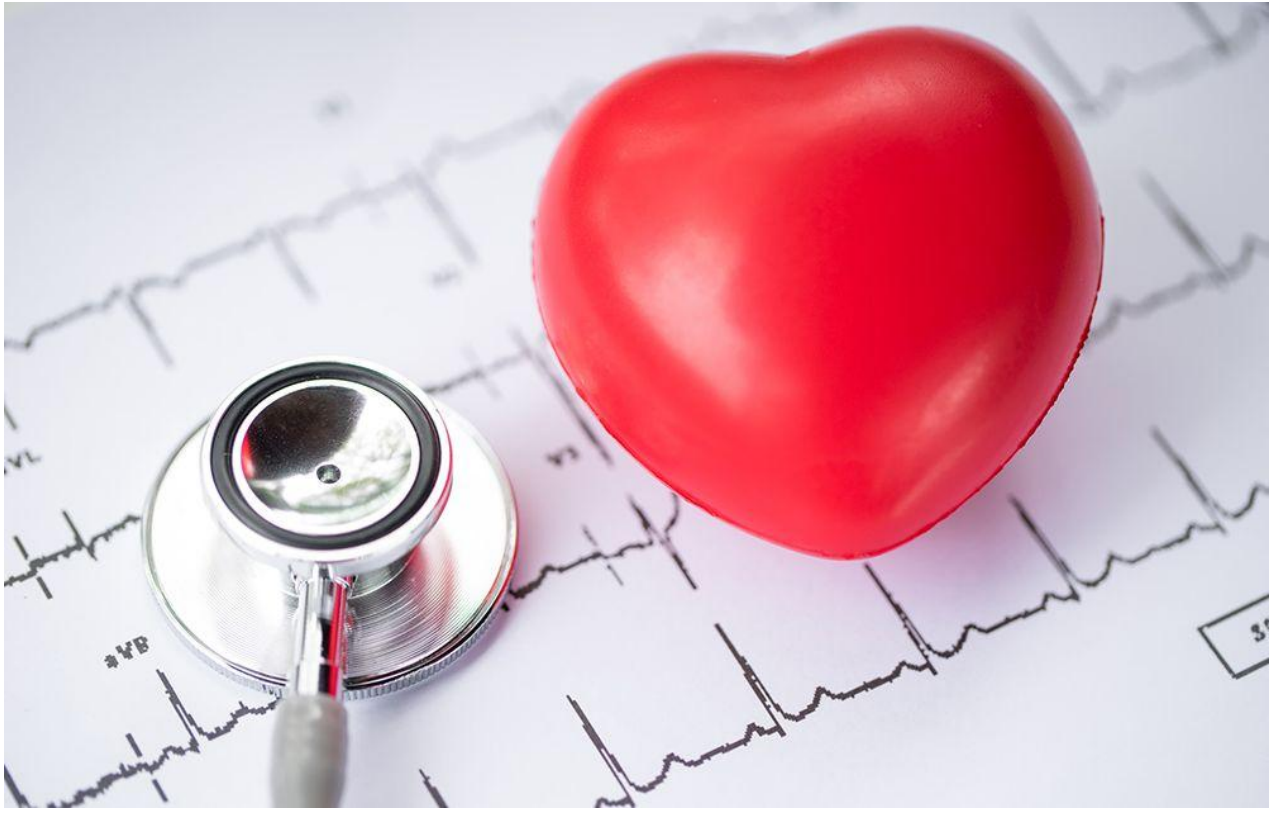
Heart Disease Classification

Group 7: Lucas Ogilvie
Mark Biernacki
Ran Ma

Background



Data Overview

1. Source: Behavioral Risk Factor Surveillance System (BRFSS) 2021

    Developed by the Centers for Disease Control and Prevention (CDC)

    Annual health-related telephone surveys to gather data across the U.S.

