HEART DISEASE CLASSIFICATION

Project Background

Originally completed in Spring 2023 in Data Mining class

Worked with two other group members

Assignment

- Explore and visualize data
- Apply machine learning algorithms
- Evaluate performance
- Present findings

Motivation and Objectives

- Why? 7.2% of Americans over 20 years old have heart disease, costing $240 billion annually[1]
- What? Use survey responses from individuals in the United States to predict if the person answers in the survey that they have heart disease
- Used two survey questions combined into one feature to train supervised models:
  - *(Ever told) (you had) angina or coronary heart disease? (CAD)*
  - *(Ever told) you had a heart attack, also called a myocardial infarction? (MI)*

[1] https://www.cdc.gov/heartdisease/facts.htm

About the Data

Full dataset available for direct download on the Center for Disease Control website

Survey conducted by the Behavioral Risk Factor Surveillance System (BRFSS) administered in 2021[2]

438,693 participants (instances), 303 survey questions (features)

Data is consistent and highly structured with no major anomalies or NA values

Most responses are coded and requires heavy reference to BRFSS Codebook[3]

No names, addresses, phone numbers, or other personal information given to deduce who took the survey

[2] https://www.cdc.gov/brfss/annual_data/annual_2021.html
[3] https://www.cdc.gov/brfss/annual_data/2021/pdf/codebook21_llcp-v2-508.pdf

Original Assumptions

- Selected features based on subject matter knowledge from the American Heart Association[4]
- Imputed missing values based on mean, median, mode, grouped by age, sex, region
- Replaced response “Don’t know” and “Refused” with “No”

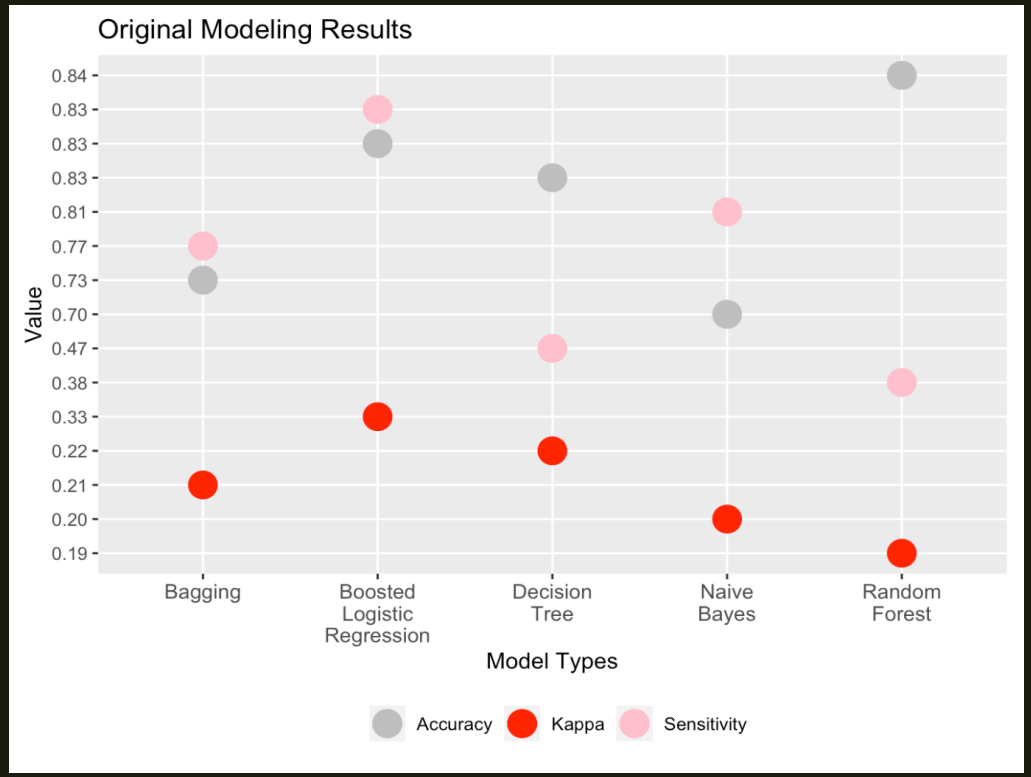
```
```{r codeEx, eval = FALSE, echo = TRUE}

round(prop.table(table(heart$X_PHYS14D, useNA = "always")) * 100, digits = 2)

