03.11.2025

# Clustering Regression for Heart Disease Prediction

## Machine Learning I
## Final Project


Team 6
Zhiwei Guo, Amy Kim, Samuel Martinez Koss, Alvin Yao

1. Research Objective
2. Data Description
3. Modeling Methodology
4. Results & Conclusion

1. Research Objective
2. Data Description
3. Modeling Methodology
4. Results & Conclusion

# Background

## Heart Disease Affects Us All

- In 2022, 702,880 people died from heart disease. That's the equivalent of **1 in every 5 deaths**. (Source)
- Heart disease remains the **leading cause of death** for men, women, and people of most racial and ethnic groups. (Source)
- Heart disease **cost about $252.2 billion** from 2019 to 2020. This includes the cost of healthcare services, medicines, and lost productivity due to death. (Source)

## Early Detection = Saving Lives & Saving Costs

- Early detection methods are not only cost-effective, but also reduces medical costs for treatment as opposed to no detection. (Source)

Our goal is to:

**Produce interpretable profiles to help physicians identify high-risk individuals for further monitoring and detection.**

1. Research Objective
2. Data Description
3. Modeling Methodology
4. Results & Conclusion

# Heart Disease Dataset

Source: University of California Irvine Machine Learning Repository

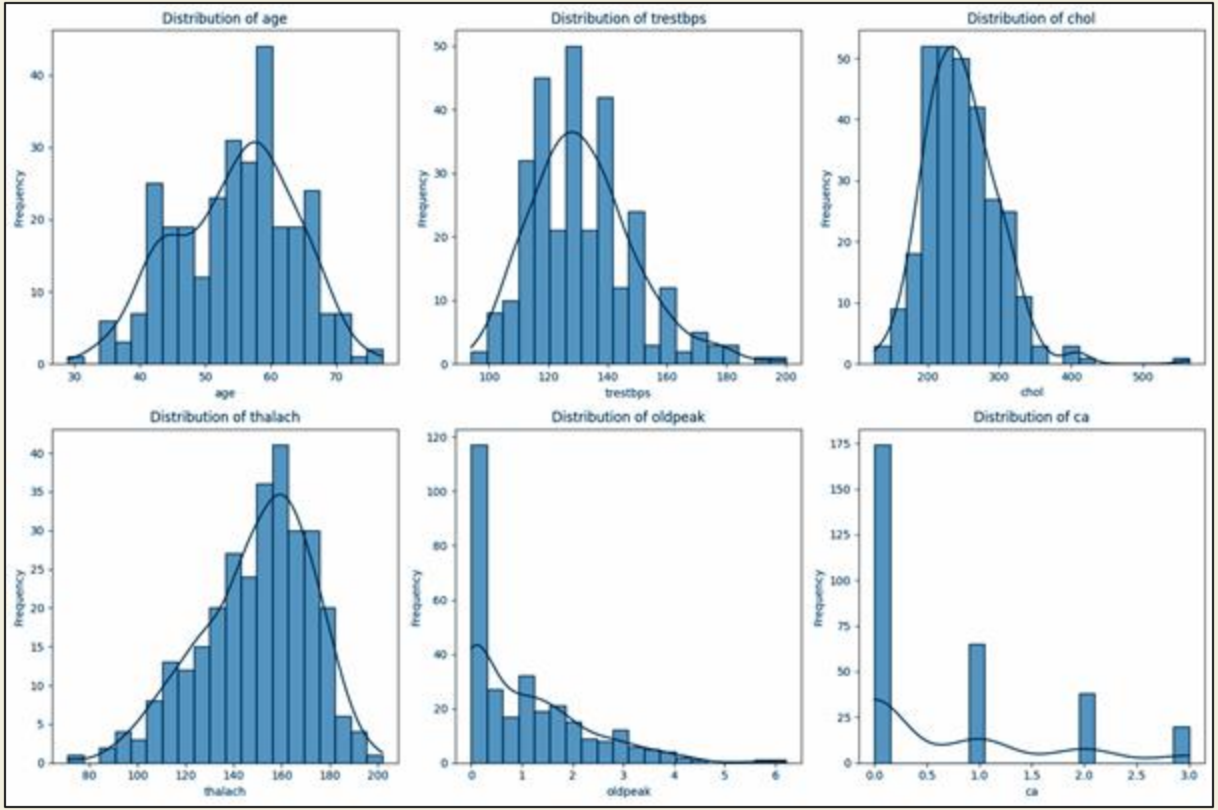
13 Primary Features, 1 Target Variable:

- **age**: Age of patient in years
- **sex**: Sex of patient in years
- **cp**: chest pain type
- **trestbps**: resting blood pressure (numeric
- **chol**: serum cholesterol in mg/dl (numeric
- **fbs**: if fasting blood sugar > 120 mg/dl (binary)
- **restecg**: resting electrocardiographic results (3 attribu
- **thalach**: maximum heart rate achieved (numeric
- **exang**: exercise induced angina (binary)
- **oldpeak**: ST depression induced by exercise relative to rest (numeric)
- **slope**: slope of the peak exercise ST segment (3 attribu
- **ca**: number of major vessels colored by fluoroscopy (numeric)
- **thal**: thallium stress test

- **num**: angiographic disease status (target v

*76 Attributes*

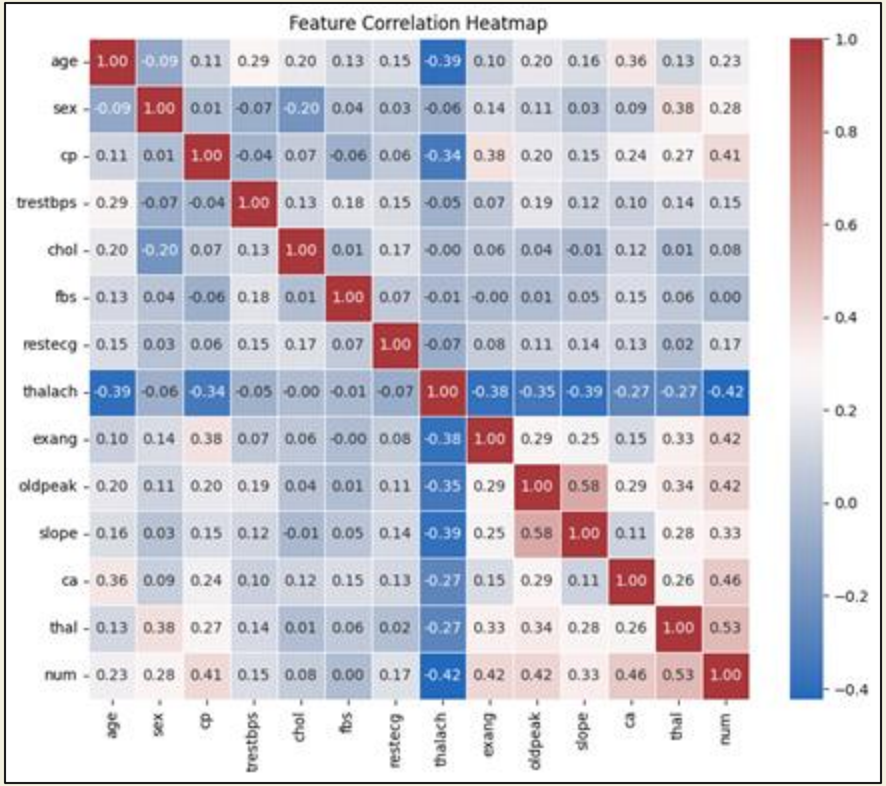
# Numeric Feature Distributions




- Ranges of features vary on orders of magnitude implying **need to scale**
- Some features appear to be **non-normally distributed** (e.g. **oldpeak** exhibits heavy right-skew)
- The distribution of **ca** is integer-valued with few distinct values – it may make more sense to **treat it as categorical**