2. Size:

    438,693 rows

    303 columns

3. Content of the data:

    Demographic detail: age, sex, race

    Behavioral factors: smoking, alcohol consumption, physical activity levels

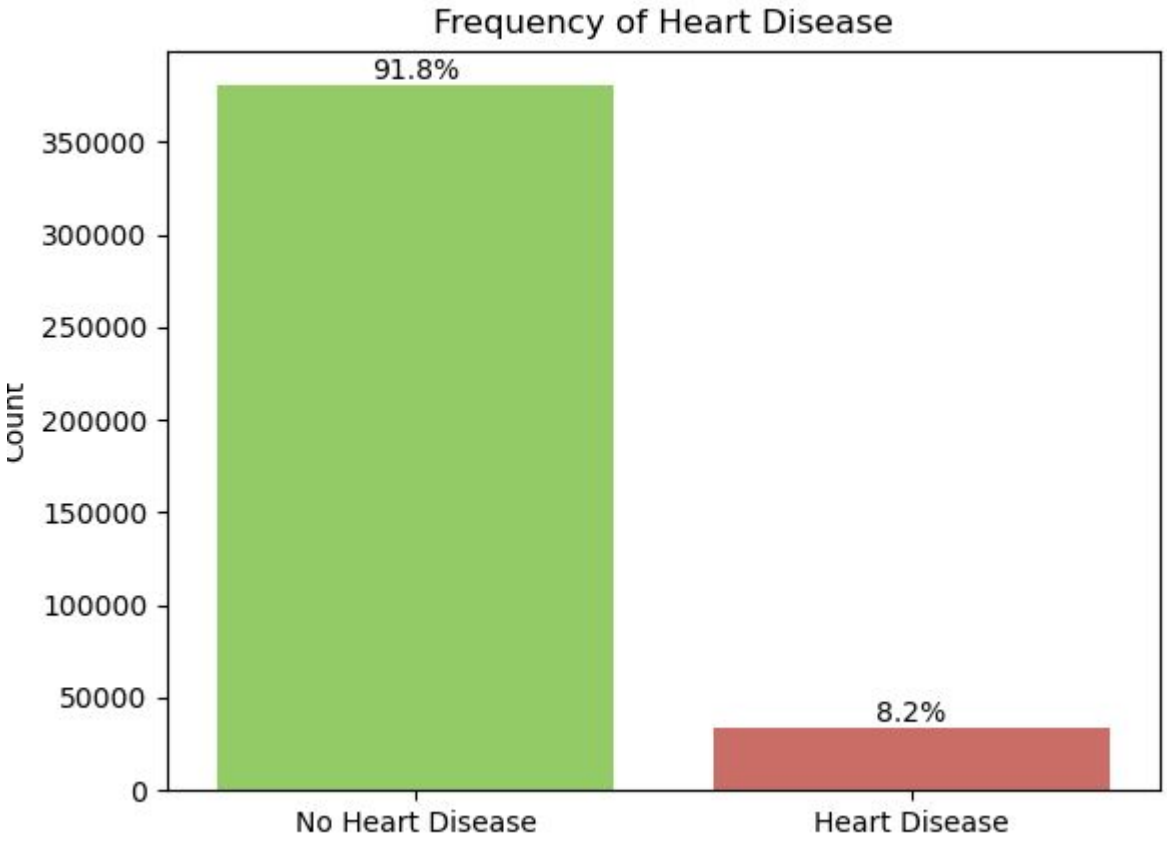
    Chronic health conditions

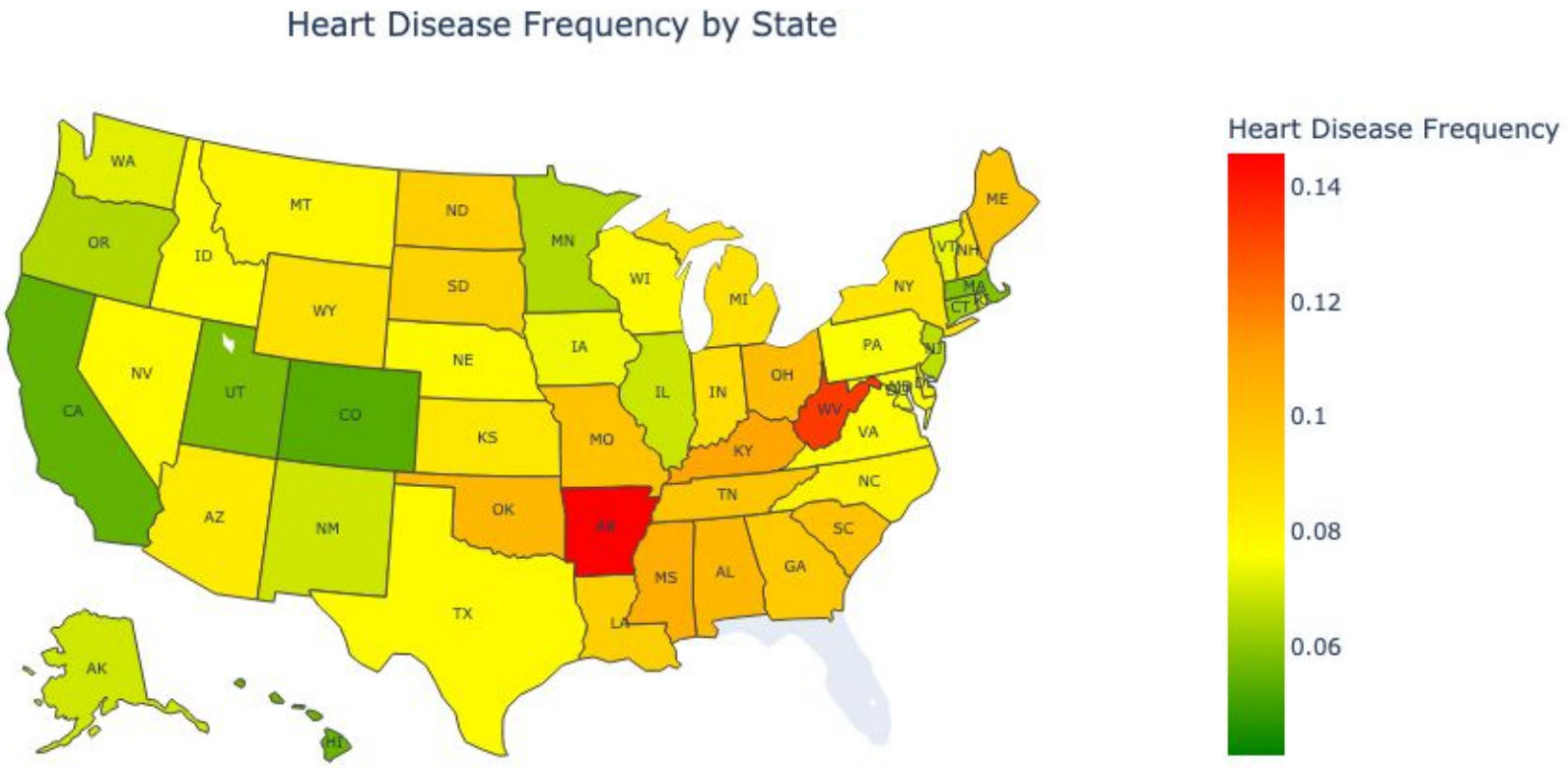
