

# Improving Heart Disease Outcomes in the United States

Colin Beveridge, Ben Grudt, Grace Seiler, Hang Zhang Cao

# Group Introductions



Colin Beveridge

---

University of Notre Dame

Bachelor of Science:  
Physics



Hang Zhang Cao

---

Stony Brook University

Bachelor of Engineering:  
Chemical Engineering  
Applied Mathematics



Ben Grudt

---

Northern Illinois  
University

Bachelor of Science:  
Applied Mathematics



Grace Seiler

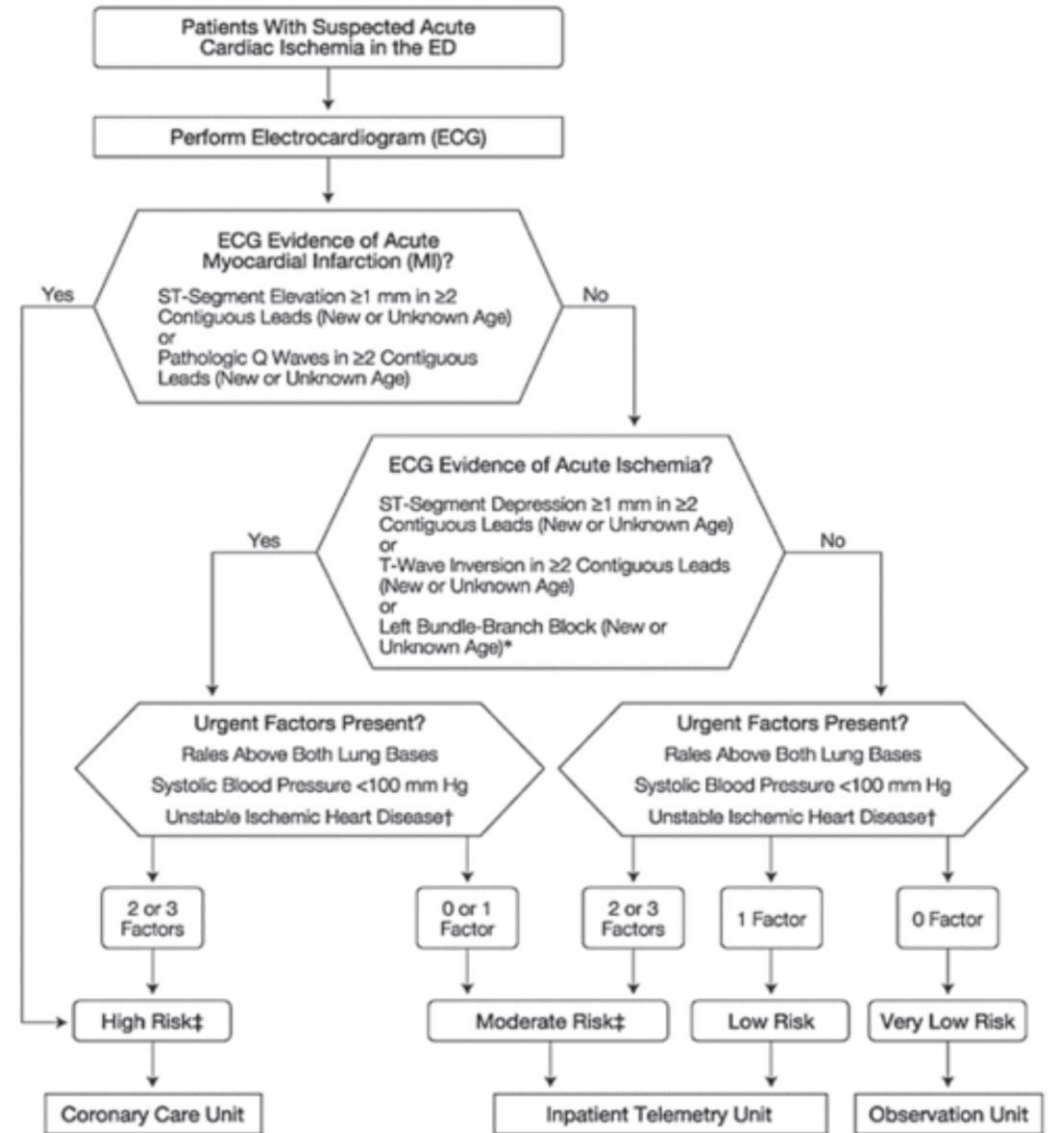
---

Luther College

Bachelor of Arts:  
Physics

# Project Idea

- ▶ ER patients with chest pain are a complex problem
  - ▶ Doctors don't always get this right
  - ▶ Huge consequences for mistakes
- ▶ Machine Learning - Decision Tree
  - ▶ Lee Goldman - 1970's
- ▶ Three key points:
  1. Accurate diagnosis is very important
  2. Doctors are not perfect at making these diagnoses
  3. A machine learning classifier has done this better than doctors