# Correlation Analysis

- Strong potential linear predictors of **num**: **cp, thalach, exang, oldpeak, ca, thal**
- Concerns of potential multicollinearity in the data, particularly based on the correlations between **thalach** and other candidate predictors
- Some weakly correlated features such as **chol** and **fbs** might not be suitable for predicting heart disease on their own, but could yet interact with other variables or exhibit non-linear relationships with the target variable

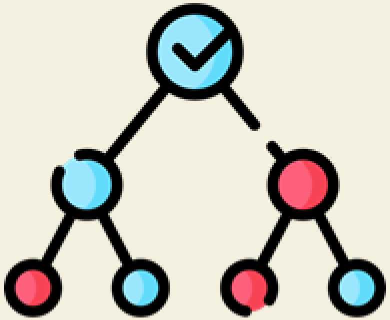


1. Research Objective
2. Data Description
3. **Modeling Methodology**
4. Results & Conclusion

# Methodology

## Machine Learning Feature Selection

Using **Classification Trees**, we can derive the most important features for us to base our clustering methods on.



## Clustering Model Selection

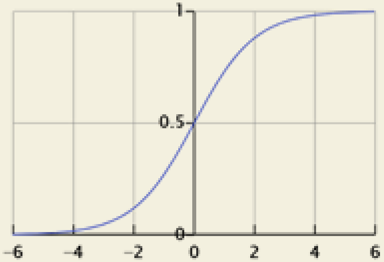
We trained multiple clustering models and compared to find the best clusters to cast on our regression.

We looked at three different methods:

**Latent Class Analysis**
**Gaussian Mixture**
**K-Prototype Clustering**

## Within-Cluster Prediction

With our selected clustering method and its derived clusters, we can train a predictive model to help us determine what the probability of heart disease given a patient's cluster.



# Feature Selection

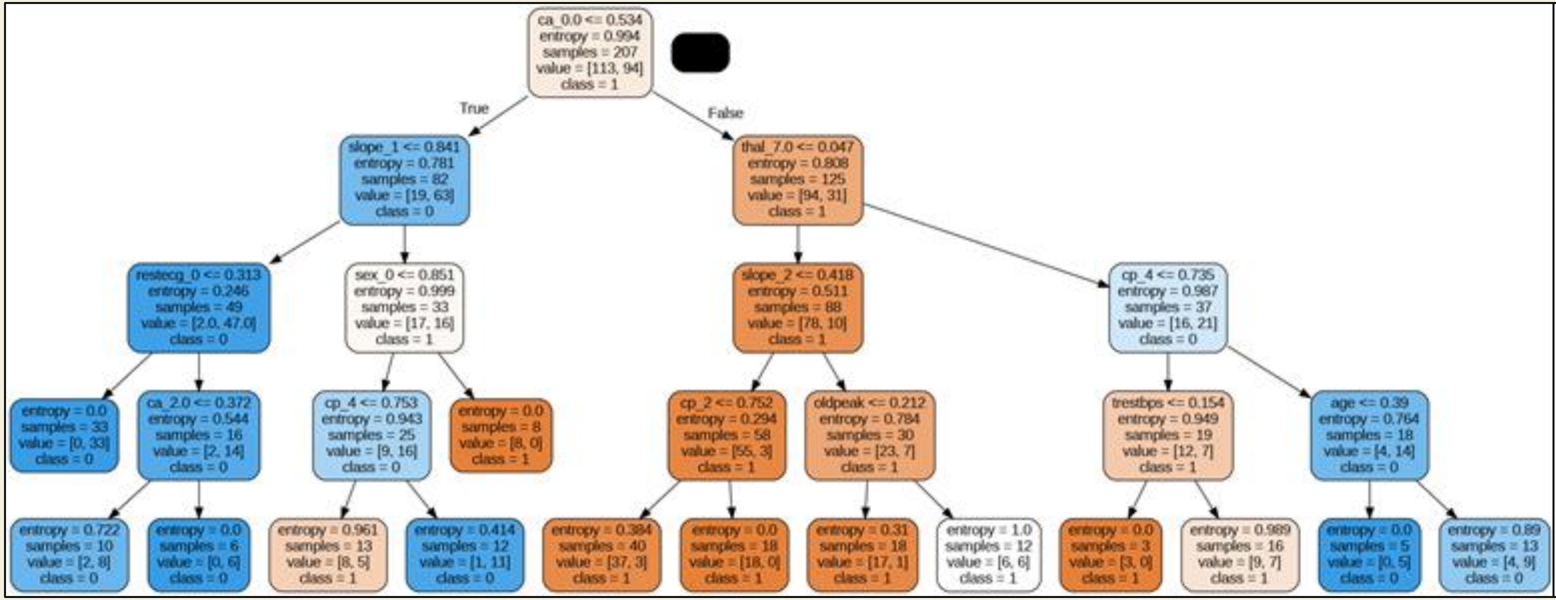
## Why Feature Selection?

