# COMPARING MACHINE LEARNING ALGORITHM TO PREDICT HEART ATTACKS

HALÉ KPETIGO / HALE.KPETIGO@GMAIL.COM

NIH / FAES – 509 / SPRING 2021

03/19/2021

# CAN HEART ATTACKS BE PREDICTED?

- Heart diseases is the leading cause of death in the US.
- Condition of a heart disease is often silent and not diagnosed
- This leads to heart attacks, arrhythmia or heart failure.
- If we can diagnose at risk population from safe populations we could save lives

# HOW CAN WE PREDICT?

- Identify risk factor for heart attack
- Identify dataset with risk factor information
- Use unsupervised machine learning algorithm to classify the populations
- Compare the different supervised approaches. What information to they provide?
- Use a supervised approach to predict condition
- Measure and compare each supervised approach

# RISK FACTORS

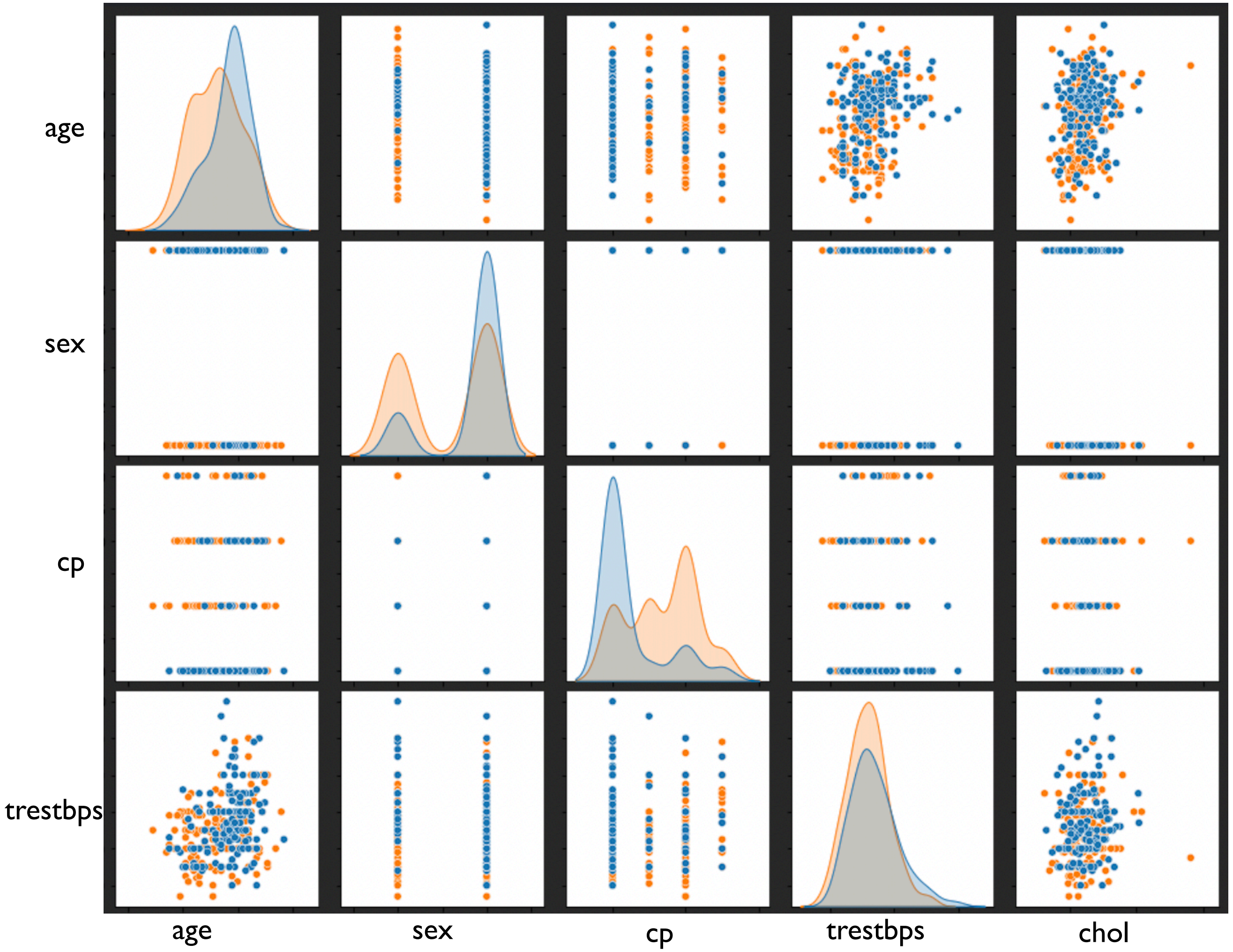
- High blood pressure
- High cholesterol
- Smoking
- Diabetes
- Obesity
- Unhealthy diet
- Physical inactivity
- Excessive use of alcohol

