Maria Jose Jimenez
Sofia Fraija
David Silvera
Carlota Huertas
Kushaal Raju Palasamudrum

Predicting and Classifying Heart Disease Subtypes Using Machine Learning: Early Detection of SCA, MI, Heart Failure, and CAD

I. Introduction/Background

Cardiovascular diseases (CVDs) are the leading cause of death globally, responsible for approximately 17.9 million deaths annually, accounting for 31% of all global deaths. Major contributors include Sudden Cardiac Arrest (SCA)—the sudden loss of all heart activity due to an irregular heart rhythm, Myocardial Infarction (MI)—the death of heart muscle tissue caused by insufficient blood flow, Heart Failure—the heart's inability to pump blood effectively, and Coronary Artery Disease (CAD)—the buildup of fats, cholesterol, and other substances that reduce blood flow to the heart muscle.

Early detection and precise classification of these conditions are essential for timely intervention and treatment. However, current diagnostic tools are often limited by accessibility, high costs, or lack of specificity, resulting in many patients remaining undiagnosed until the disease progresses to a critical stage. This project aims to develop a machine learning model capable of predicting heart disease and classifying it into one of these specific conditions. Such a model can assist healthcare professionals by providing precise diagnostic tools, facilitating timely interventions, reducing mortality rates, and alleviating the burden of cardiovascular diseases on individuals and healthcare systems.

We utilize the Heart Failure Prediction Dataset from Kaggle, which includes 11 clinical features—such as age, sex, chest pain type, and cholesterol levels—across 918 observations sourced from datasets like Cleveland, Hungarian, and Switzerland. More details can be found here.

Dataset link: Heart Failure Prediction Dataset on Kaggle

II. Problem Definition

While predicting heart disease is valuable, distinguishing between different heart conditions such as SCA, MI, Heart Failure, and CAD is even more critical for clinical decision-making. Each condition requires a unique treatment approach, and an accurate classification can significantly improve patient outcomes. For example, SCA has a very low survival rate (around 10%) if not treated immediately, whereas early intervention in MI cases can prevent heart damage. The motivation behind this project is to develop a machine learning model that predicts not only the presence of heart disease but also classifies it into these subtypes, providing actionable insights for personalized treatment.

III. Methods

1. Data Preprocessing Steps

To prepare the Heart dataset for analysis and modeling, two major data preprocessing steps were applied:

1.1 One-Hot Encoding

The Heart dataset includes several categorical features that are essential for identifying cardiovascular disease, such as *sex*, *chest pain type*, and *resting ECG*. These categorical variables were transformed through one-hot encoding, which converts each category into individual binary columns. This process enables our models to interpret and use categorical data more effectively by representing each category as a distinct feature.

1.2 Standardization

By the same token, the dataset contains numerous numerical features, such as *age*, *cholesterol*, and *resting BP*. Each of these possesses different ranges that could unevenly influence model performance. To address this, we standardized the data by centering it (mean = 0) and scaling it to unit variance (standard deviation = 1). This step improves the performance and convergence of certain classification models by ensuring that all numerical features contribute proportionately.

2. Implemented Models

We utilized three machine learning models to tackle the problem, combining supervised and unsupervised approaches. The two supervised models, Logistic Regression and Support Vector Machines (SVM), were designed to predict whether an individual has heart disease based on clinical features. Supervised learning relies on labeled data, where the model is trained with inputs (features like age or cholesterol levels) paired with outputs (heart disease presence or absence). Logistic Regression offers interpretability, while SVM uses hyperplanes to classify data, excelling in handling high-dimensional and complex relationships with both linear and non-linear kernels. On the other hand, K-Means Clustering represents an unsupervised learning approach, where the model identifies patterns and groups the data into clusters without predefined labels. This was used to classify the type of heart disease among individuals already diagnosed, providing insights into potential subtypes based on clinical feature groupings. Together, these models offer a comprehensive framework for prediction and classification in heart disease analysis.