- Increase model interpretability as well as application generalizability by limiting the number of attributes a physician needs to analyze
- Reduces computational cost as well as noise in the data by focusing only on features which contribute meaningful information
- With the dummy-coding of many categorical variables, reduced feature set helps avoid the curse of dimensionality

## Why Classification Tree?

- Captures non-linear relationships between features and the target variable, as well as naturally accounting for feature interactions
- Addresses the multicollinearity concerns from the correlation analysis by choosing additional features based on marginal information gain
- Clearly interpretable as well as less computationally expensive compared to wrapper methods

# Feature Selection via Classification Tree




| | Feature | Importance |
|---|---|---|
| 1 | ca_0.0 | 0.3371 |
| 2 | thal_7.0 | 0.1620 |
| 3 | slope_1 | 0.1574 |
| 4 | cp_4 | 0.0896 |
| 5 | sex_0 | 0.0780 |
| 6 | oldpeak | 0.0492 |
| 7 | slope_2 | 0.0365 |
| 8 | restecg_0 | 0.0278 |
| 9 | trestbps | 0.0184 |
| 10 | age | 0.0181 |
| 11 | cp_2 | 0.0138 |
| 12 | ca_2.0 | 0.0122 |

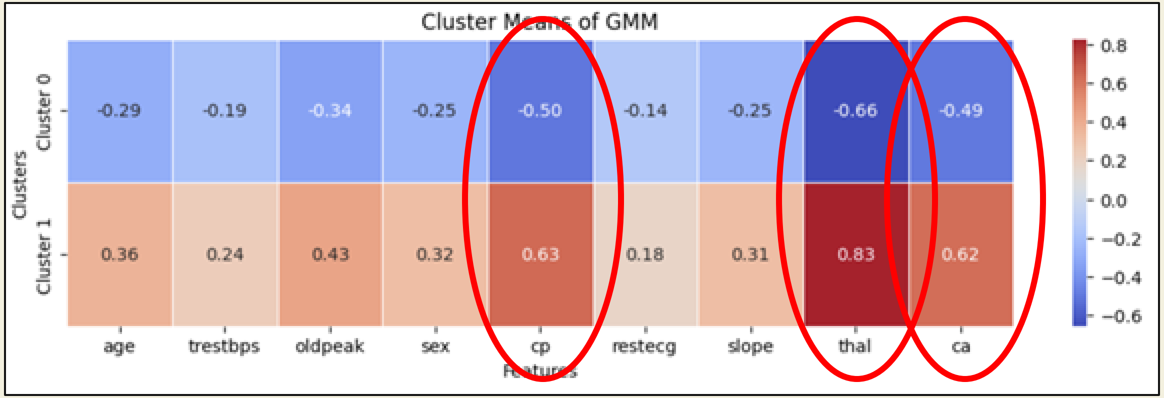
- **GridSearchCV** used to hyper-parameter tune this model
- Features with **non-zero importance** are selected for further model training
- Train Accuracy: 0.86, Test Accuracy: 0.86, Avg. Precision: 0.86, Avg. Recall: 0.86, Avg. F1: 0.86

# Clustering via Gaussian Mixture (1/2)

1. **Defined Number of Clusters** (Unsupervised Learning)