    Preventive health measures

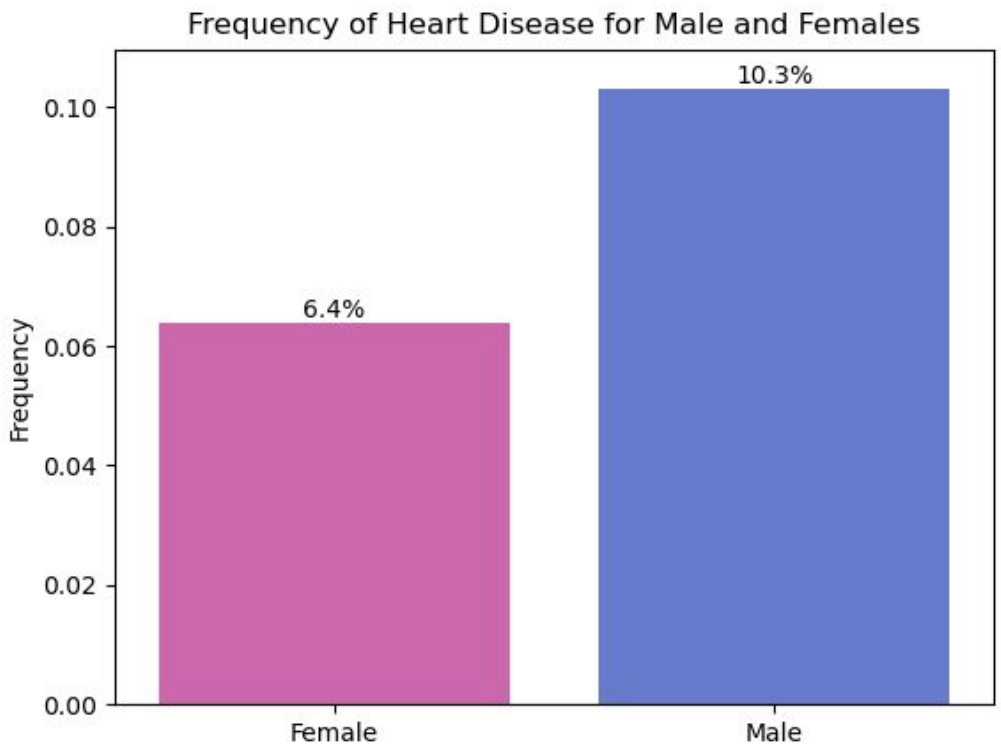
Data Cleaning

- Identify and remove irrelevant columns
- Remove columns with excessive missing values
- Missing values replacement
- Numeric to category conversion
- Missing Value imputation