# IDENTIFIED DATASET

| Features | Description |
|---|---|
| Age | Age of patient in years |
| Sex | Sex of patient (1 = male; 0 = female) |
| cp | Chest pain type (4 values) |
| trestbps | Resting blood pressure |
| chol | Serum cholesterol in mg/dl |
| fbs | Fasting blood sugar > 120 mg/dl |
| restecg | Resting electrocardiographic results (values 0,1,2) |
| thalach | Maximum heart rate achieved |
| exang | Exercise induced angina |
| oldpeak | Oldpeak = ST depression induced by exercise relative to rest (ECG) |
| slope | The slope of the peak exercise ST segment (ECG) |
| ca | Number of major vessels (0-3) colored by flourosopy |
| thal | Thalassemia: 0 = normal; 1 = fixed defect; 2 = reversible defect |
| target | 0= less chance of heart attack 1= more chance of heart attack |

- About data source: This database contains 76 attributes, but all published experiments refer to using a subset of 14 of them. Ref: https://archive.ics.uci.edu/ml/datasets/heart+disease
- Dataset used – Cleveland only:
  - https://www.kaggle.com/johnsmith88/heart-disease-dataset
- Similar dataset – All:
  - https://www.kaggle.com/nareshbhat/health-care-data-set-on-heart-attack-possibility