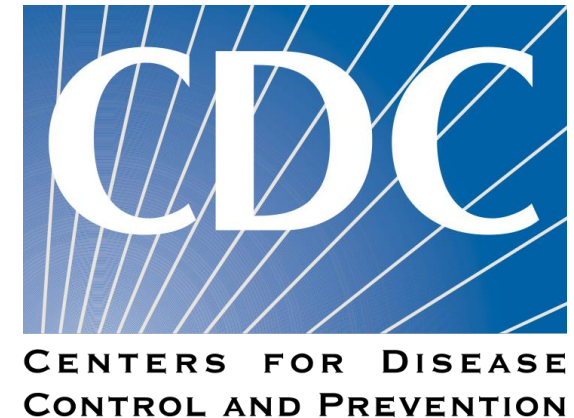


# Project Goal

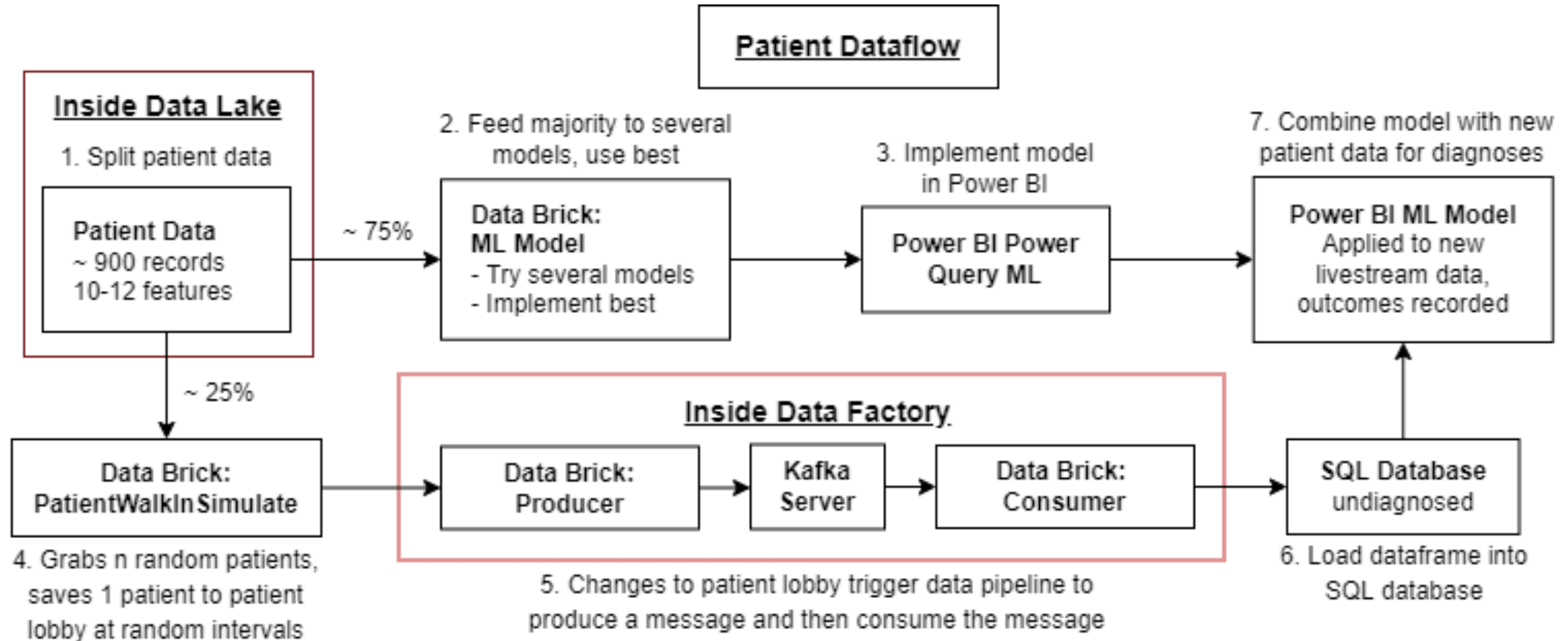
- ▶ Replicate the results of this decision tree with our own machine learning model
- ▶ Improve heart disease outcomes
- ▶ Hypothetical Client:
  - ▶ Goal: improve heart health in the United States and relieve heart patient bottleneck at hospitals
  - ▶ Method: Establish clinics in areas of highest need
  - ▶ Make recommendations for clinic sites and use our ML Model in clinics to help diagnose patients cheaply and quickly
    - ▶ What risk factors most influence heart disease rates?
    - ▶ How do we use these factors to decide areas in need of heart health services?

# Data Sources

- ▶ Census Insurance Data
  - ▶ Source: United States Census Bureau
  - ▶ Access: API Call
  - ▶ Cleaning: only call necessary columns
- ▶ CDC Behavioral Risk Factor Surveillance System
  - ▶ Source: Centers for Disease Control
  - ▶ Access: Download URL call in python
  - ▶ Cleaning: lots of extra columns to drop, filter recent data
- ▶ CDC Metropolitan Survey
  - ▶ Source: Centers for Disease Control
  - ▶ Access: Download URL call in python
  - ▶ Cleaning: lots of extra columns to drop, filter recent data
- ▶ Heart Disease Prediction Dataset
  - ▶ Source: Kaggle (combination of UCI ML Repository datasets)
  - ▶ Access: Direct download from Kaggle



# Live Patient Data Streaming



# Data Pre-Processing

- ▶ 688 total records of data available for training (75% of total data)
- ▶ 128 records (~19 %) of missing data on Cholesterol column.
  - ▶ Impute using mean from same gender
  - ▶ Age +- 10
  - ▶ Reduced importance in predicting heart disease
- ▶ Unbalanced dataset
  - ▶ 387 records of heart disease
  - ▶ 301 records of non-heart disease
  - ▶ Balanced using SMOTE module
- ▶ Data is normalized using StandardScaler
- ▶ Dummy variables are created for categorical columns

# ML Model

- ▶ Binary classification problem
- ▶ 5-fold cross validation to determine the optimum model
- ▶ Support-vector classification (SVC) won by a slight margin
- ▶ SVC is effective for small datasets with multiple dimensions.

Model	Accuracy	Precision	Recall	ROC
LogisticReg	0.8644	0.8650	0.8680	0.9349
KNeighbors	0.8687	0.8530	0.8888	0.9172
DecisionTree	0.8075	0.8030	0.8165	0.8075
RandomForest	0.8774	0.8639	0.8990	0.9297
BernoulliNB	0.8618	0.8643	0.8604	0.9257
GaussianNB	0.8618	0.8602	0.8681	0.9156
SVC	0.8799	0.8717	0.8965	0.9308