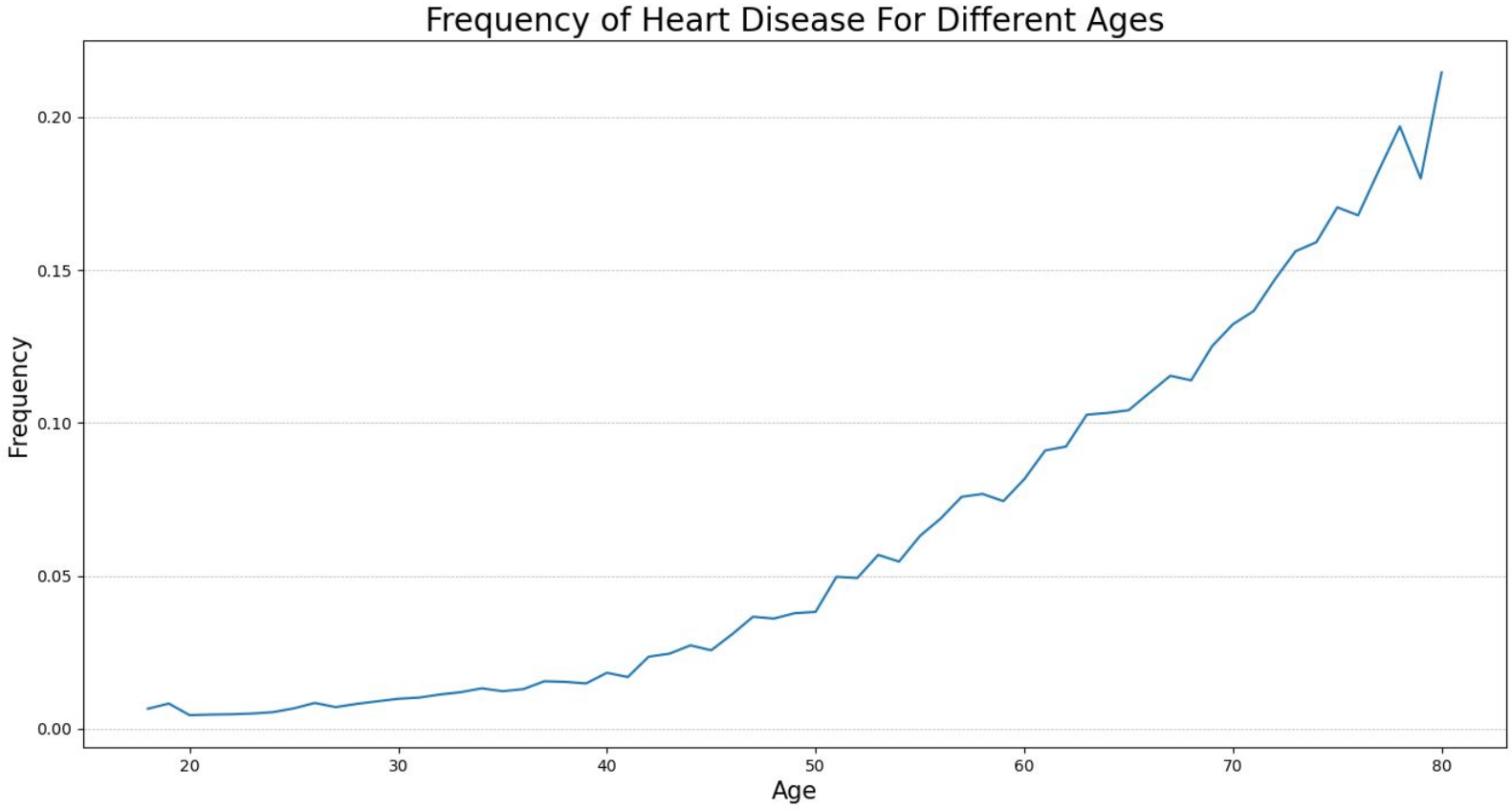
Imbalanced Data

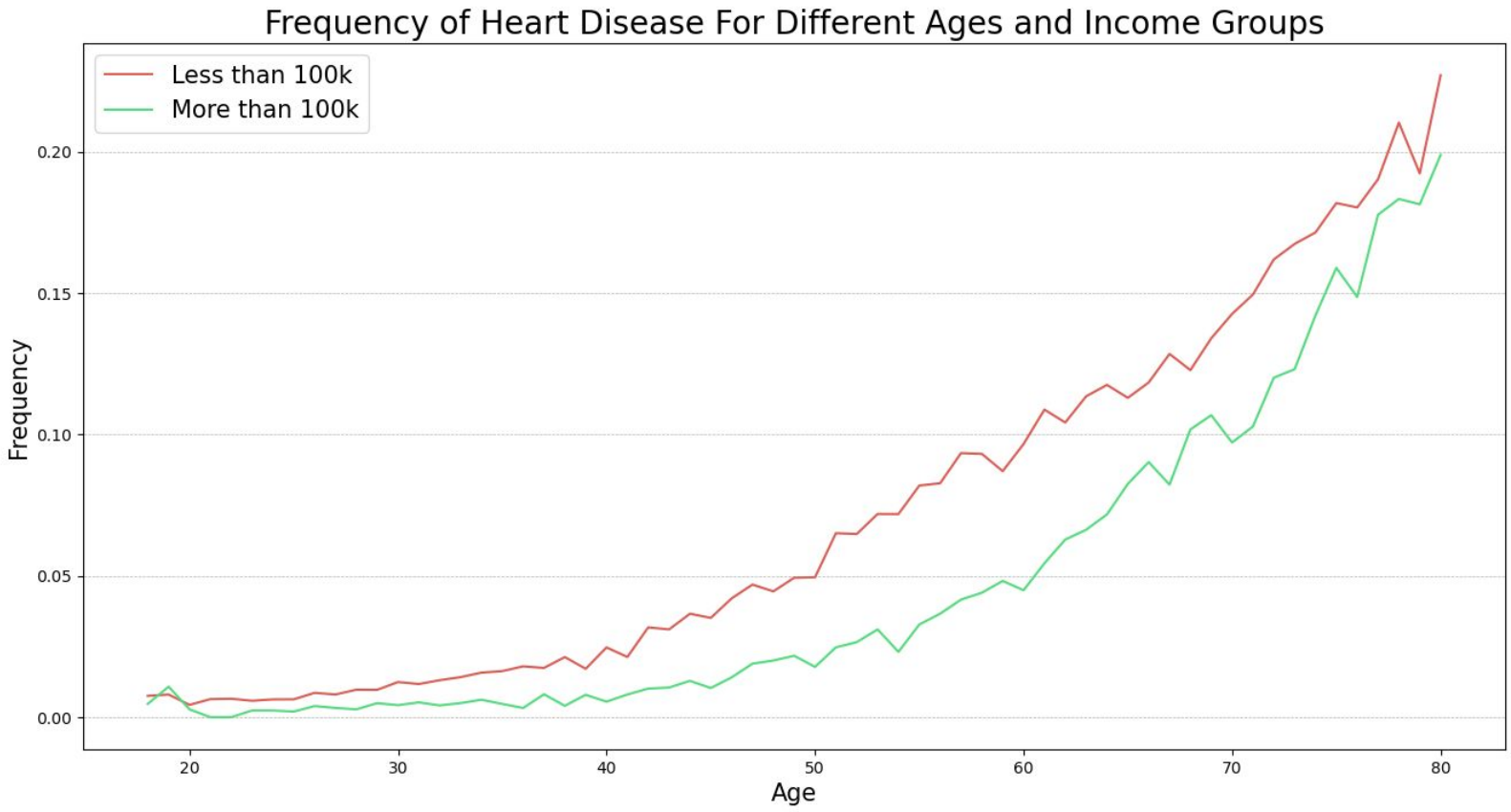


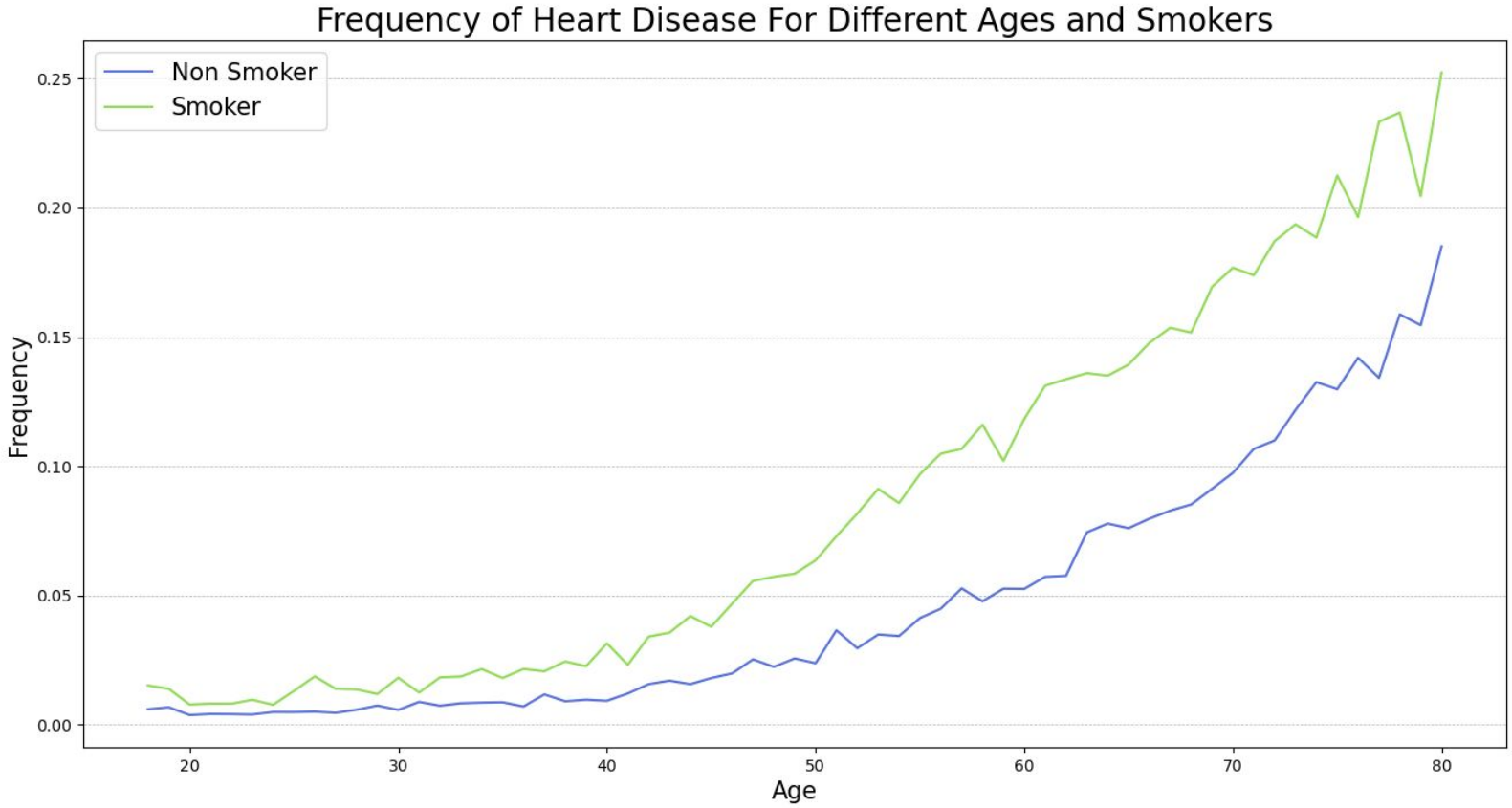