   **2 (Positive vs Negative for heart disease)**

1. **Cluster Means Analysis** (Strong differencing factors)
   a. **'Thal': Thallium Stress Test Result** (Indicates how well blood flows into your heart while exercising or at rest)
   b. **'cp':** Chest pain type
   c. **'ca':** Number of major vessels (0-3)

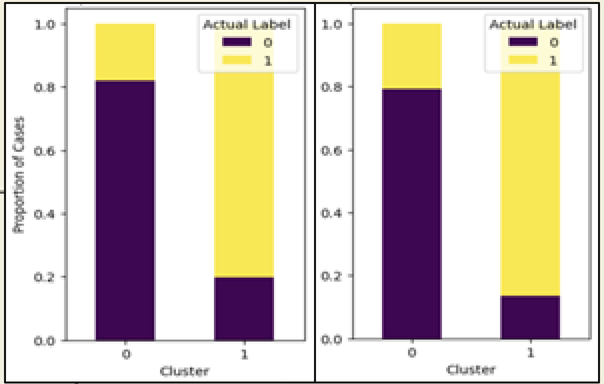



*** Chest pain type ('cp')**
➢1: Typical angina
➢2: Atypical angina
➢3: Non-anginal pain
➢4: Asymptomatic

# Clustering via Gaussian Mixture (2/2)

3. Mapping of cluster results to the actual data based on majority class
    a. **Cluster 0:** Negative on Heart Disease
    b. **Cluster 1:** Positive on Heart Disease

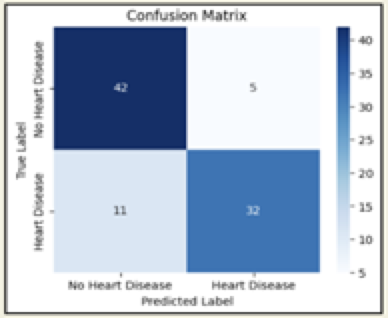
**Distribution of Actual Labels Across Clusters**




Training Data Testing Data

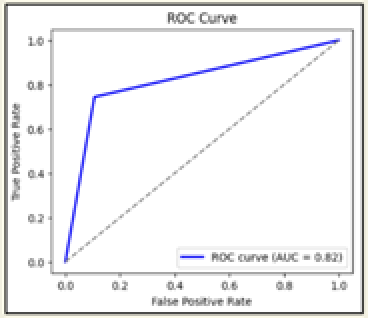
3. Evaluation:

**Precision & Recall:**




| | Precision | Recall | F1 Score |
|---|---|---|---|
| No Disease | 0.79 | 0.89 | 0.84 |
| Disease | 0.86 | 0.74 | 0.80 |

**ROC Curve (AUC = 0.82):**




Slight trade-off between detecting true cases and avoiding false positives, but demonstrates good discrimination.

**Overall Accuracy:**

- Train: 0.81
- Test: 0.82

# Clustering via LCA

Method:

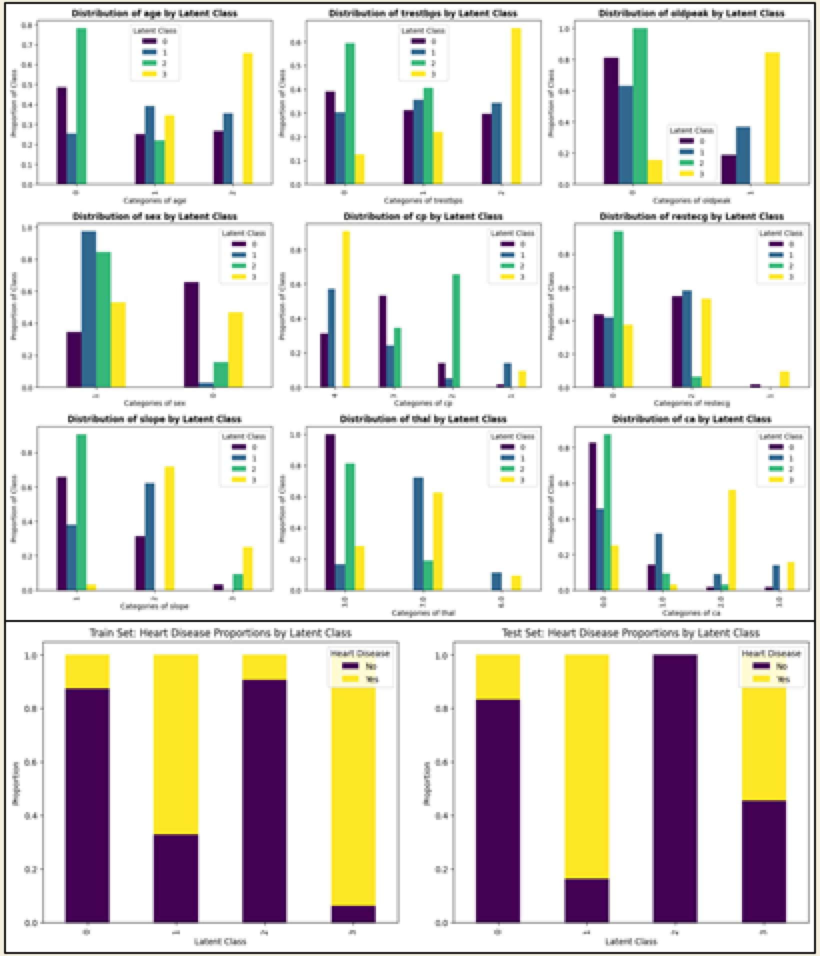
- Due to LCA requiring categorical variables, we processed the continuous variables into bins to “categorize” the variables
- By testing numerous class counts and measuring AIC and BIC, we found **4 classes to be the most optimal** using the elbow method

Cluster Profiles:

- Cluster 0 **(low risk)**: majority female, mostly normal heart conditions
- Cluster 1 **(high risk)**: mostly male, with some minimal to concerning heart conditions
- Cluster 2 **(low risk)**: mostly male, with minimal conditions other than some atypical cardiac chest pain
- Cluster 3 **(high risk)**: individuals with concerning heart conditions

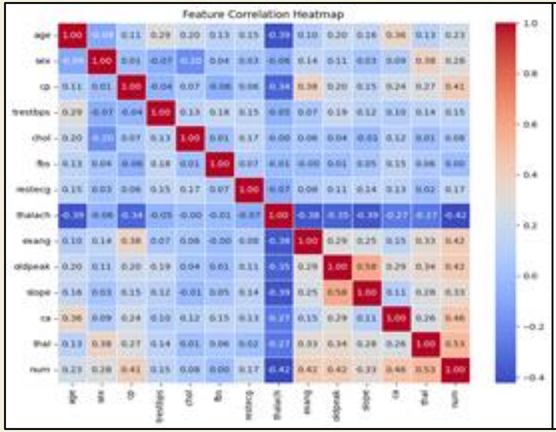
Performance Metrics:

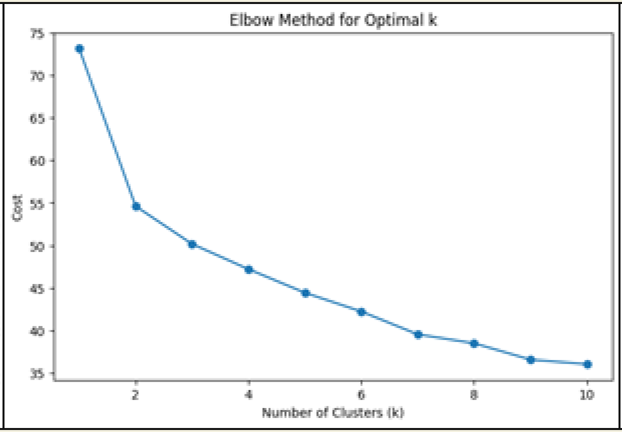
| | Precision | Recall | F1 Score |
|---|---|---|---|
| No Disease | 0.89 | 0.75 | 0.81 |
| Disease | 0.75 | 0.88 | 0.81 |