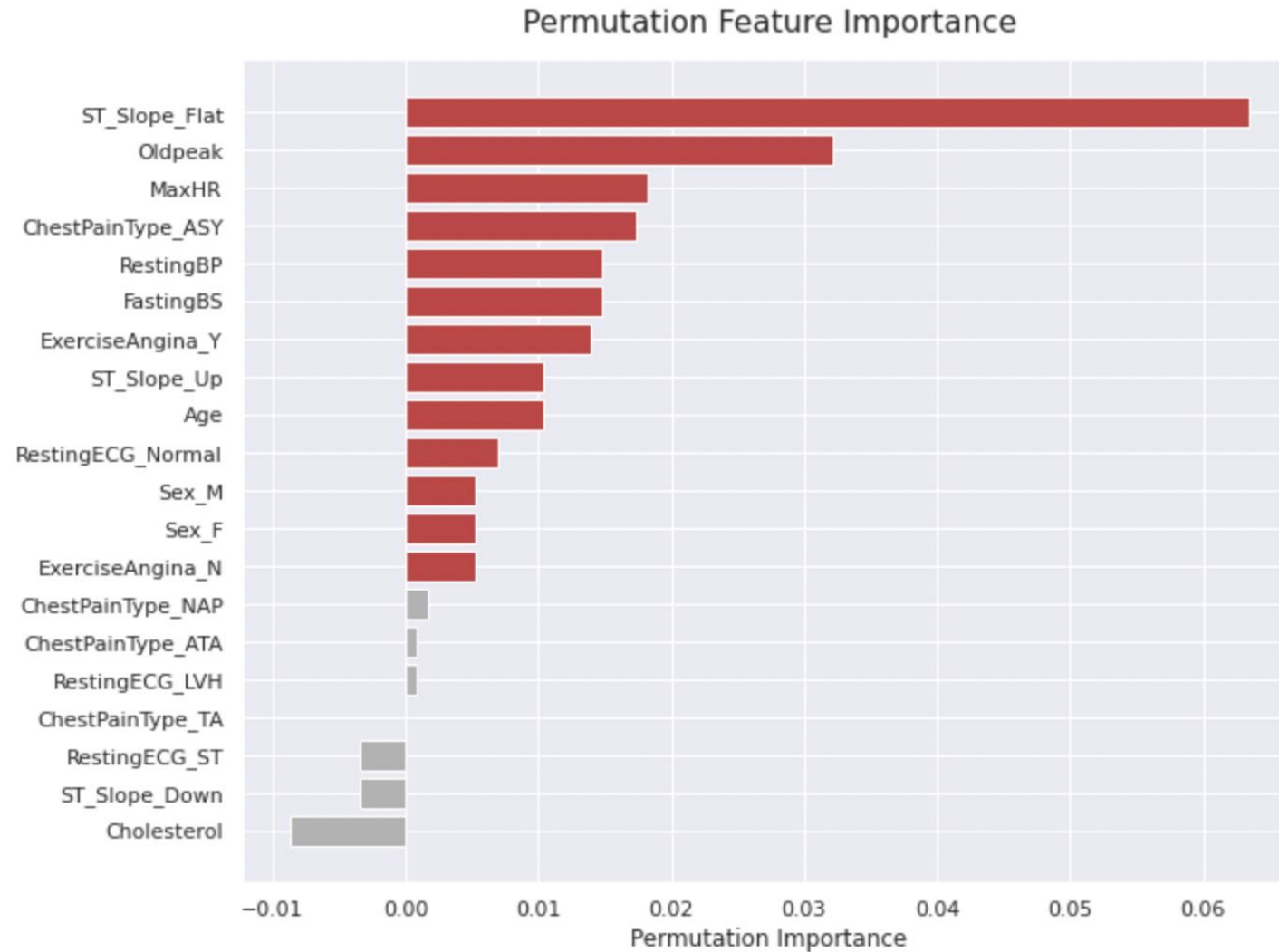
# SVC Tuning - Hyperparameters

- ▶ Fine tuning of hyperparameters are performed using Grid Search
  - ▶  $C = 10$ ,  $\gamma = 0.05$ , kernel = poly
  - ▶ Slight improvement on precision and sensitivity
  - ▶ Two extra cases of TP but 6 more cases of FN
- ▶ Hyperparameters tuning does not improve the performance of our model

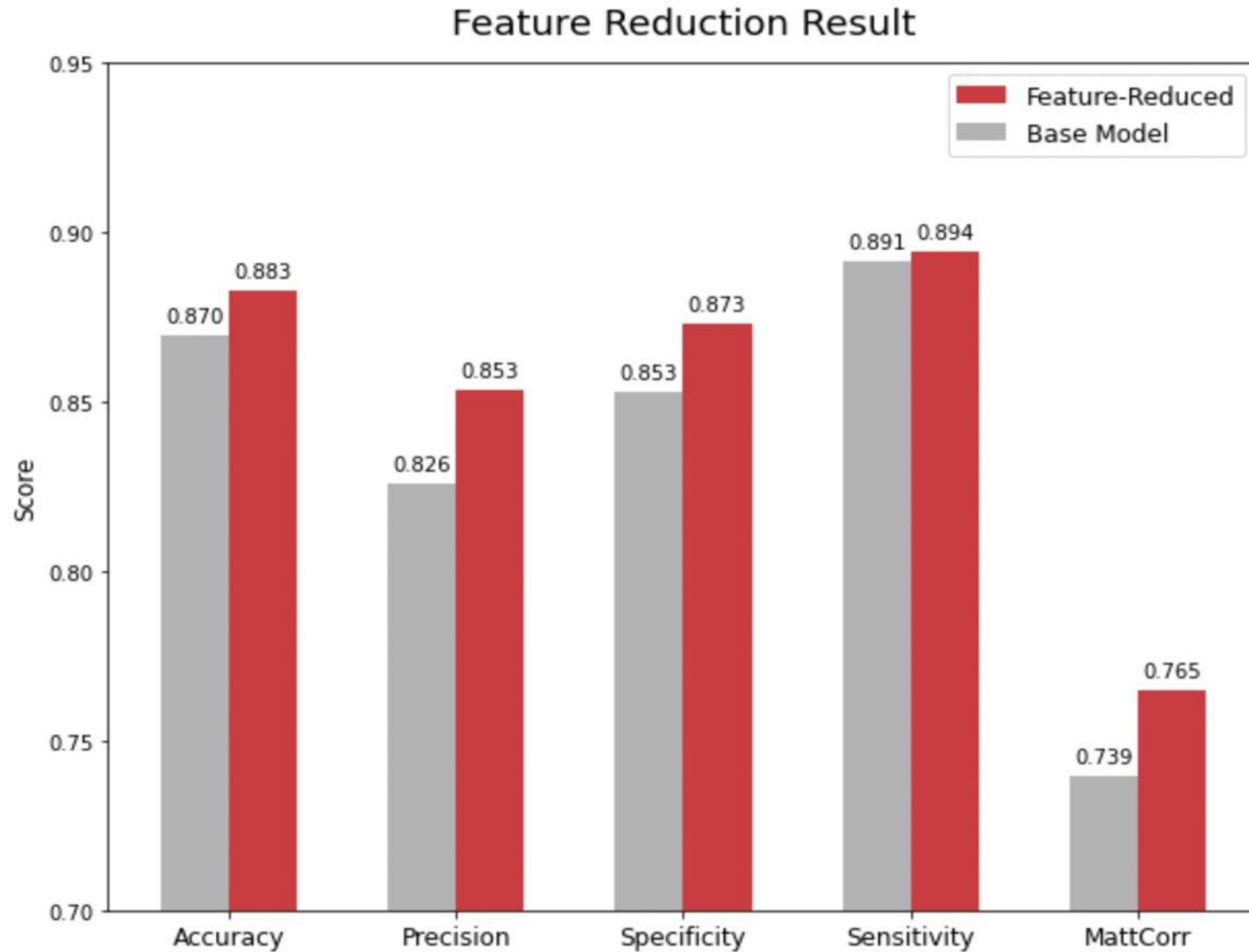
TP 92/94	FP 17/15
FN 12/18	TN 109/103

Model	Accuracy	Precision	Specificity	Sensitivity	MattCorr
SVC	0.8739	0.8440	0.8651	0.8846	0.7473
SVC_Tuned	0.8565	0.8624	0.8729	0.8393	0.7129