Heart Disease Frequency by State
Heart Disease Frequency
0.14
0.12
0.1
0.08
0.06
WA
OR
CA
NV
ID
MT
WY
UT
AZ
CO
NM
ND
SD
NE
KS
OK
TX
MN
IA
MO
AR
LA
WI
IL
MI
IN
KY
TN
MS
AL
OH
WV
VA
NC
SC
GA
FL
PA
NY
VT
NH
ME
MA
CT
RI
NJ
DE
MD
AK
HI



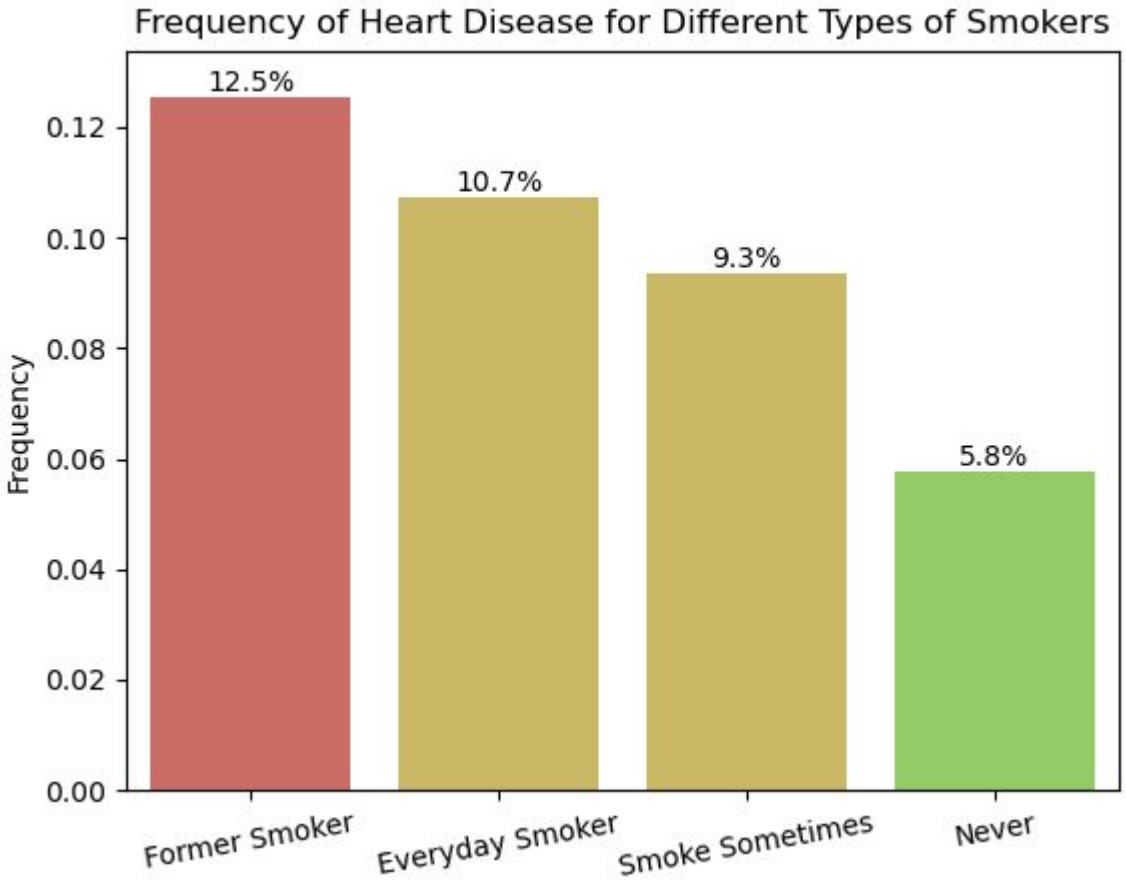
Women are less likely to have heart disease

