Selected: ['HighBP', 'HighChol', 'GenHlth', 'Sex', 'Age']
- 60-40 Selected: ['HighBP', 'HighChol', 'Stroke', 'GenHlth', 'Sex', 'Age']



Neural Networks

Dataset	Model	Accuracy	AUC	Runtime
Full Dataset	NN w/ Feature Selection	0.91 (+/- 0.00)	0.84 (+/- 0.01)	36 sec
50-50 Dataset	NN w/ Feature Selection	0.76 (+/- 0.01)	0.84 (+/- 0.01)	18 sec
60-40 Dataset	NN w/ Feature Selection	0.78 (+/- 0.01)	0.84 (+/- 0.01)	21 sec
Full Dataset	NN w/o Feature Selection	0.91 (+/- 0.00)	0.85 (+/- 0.01)	113 sec

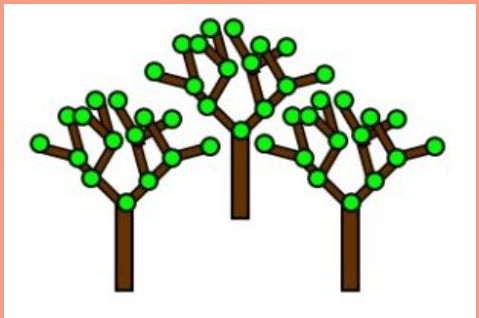
- Best Parameter Setting Results Displayed: solver, activation, alpha
- 5-fold CV Reported, solver='adam', activation='logistic', alpha=0.0001

- Full Dataset Selected: ['HighBP', 'HighChol', 'Stroke', 'GenHlth', 'DiffWalk', 'Sex', 'Age']
- 50-50 Selected: ['HighBP', 'HighChol', 'GenHlth', 'Sex', 'Age']
- 60-40 Selected: ['HighBP', 'HighChol', 'Stroke', 'GenHlth', 'Sex', 'Age']

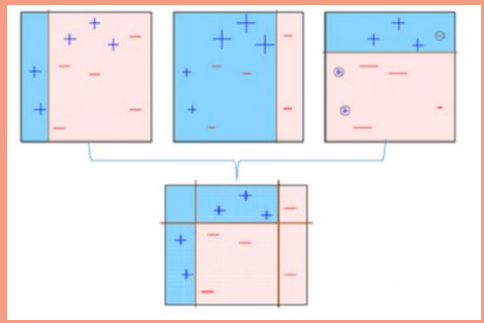
To Review

- 4 Models
- w/ Feature Selection
- w/o Feature Selection
- Full Dataset, 50-50, 60-40
- Accuracy, AUC, Runtime

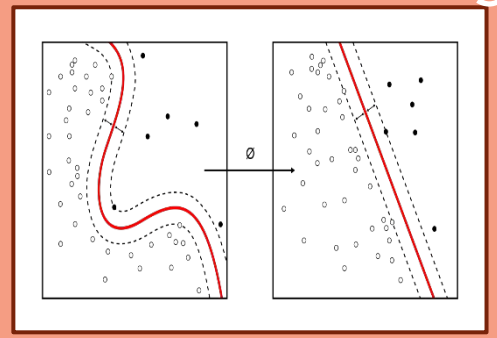
Random Forests



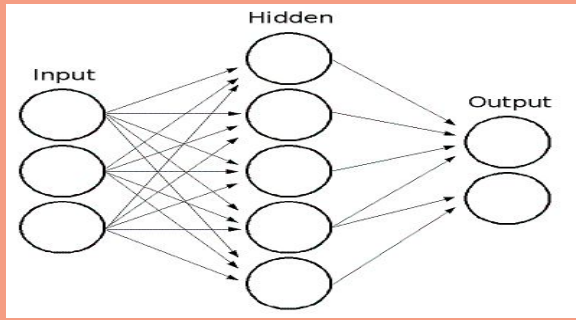
AdaBoost

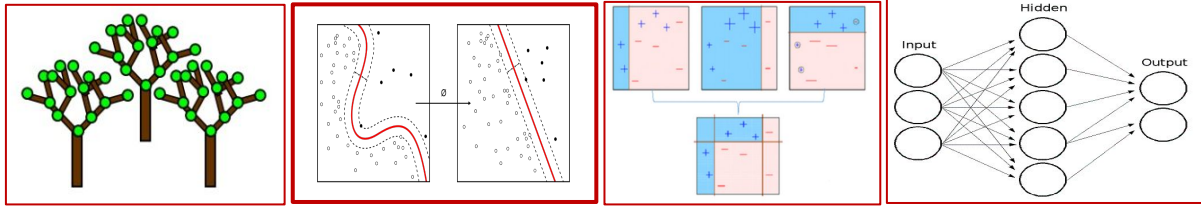


Gradient Boosting



Neural Networks





The Best of the Best

Dataset	Model	Accuracy	AUC	Runtime
Full Dataset	**NN w/ Feature Selection**	**0.91 (+/- 0.00)**	**0.84 (+/- 0.01)**	**36 sec**
50-50 Dataset	GB w/ Feature Selection	0.76 (+/- 0.01)	0.84 (+/- 0.01)	8 sec
60-40 Dataset	GB w/ Feature Selection	0.78 (+/- 0.01)	0.84 (+/- 0.01)	13 sec
Full Dataset	Ada w/o Feature Selection	0.91 (+/- 0.00)	0.84 (+/- 0.01)	92 sec

- Near identical performance between Gradient Boosting, AdaBoost, and Neural Networks
- Best Models Selected by Accuracy, AUC, Runtime

Important Features

RF Selected Features:

- BMI
- GenHlth
- MentHlt
- PhysHlt
- Age
- Education
- Income

GB, Ada, NN Selected Features

- HighBP
- HighChol
- Stroke
- GenHlth
- DiffWalk
- Sex
- Age

Remember…

Eat Healthy Foods.

Increase your Physical Activity

Especially Important if you're over age 65!

Don't Forget About your Mental Health, just breath…

Today we looked into Building Predictive Models for Heart Disease using the BRFSS 2015.

Random Forests, Gradient Boosting, AdaBoost, Neural Networks

Dataset	Model	Accuracy	AUC	Runtime
Full Dataset	NN w/ Feature Selection	0.91 (+/- 0.00)	0.84 (+/- 0.01)	36 sec

Alex Teboul
DSC 540: Advanced Machine Learning
Professor: Casey Bennett