# SVC Tuning - Feature Reduction



# SVC Tuning - Feature Reduction



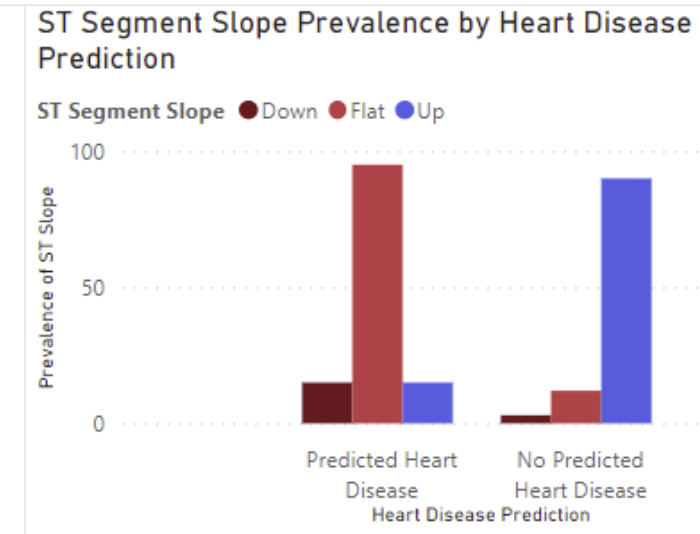
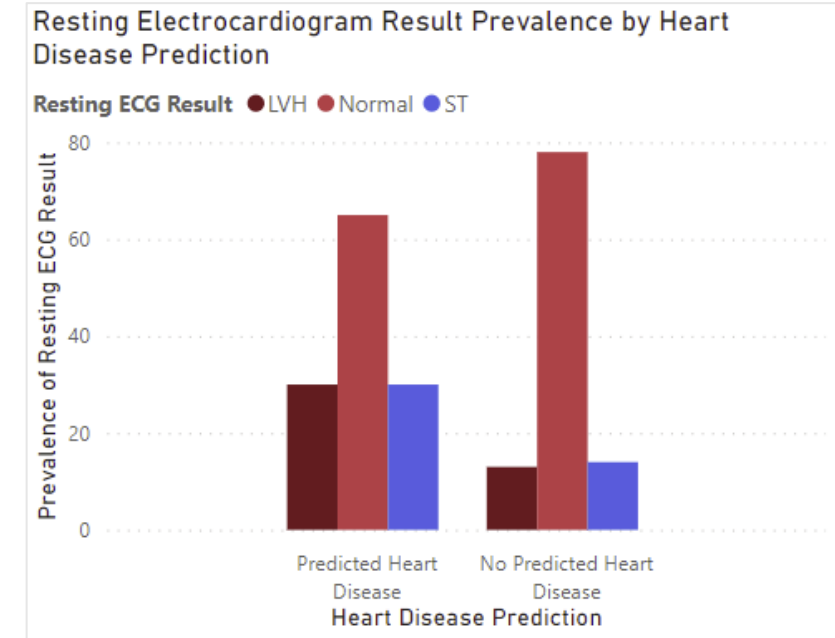
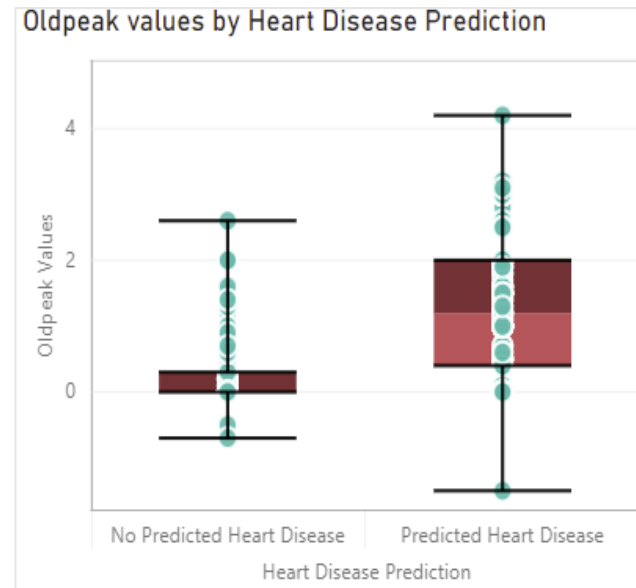
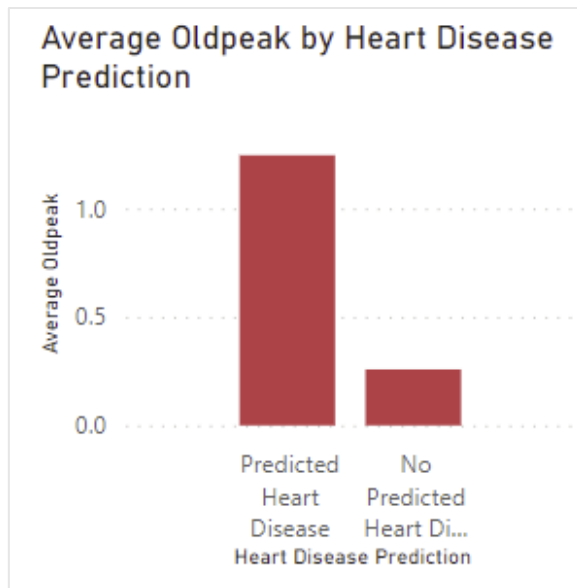
# Simulating Patient Data Overview

- ▶ SVC machine learning model trained from 75% of original dataset saved as model
- ▶ Loaded remaining 25% of original dataset not used for training from SQL database into PowerBI as a test dataset
- ▶ Applied SVC machine learning model to test dataset using PowerBI's python script capabilities
- ▶ Generated graphics based on the SVC prediction to better analyze what factors could be considered important for predicting heart disease