# Clustering via K-Prototype Clustering

1. Define the elbow value of 3
2. Fit the model and examine the clusters
    - K-prototype works well with both categorical and numerical variables
3. Assign majority class for prediction
4. Evaluation




```
Accuracy on Test Set: 0.8222

Confusion Matrix:
[[43  4]
 [12 31]]

Classification Report:
              precision    recall  f1-score   support

           0       0.78      0.91      0.84        47
           1       0.89      0.72      0.79        43

    accuracy                           0.82        90
   macro avg       0.83      0.82      0.82        90
weighted avg       0.83      0.82      0.82        90
```

# Within-Cluster Predictive Modeling

## Classification Tree

- The Classification Tree used for feature selection performs with high accuracy, precision, recall, and F1 scores, suggesting it may model this dataset efficiently already
- Retains interpretability and readily captures non-linear relationships as well as interactions between different patient attributes

## Logistic Regression

- Comparisons of baseline models in the literature on this dataset suggest Logistic Regression performs better than SVM, Random Forest, and Neural Net Classification models for this data on average
- A statistical model allows for probabilistic results interpretations and does not rely on heuristics

# Model Comparison - Training Accuracy

| Cluster Regression Combination | Latent Class Analysis | Gaussian Mixture | K-Prototype |
|---|---|---|---|
| Classification Tree | **0.92** | **0.91** | **0.91** |
| Logistic Regression | **0.89** | **0.88** | **0.89** |

# Model Comparison – Testing Accuracy

| Cluster Regression Combination | Latent Class Analysis | Gaussian Mixture | K-Prototype |
|---|---|---|---|
| Classification Tree | **0.82** -0.10 | **0.73** -0.18 | **0.80** -0.11 |
| Logistic Regression | **0.81** -0.08 | **0.87** -0.01 | **0.81** -0.08 |

1. Research Objective
2. Data Description
3. Modeling Methodology
4. **Results & Conclusion**

# Final Model Interpretation

**Selected Clustering Method:** Gaussian Mixture

- 2 clusters (High Risk, Low Risk): easy to profile and interpret
- Allows for calculating key metric: Relative Risk

**Key Indicator:** Relative Risk (*RR*)

$$\frac{\text{mean } \mathbb{P}(\mathbf{num} = 1 | C_1)}{\text{mean } \mathbb{P}(\mathbf{num} = 1 | C_0)} \approx \frac{\text{proportion of } \mathbf{num} = 1 \text{ in } C_1}{\text{proportion of } \mathbf{num} = 1 \text{ in } C_0} = RR = 4.43$$

- A patient classified in the "high-risk" cluster ($C_1$) is 4.4 times more likely to develop heart disease than a patient classified in the "low risk" cluster ($C_0$)
- Our logistic regression model's predictive probabilities of heart disease converge to the empirical proportions, indicating model stability and well-separated clusters

```
Relative Risk (RR) Analysis
Heart Disease RR by Cluster
******************************
LogReg Probabilities
    RR in Cluster 1: 4.4309
    RR in Cluster 0: 0.2257
    P(HD|Cluster 1): 0.8022
    P(HD|Cluster 0): 0.1810
******************************
Frequentist Proportions
    RR in Cluster 1: 4.4312
    RR in Cluster 0: 0.2257
    P(HD|Cluster 1): 0.8022
    P(HD|Cluster 0): 0.1810
******************************
```

# Conclusion

With easy-to-interpret clusters with distinct risk levels for developing heart diseases, we believe **our model can be implemented as a "Red Flag" system** to **assist physicians** in **detecting patients who are at risk of developing heart diseases** based on their health condition.

Further development of the model include:

- Incorporating patients' lifestyle traits to further develop the model to recognize lifestyle patterns that might contribute towards risk levels
- Expanding the training dataset to further refine and validate our model

Moral of the Story:

# The Machine Never Stops Learning

Thanks for listening!