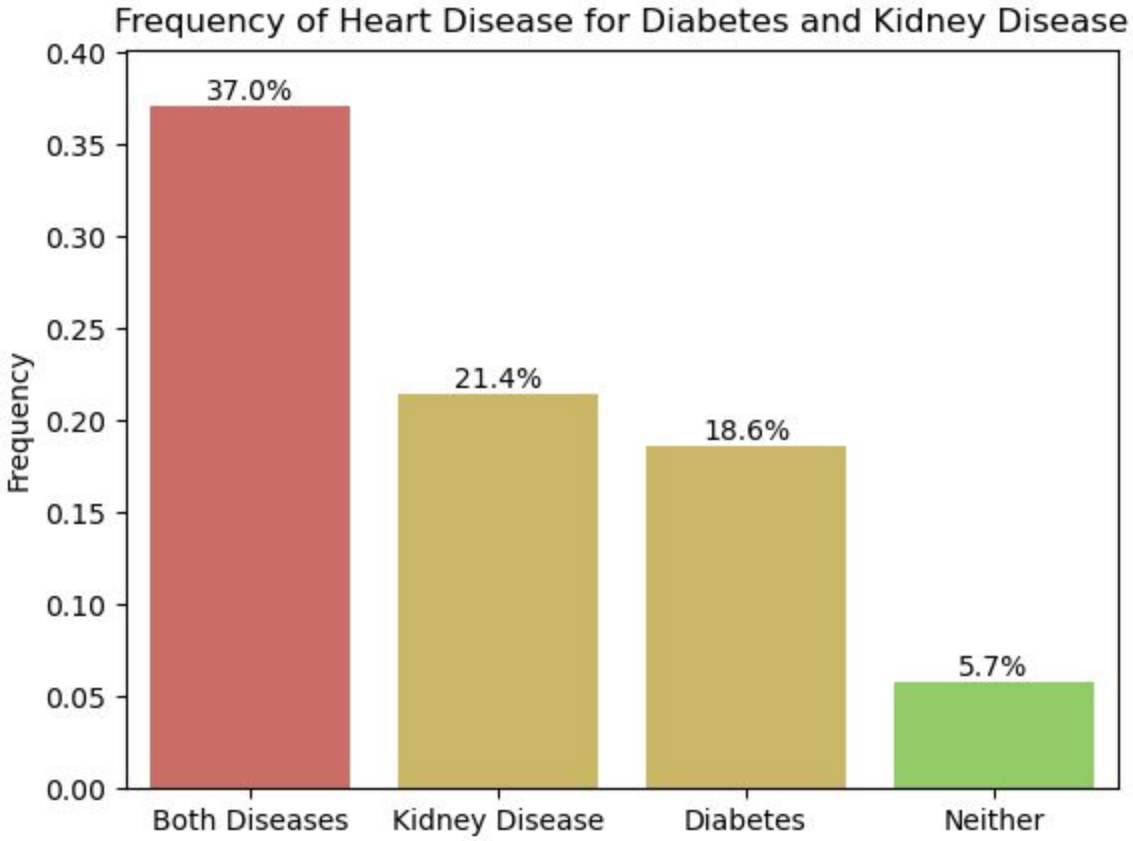

Frequency of Heart Disease For Different Ages
Frequency
0.20
0.15
0.10
0.05
0.00
20
30
40
50
60
70
80
Age


Frequency of Heart Disease For Different Ages and Income Groups
Less than 100k
More than 100k
Frequency
0.20
0.15
0.10
0.05
0.00
20
30
40
50
60
70
80
Age


Frequency of Heart Disease For Different Ages and Smokers
Non Smoker
Smoker
Frequency
0.25
0.20
0.15
0.10
0.05
0.00
20
30
40
50
60
70
80
Age

Smokers are more likely to have Heart Disease





“When it rains, it pours”