#Mode by groupings to check which mode to use
heart %>% group_by(X_AGE_G) %>% summarise(mode = getmode(X_PHYS14D))
heart %>% group_by(SEXVAR) %>% summarise(mode = getmode(X_PHYS14D))
heart %>% group_by(X_REGION) %>% summarise(mode = getmode(X_PHYS14D))

#Impute using mode
heart <- heart %>% mutate(X_PHYS14D = ifelse(X_PHYS14D == 9, NA,
                                             X_PHYS14D))
heart <- heart %>% mutate(X_PHYS14D = ifelse(is.na(X_PHYS14D), getmode(X_PHYS14D),
                                             X_PHYS14D))
```

[4] https://www.heart.org/en/health-topics/heart-attack/understand-your-risks-to-prevent-a-heart-attack
```



Initial Results

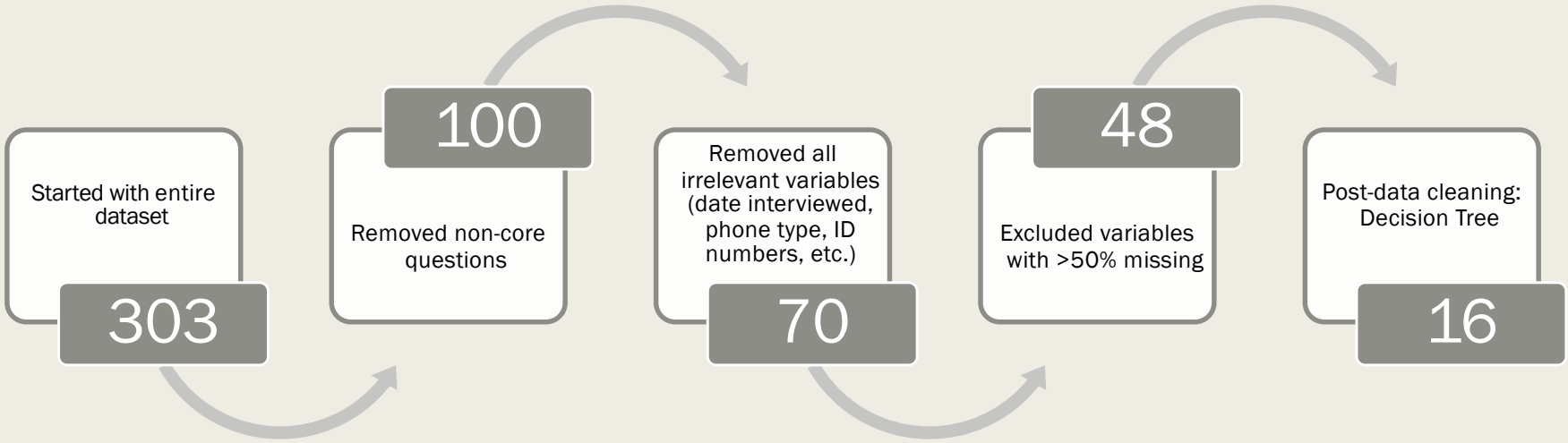
- Chose sensitivity to evaluate models
- Kappa considered because of class imbalance (8:92)
- Trained five different machine learning algorithms
- Logistic regression model gave the best outcome:
  - 83% accuracy
    - 10% less than no information rate (if you always guessed the person didn't have heart disease)
  - 83% sensitivity (true positive rate)
  - Kappa value 0.33 (adjusts accuracy to discount chance alone)