# Simulating Patient Data Visualizations

## ► Electrocardiogram Features

- Oldpeak
- Result
- ST Segment Slope

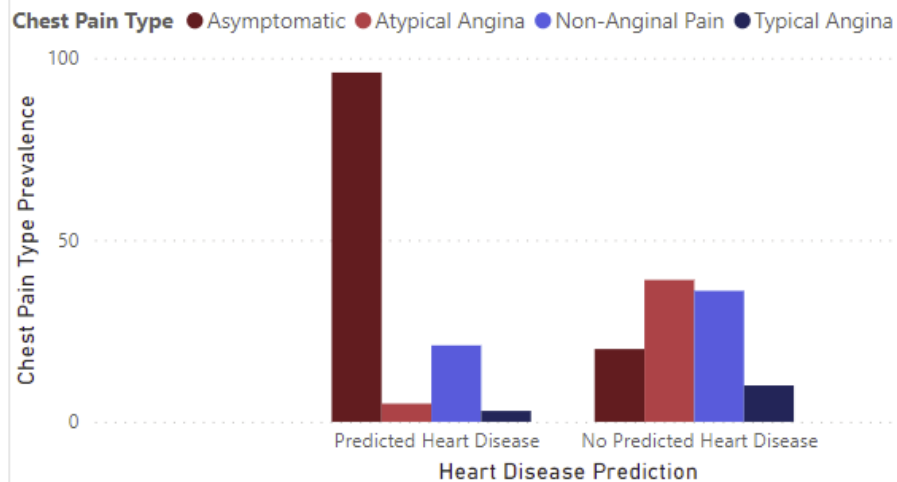


# Simulating Patient Data Visualizations

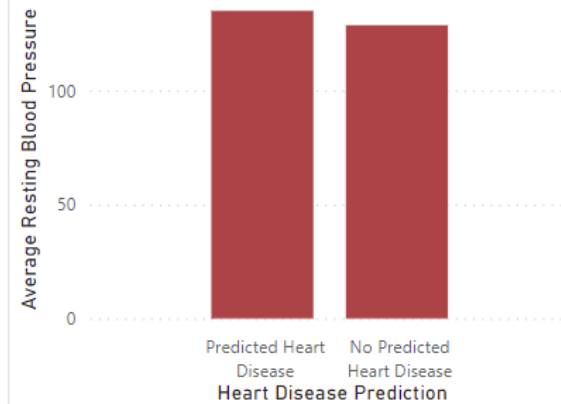
## ► Patient Vital Features

- Chest Pain
- Blood Pressure
- Heart Rate
- Blood Sugar
- Exercise-Induced Angina

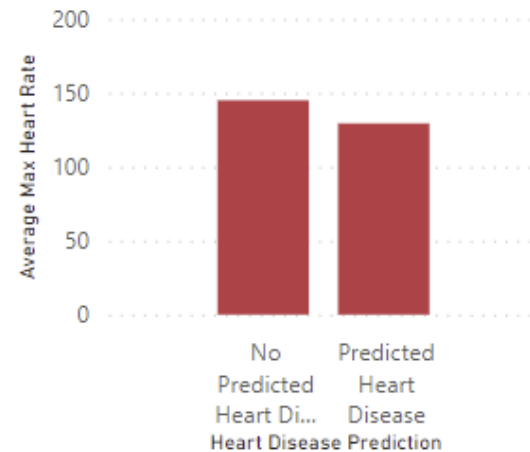
Chest Pain Type Prevalence by Heart Disease Prediction



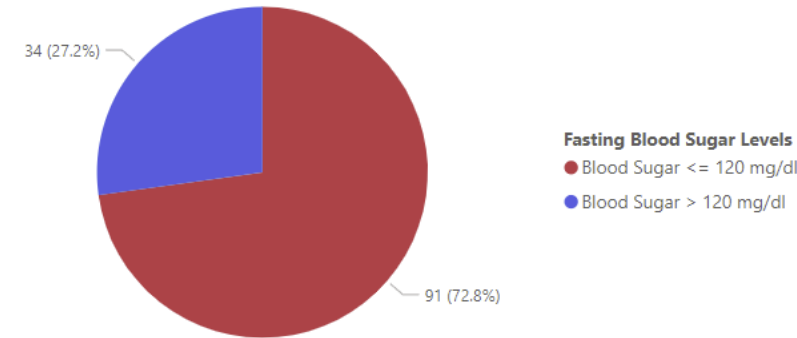
Average Resting Blood Pressure by Heart Disease Prediction



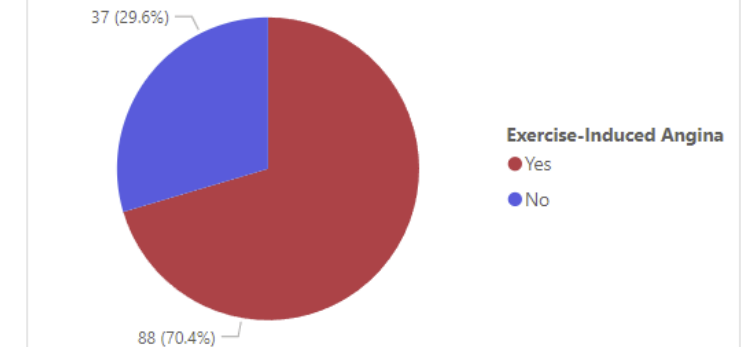
Average Max Heart Rate by Heart Disease Prediction