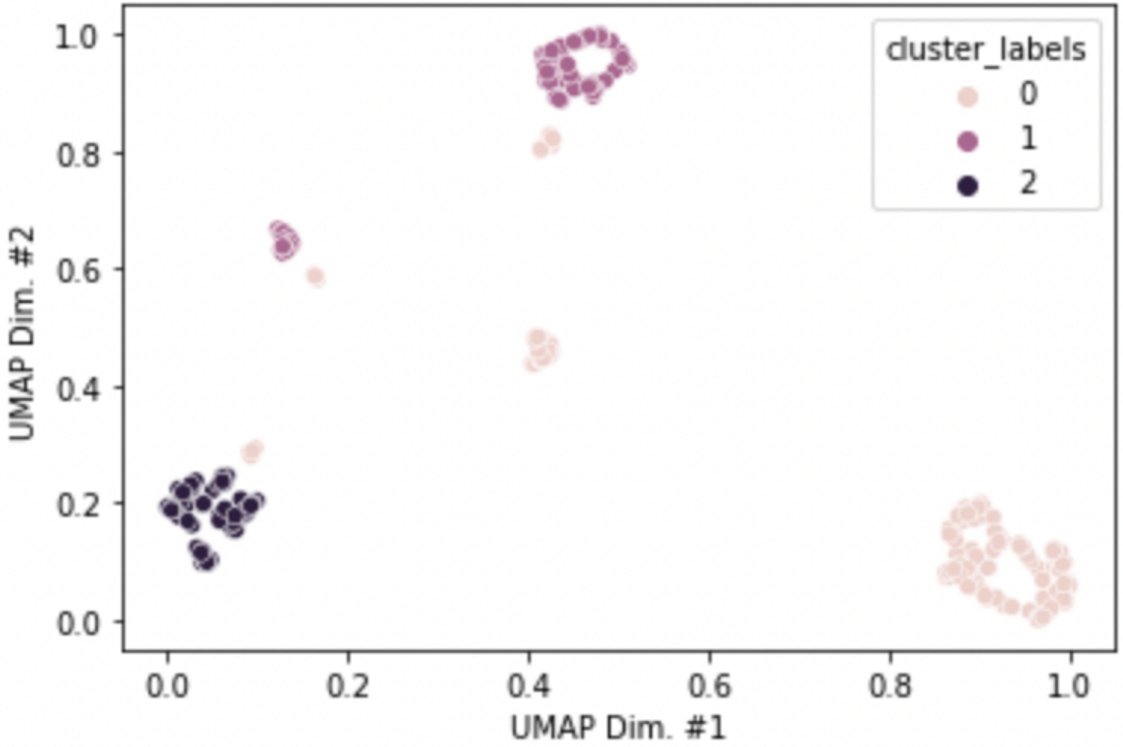
# DATA ANALYSIS – PAIR PLOT

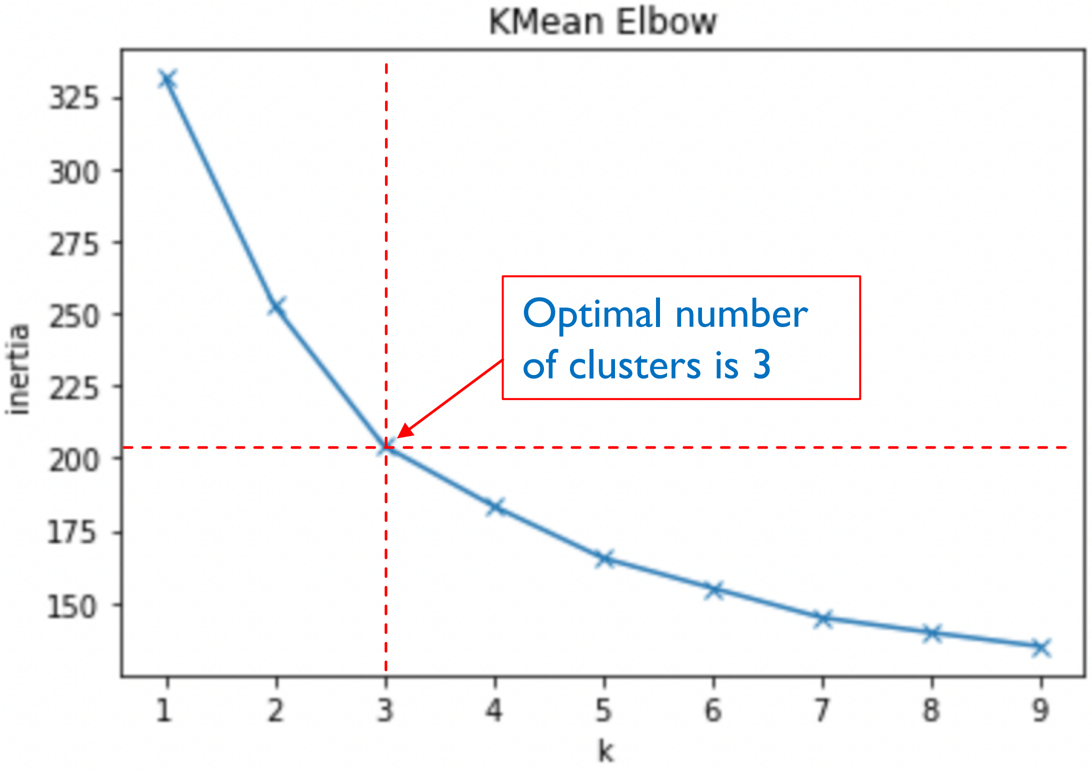


- Legend:
  - **Blue**: less chance of heart attack
  - **Orange**: more chance of heart attack
- Pair plot shows relationship between the features of the dataset.
- We can observe that pairs of the features that have a non continuous value (i.e. sex or cp) the clusters are not globular but instead linear. Each line representing the male or female for sex or each level of chest pain for cp.
- From the diagonal distribution plot, we can observe that at risk population has a shorter life expectancy.
- We can also observe that our dataset has less women than men, but of the studied women population, women are more at risk compared to the men within their studied population.