Goals for Improvement

Automate recoding process to reduce manual clean-up time

Leave ‘Don’t Know’ and ‘Not Sure’ responses to see if model performance improves

Begin with whole dataset and use decision tree to reduce number of features prior to modeling



Data Reduction, Revisited

Code file:

```
PRIMINSR
    .  = "Not asked or Missing"
    .D = "DK/NS"
    .R = "REFUSED"
    1  = "A plan purchased through an employer or union (including plans purchased through another person´s employer)"
    2  = "A private nongovernmental plan that you or another family member buys on your own"
    3  = "Medicare"
    4  = "Medigap"
    5  = "Medicaid"
    6  = "Children´s Health Insurance Program (CHIP)"
    7  = "Military related health care: TRICARE (CHAMPUS) / VA health care / CHAMP-VA"
    8  = "Indian Health Service"
    9  = "State sponsored health plan"
    10 = "Other government program"
    77 = "Don't know/Not Sure"
    88 = "No coverage of any type"
    99 = "Refused"
    ;
```

Dataset pre/post recode:

PRIMINSR	PERSDOC3	MEDCOST1	CHECKUP1	EXERANY2	BPHIGH6
3	1	2	2	2	3
1	2	2	1	1	1
2	2	2	1	2	1
2	1	2	1	1	1
3	1	2	1	1	4

PRIMINSR	PERSDOC3	MEDCOST1	CHECKUP1	EXERANY2	BPHIGH6
Medicare	Yes, only one	No	Within past 2 years (1 year but < 2 years ago)	No	No
A plan purchased through an employer or union (incl...	More than one	No	Within past year (anytime < 12 months ago)	Yes	Yes
A private nongovernmental plan that you or another f...	More than one	No	Within past year (anytime < 12 months ago)	No	Yes
A private nongovernmental plan that you or another f...	Yes, only one	No	Within past year (anytime < 12 months ago)	Yes	Yes
Medicare	Yes, only one	No	Within past year (anytime < 12 months ago)	Yes	Told borderline high or pre-hypertensive or elevated ..

Automate Recoding Process

- Downloaded SAS file of variable codes and descriptions[5]
- Cleaned code file
- Created function to loop through data file and recode variables
- Some variables still required manual cleaning
- Results much easier to interpret
- Very time consuming to design