Predicted Heart Disease by Fasting Blood Sugar Levels



Predicted Heart Disease by Exercise-Induced Angina

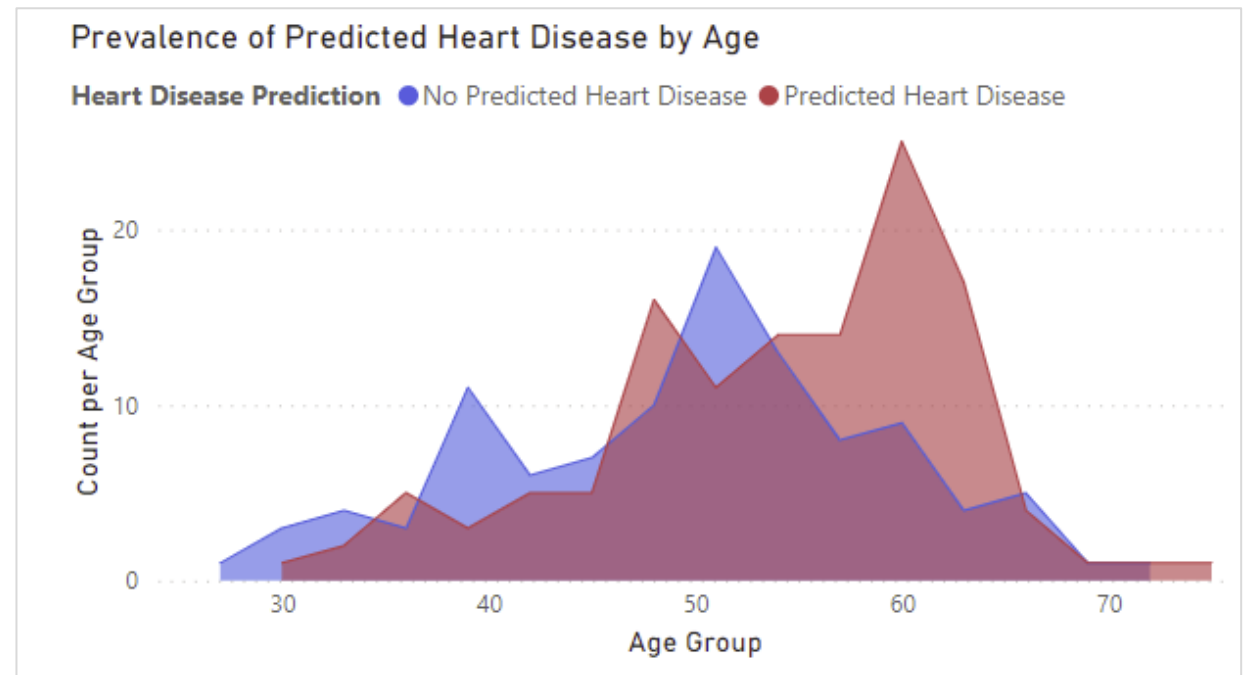
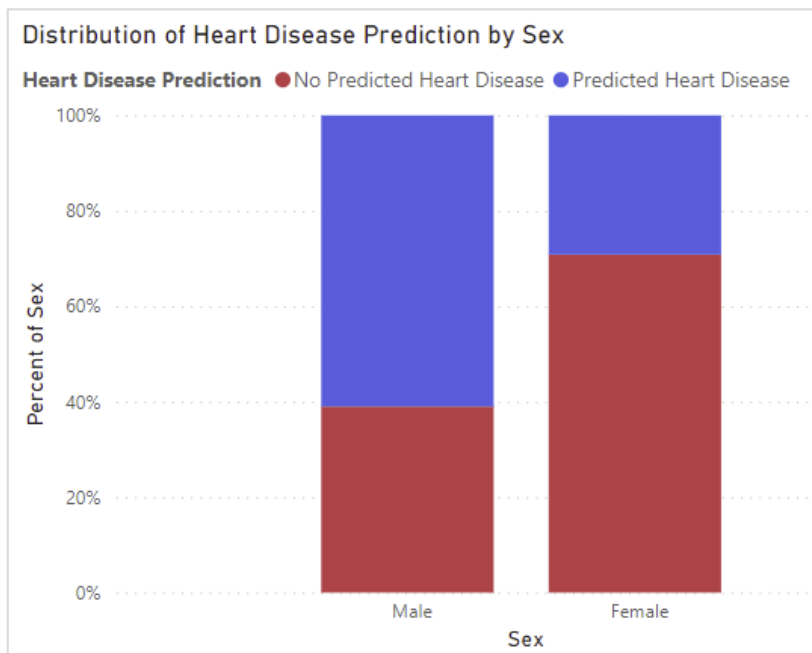


# Simulating Patient Data Visualizations

## ► Demographic Features

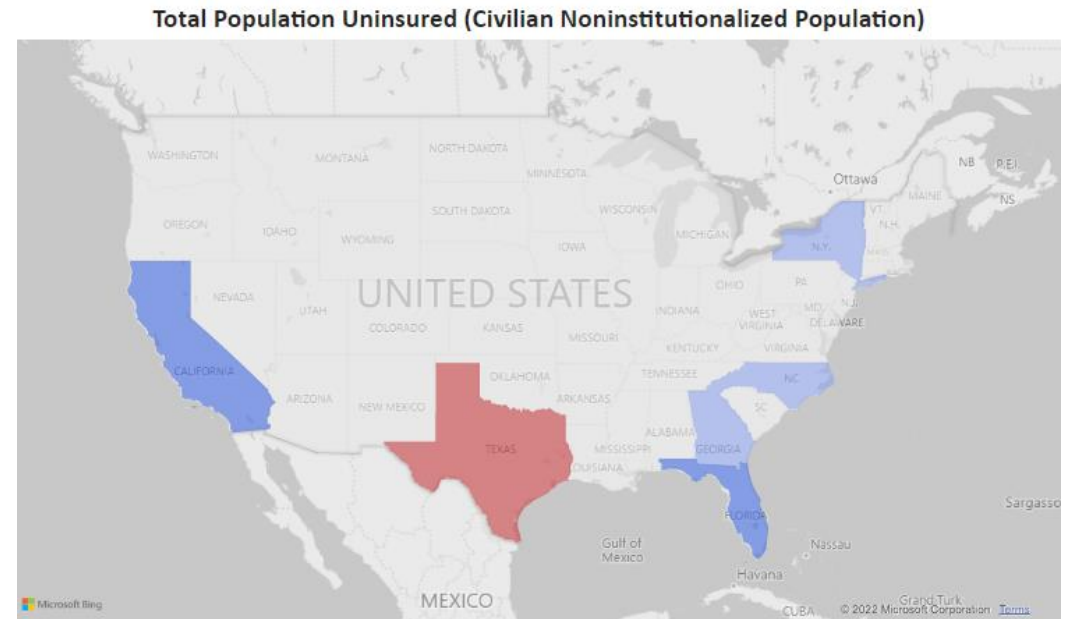
► Sex

► Age



# Exploration: State Level

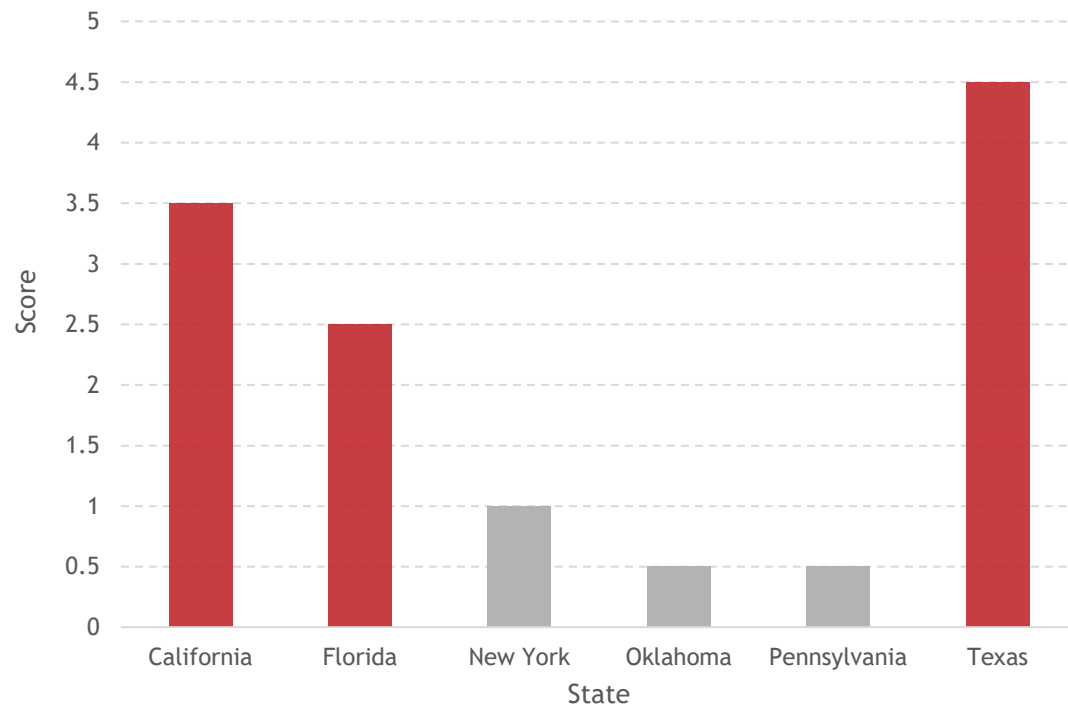
- ▶ Economic & Demographic Factors:
  - ▶ Total Population Below 100 Percent of the Poverty Threshold
  - ▶ Total Population Uninsured (CNP)
  - ▶ Total Population of Males
  - ▶ Total Population of People Aged 65 Years and Older
  - ▶ Percent Uninsured Male
  - ▶ Percent Uninsured 64 Years and Older
- ▶ Behavioral Risk Factors:
  - ▶ Prevalence of Current Smoking Among US Adults (18+)
  - ▶ Prevalence of Obesity Among US Adults (18+)
  - ▶ Prevalence of Physical Inactivity Among US Adults (18+)
  - ▶ Prevalence of Major Cardiovascular Disease Among US Adults (18+)