# UNSUPERVISED APPROACH – K-MEANS

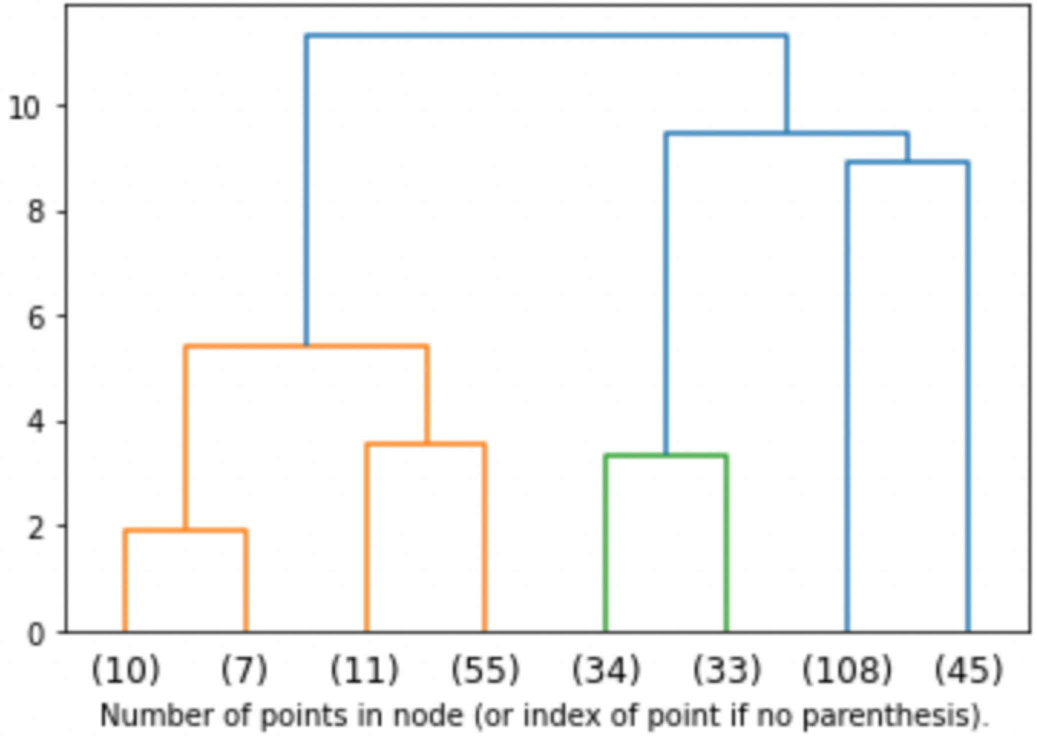
- K-Means helped identify 3 clusters
- Elbow confirms that the optimal number of cluster is 3