You are more likely to have heart disease if you have other underlying diseases

Modeling Techniques

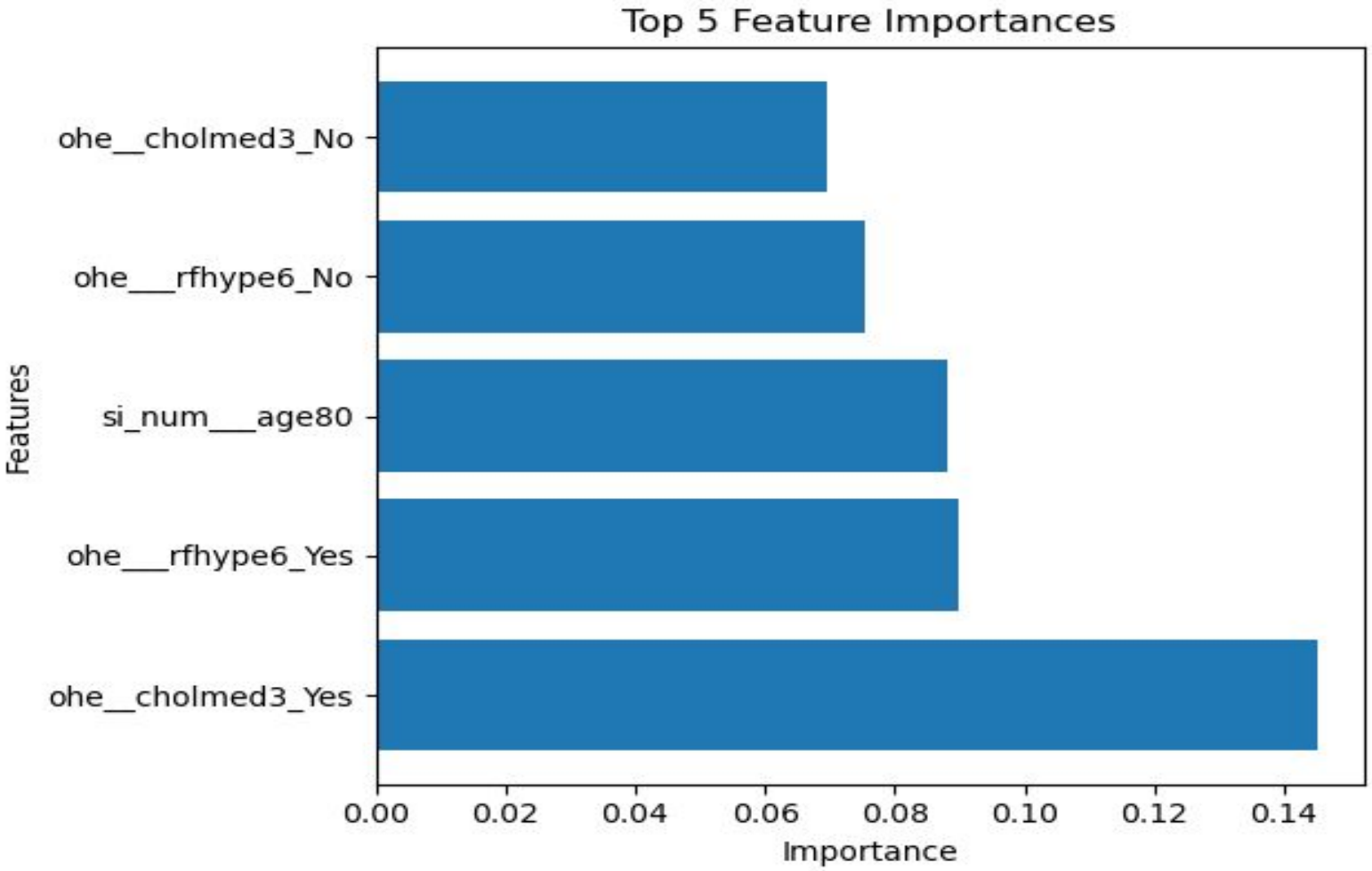
- Models: Multinomial Bayes, Logistic, XGBoost, and Random Forest classification models.
- Handling imbalance classes: Oversampling and Overweight minority class, Resampling with SMOTE and ADASYN.

Modeling Evaluation

Recall = True Positive Rate

Type of Model	balanced_accuracy	recall
Best NB Model	0.751111	0.754287
Best Logistic Model	0.784727	0.811017
XGBoost Model	0.786722	0.819591
Best Random Forest Model	0.772020	0.826056

Random Forest Model = Best Model


Top 5 Feature Importances
ohe__cholmed3_No
ohe___rfhype6_No
si_num___age80
ohe___rfhype6_Yes
ohe__cholmed3_Yes
Features
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Importance

Conclusion

- Model Application: Predicting the probability that a patient has heart disease will be useful as a screening process for healthcare professionals.
- Our model had a high balanced accuracy and an even higher recall score.
- Since the model was chosen to be sensitive towards positives, this can be used as a tool to identify people who are likely to have heart disease.