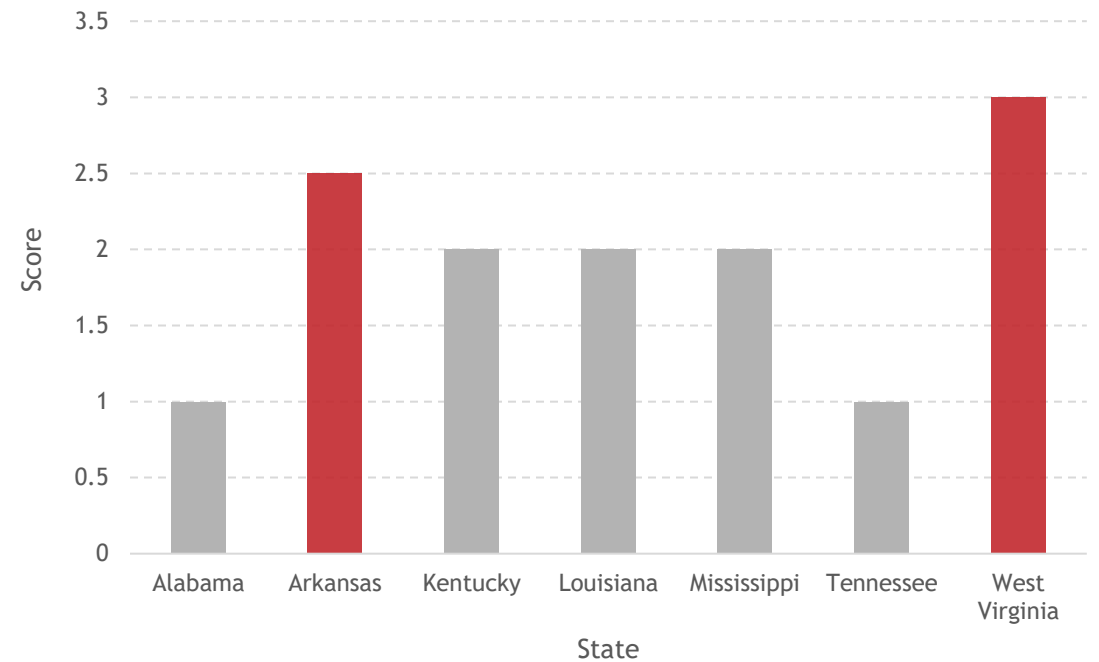


# Exploration: State Level

States with High Populations of High Risk or High Need Economic and Demographic Groups

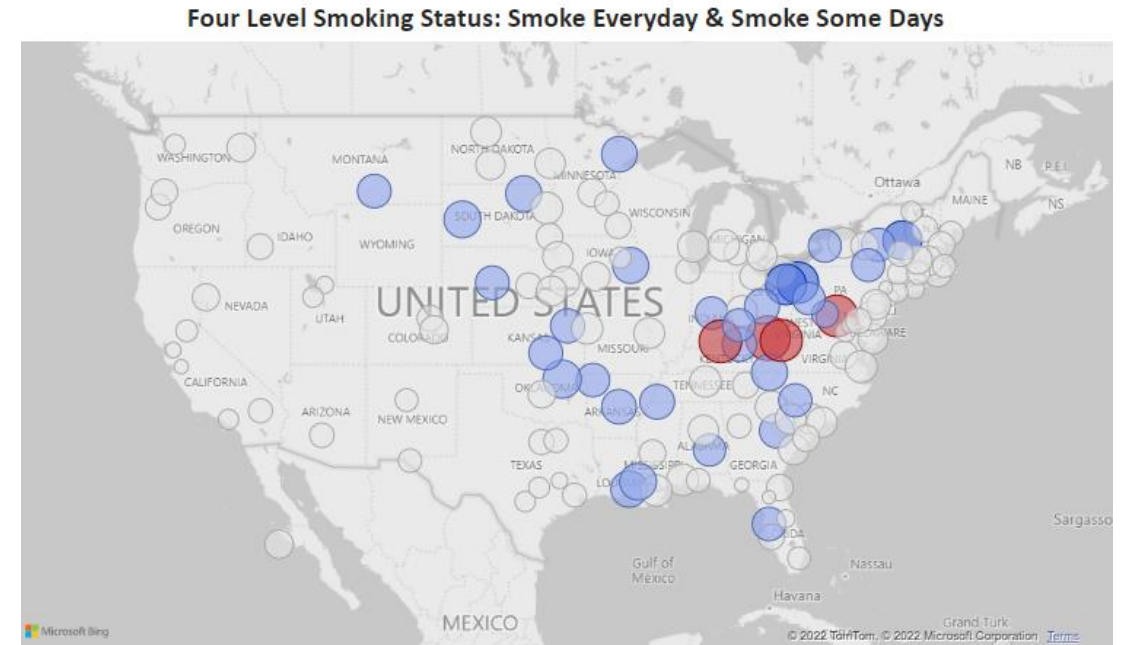


States with High Proportion of Behavioral Risk Factors Identified as High Risk for Heart Disease