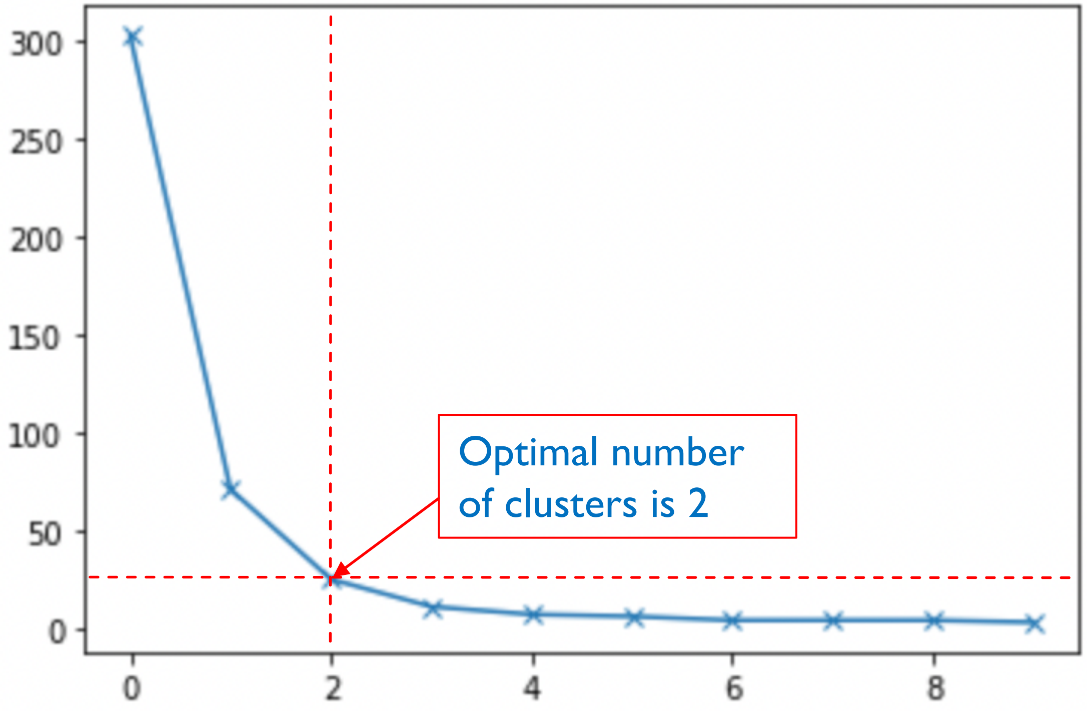




Optimal number of clusters is 3

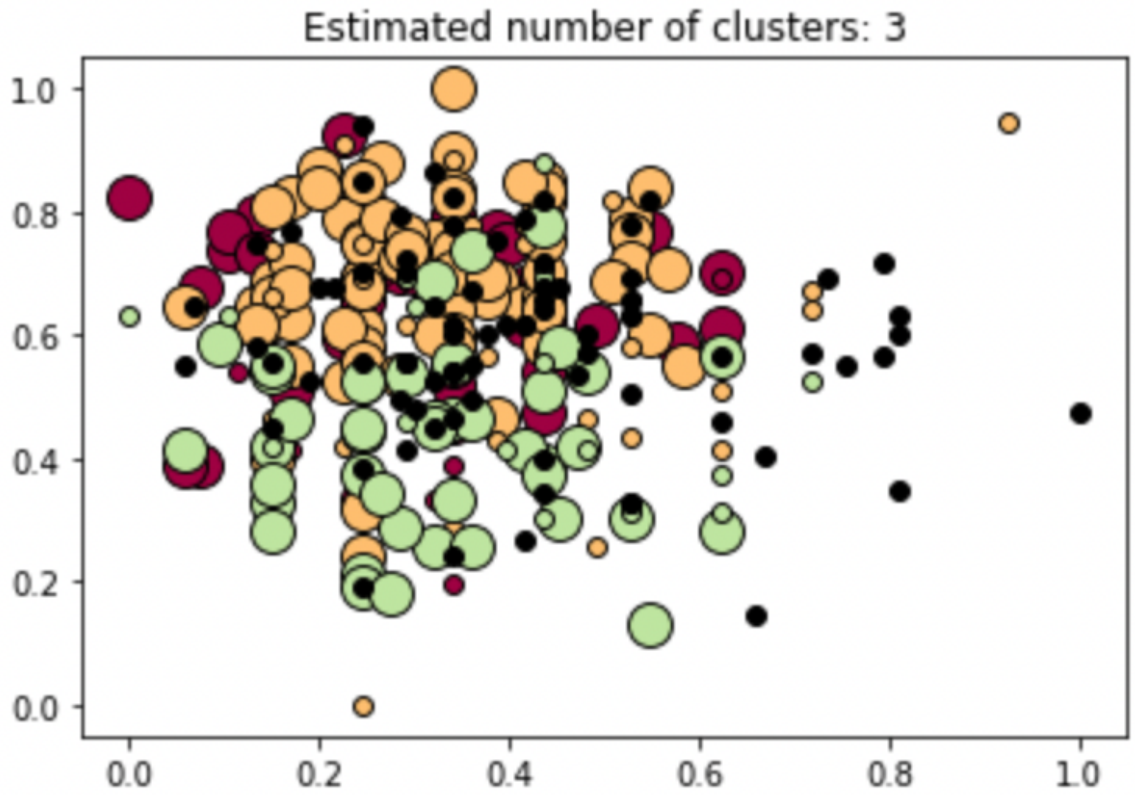
# UNSUPERVISED APPROACH – HIERARCHICAL CLUSTERING