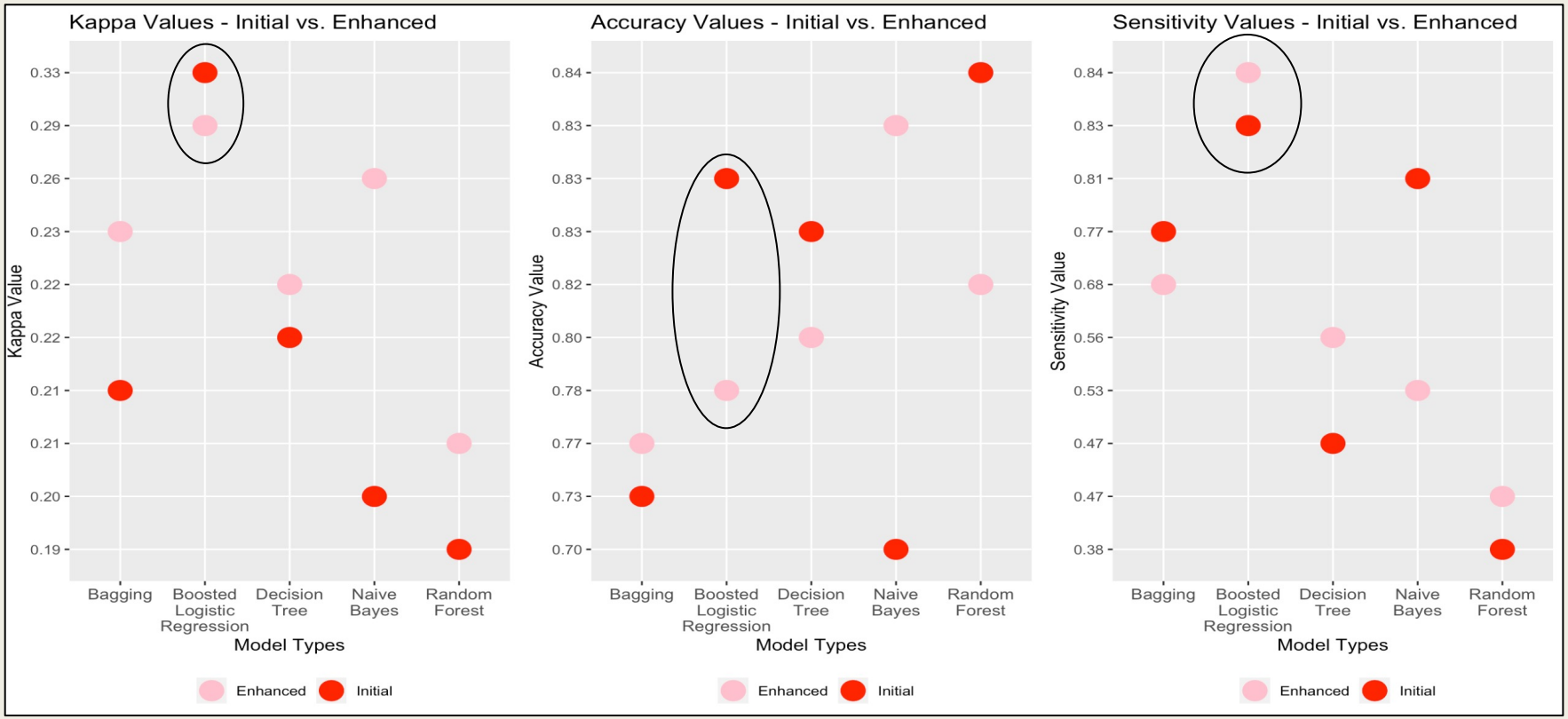
5 https://www.cdc.gov/brfss/annual_data/annual_2021.html

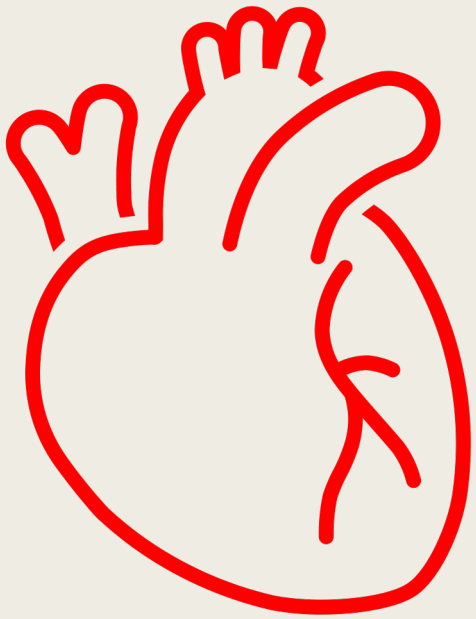
Data Pre-Processing, Summarized

Major differences:

- Started with entire dataset, rather than pre-selected variables
- Converted everything to code descriptions rather than codes
- Applied decision tree algorithm
- Final dataset for modeling: 16 features, ~439k instances

Model Results: Mixed Bag





Conclusion, Challenges and Further Work

- Re-coding automation was largely a success, but data cleaning still proves to be a major time sink in this dataset
- Large dataset is still a computational issue
  - *Need exposure to Big Data software or run processes on more powerful computer*
- Complex patterns in dataset may lead to “curse of dimensionality”
- Rather than use data for predictions, visualization and descriptive analytics would be the best use of this dataset