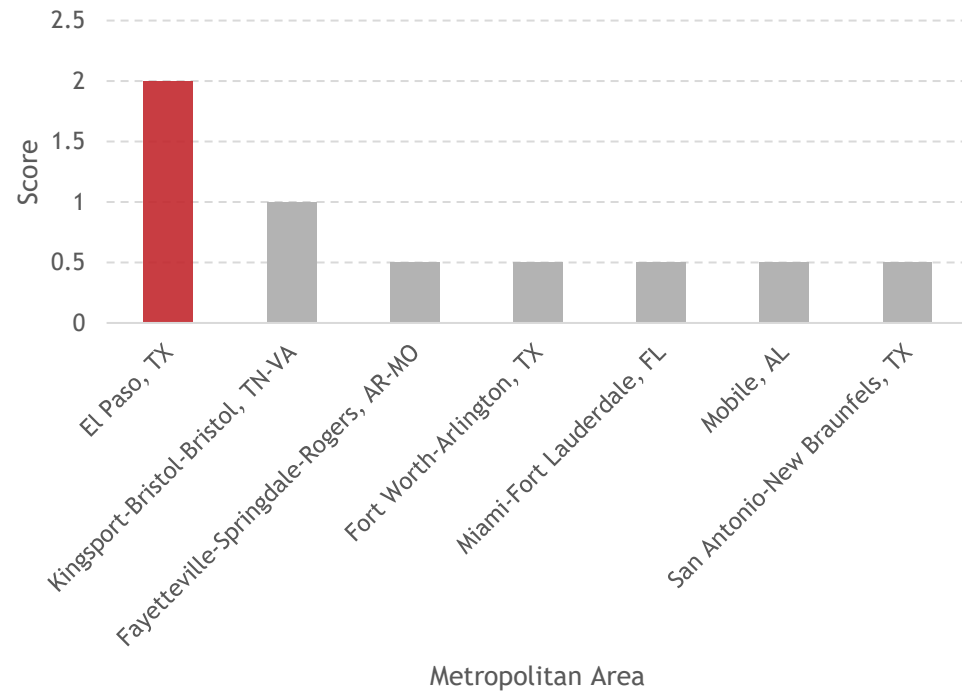
# Exploration: Metropolitan Level

- ▶ Economic Factors:
  - ▶ Adults Aged 18-64 Who Have No Health Care Coverage
  - ▶ Answered YES: Was There a Time in the Past 12 Months When You Needed to See a Doctor But Could Not Because of Cost?
- ▶ Behavioral Risk Factors:
  - ▶ Weight Classification by Body Mass Index (BMI): Obese (BMI 30.0-99.8)
  - ▶ Four Level Smoking Status: Smoke Everyday & Smoke Some Days
  - ▶ Respondents that Have Ever Reported Having Coronary Heart Disease (CHD) or Myocardial Infarction (MI)

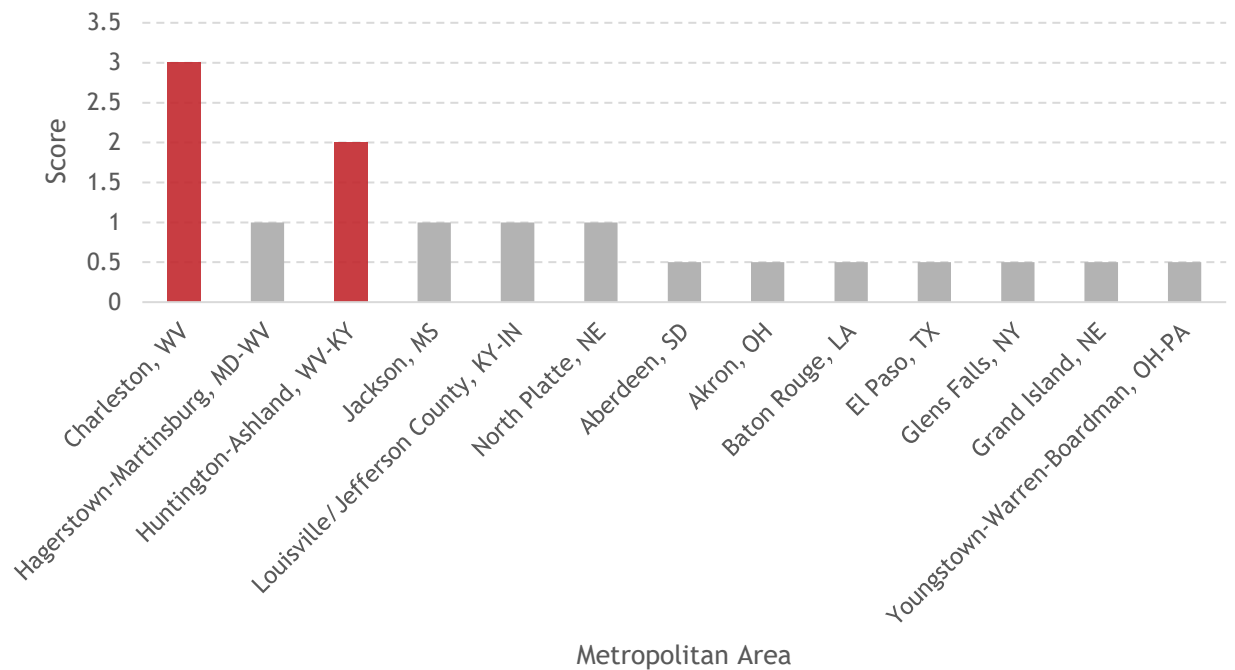


# Exploration: Metropolitan Level

Metropolitan Areas with High Proportion of Economic Factors Used to Identify High Need Populations



Metropolitan Areas with High Proportion of Behavioral Risk Factors Identified as High Risk of Heart Disease



# Recommendations

► 
$$\text{Metric} = \frac{\text{Score} \times \text{Metropolitan Area's Population}}{1,000,000}$$

Ranking	Metropolitan Area	Metric
1	Fort Worth-Arlington, TX	3.85
2	Miami-Fort Lauderdale, FL	3.00
3	El Paso, TX	2.17
4	San Antonio-New Braunfels, TX	1.30
5	Louisville/Jefferson County, KY-IN	1.26
6	Huntington-Ashland, WV-KY	0.72
7	Charleston, WV	0.62
8	Jackson, MS	0.59
9	Baton Rouge, LA	0.44
10	Akron, OH	0.35

# Conclusion

## Key Points:

- ▶ Our ML Model and clinics relieve pressure on doctors, specifically in underserved communities

## ML Model:

- ▶ Final model achieved a high accuracy of 88%

## Simulating Patient Data:

- ▶ Heart disease predictions show that Age and Sex are strong indicators, confirming our research results

## Clinic Locations:

- ▶ Locations in Texas have the greatest commonality between all needs, risk factors, and population requirements
- ▶ Considerations could be made for lower population, higher risk metropolitan areas

## Next Steps:

- ▶ ML model with real patient data
- ▶ Investigate hospital wait times to determine areas for clinics