- Using the optimal number of clusters recommended by Elbow (2) we notice a dominant population with 108 record. It would be interesting to analyze that population to see the particularities of the features present in that population as opposed to the other groups.

# UNSUPERVISED APPROACH – DENSITY CLUSTERING



- Number of clusters: 3
- Using all features:
  - Max accuracy 0.5478547854785478 with model DBSCAN(eps=1.1, min_samples=1)
- Removing sex and age:
  - Max accuracy 0.6039603960396039 with model DBSCAN(eps=0.8, min_samples=6)
- Accuracy is slightly improved by ignoring sex and age. Those two features may have been overfitting the model

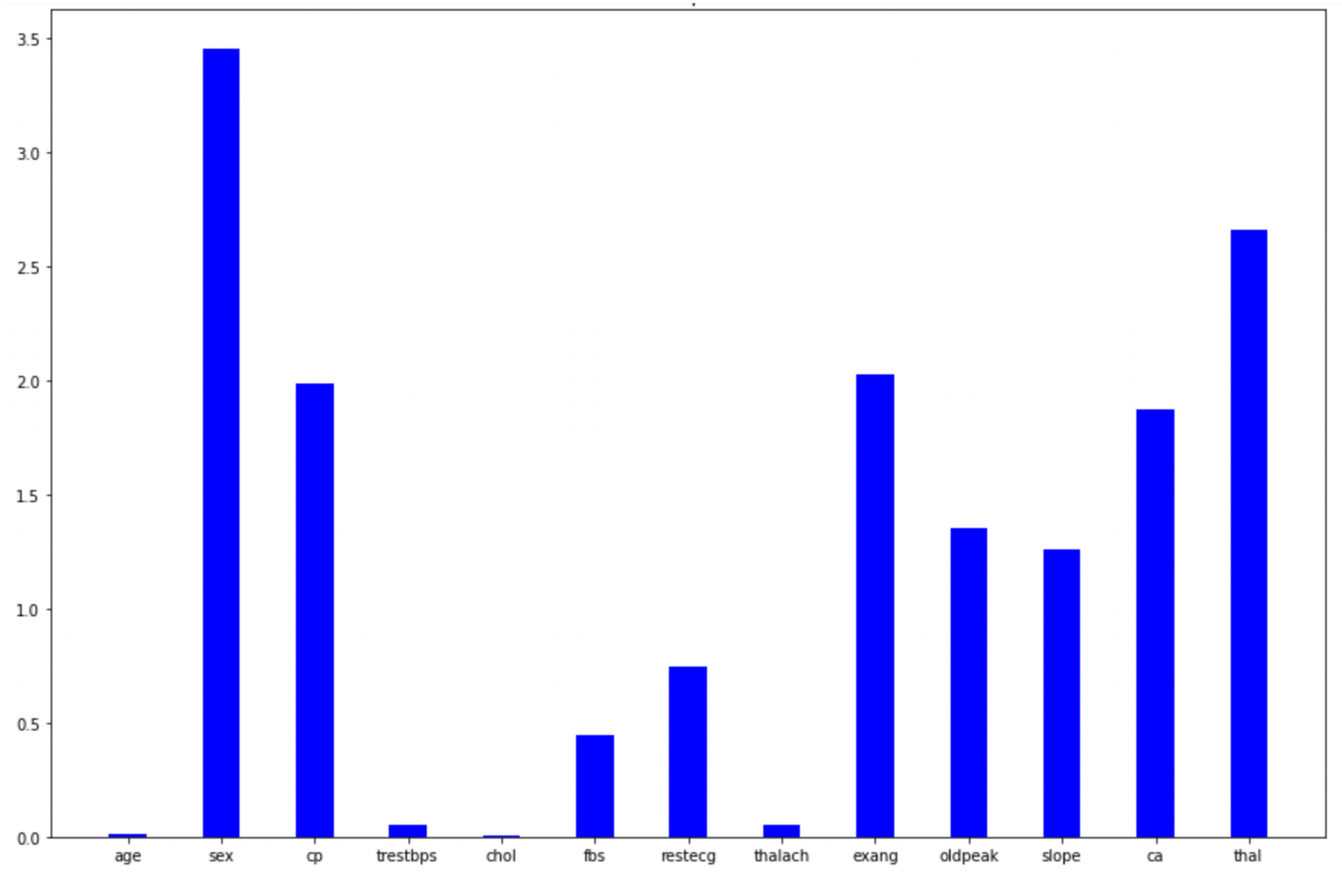
# COMPARING UNSUPERVISED APPROACH

| Algorithm | # Clusters |
|---|---|
| K-MEANS | 3 |
| Hierarchical Clustering | 2 |
| Density Clustering | 3 |

- K-MEANS and Density Clustering identified 3 clusters when in fact we should be expecting 2 clusters
- It would be interesting to analyze the data to have better insight in the similarities that exists within each population
- I would also be interested in further analyzing these algorithms with a larger set of data to see the impact on the number of clusters identified

# SUPERVISED - SVM

- PREDICTION ACCURACY 0.8150273224043716



# SUPERVISED – DECISION TREE

- Decision Tree Dataset size: 303
- Accuracy Score on train data: 1.0
- Accuracy Score on test data: 0.77
- Accuracy Score on train data: 0.7635467980295566
- Accuracy Score on the test data: 0.75

- Dataset size: 1025
- Accuracy Score on train data: 1.0
- Accuracy Score on test data: 0.9823008849557522
- Accuracy Score on train data: 0.8833819241982507
- Accuracy Score on the test data: 0.8584070796460177

# SUPERVISED – GAUSSIAN NAIVE BAYES

- Gaussian NB
- Dataset size: 303
- Accuracy Score on train data: 0.8522167487684729
- Accuracy Score on test data: 0.8

- Gaussian NB
- Dataset size: 1025
- Accuracy Score on train data: 0.8338192419825073
- Accuracy Score on test data: 0.8141592920353983

# COMPARING SUPERVISED APPROACH

| Algorithm | Accuracy (303) | Accuracy (1025) |
|---|---|---|
| SVM | 81% | 84% |
| Decision Tree | 75% | 85% |
| Bayesian – Gaussian Naïve Bayes | 80% | 81% |

| Algorithm | Accuracy (303) on (1025) trained |
|---|---|
| SVM | 83% |
| Decision Tree | 75% |
| Bayesian – Gaussian Naïve Bayes | 80% |

- The performance of the 3 algorithms are comparable ranging from 75% to 85%.
- SVM performed better on a larger data set
- We have also observed that SVM performance improved on the smaller dataset after being trained on a larger training set
- Decision Tree offered the most accuracy on a larger dataset, and performed the poorest on a smaller dataset. This may be due to overfitting.
- Gaussian Naïve Bayes remained the most consistent across dataset sample size.

# REFERENCES

- Source code: https://github.com/halekpetigo/BIOF509
- Data Description: https://archive.ics.uci.edu/ml/datasets/heart+disease
- Data used: https://www.kaggle.com/johnsmith88/heart-disease-dataset
- Data used: https://www.kaggle.com/nareshbhat/health-care-data-set-on-heart-attack-possibility

# QUESTIONS?

- Thank you!