2.1 Logistic Regression Model

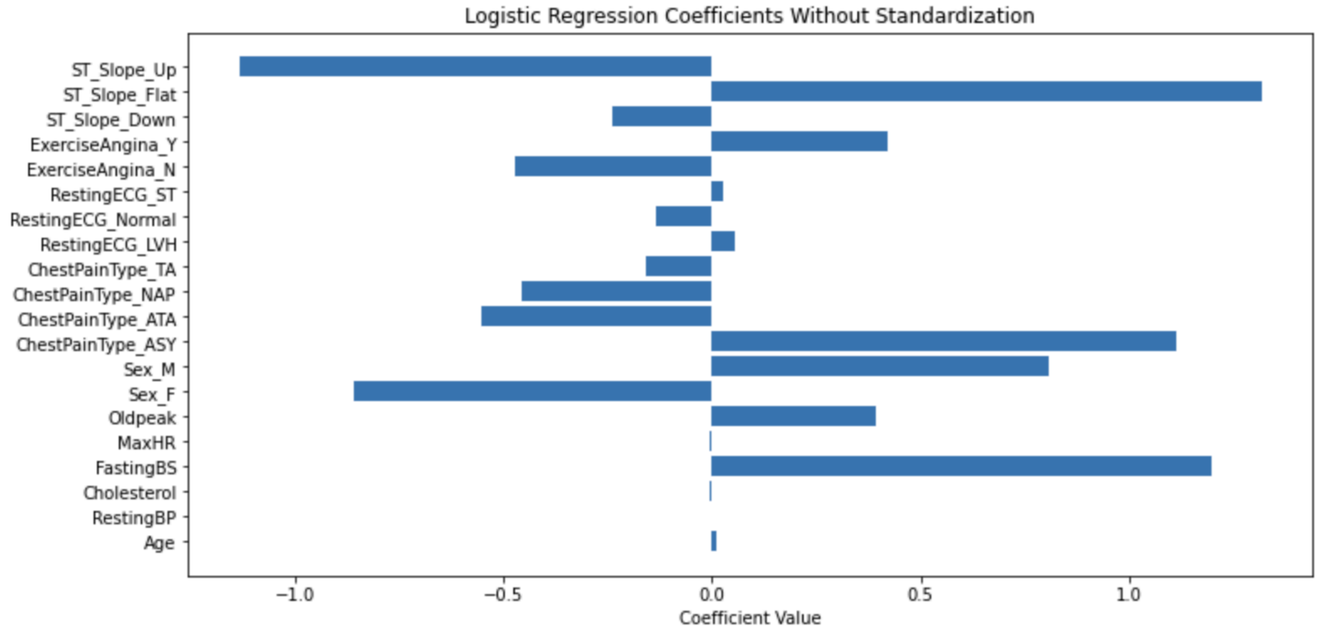
To accurately classify cardiovascular disease, we implemented a logistic regression model. Logistic regression is a supervised machine learning algorithm used for binary classification tasks, where the goal is to predict one of two possible outcomes. Unlike linear regression, which predicts a continuous value, logistic regression estimates the probability that a given input belongs to a particular class (in this case, the presence or absence of heart disease). It does so by applying the logistic (sigmoid) function, which outputs a probability value between 0 and 1. A probability threshold is then used to classify each instance into one of the two classes.

We decided to pursue this model for two key reasons:

1. **Binary Classification Requirement**: The values of the *HeartDisease* column in our dataset are binary, where 1 indicates the presence of heart disease and 0 represents the lack thereof. Logistic regression is well-suited for such binary classification tasks because it directly estimates the probability of each class.
2. **Interpretability**: Logistic regression is highly interpretable, providing coefficients for each feature that indicate the direction and strength of their relationship with the likelihood of heart disease. This feature makes it easier to understand the impact of each variable, such as *cholesterol* or *age*, on heart disease risk, which is valuable in a medical context where interpretability is often crucial.

A. Logistic Regression Model - *Visualization*

The **feature importance bar plot** provides a clear visual representation of how each feature influences the likelihood of a patient having heart disease, as predicted by the logistic regression model. The size and direction of the coefficients reflect the impact of each feature on the prediction outcome



B. Logistic Regression Model - *Quantitative Metrics*

To assess the effectiveness of the logistic regression model in predicting heart disease, several quantitative metrics were used. These metrics provide insights into different aspects of the model's performance, helping us understand its accuracy and potential areas for improvement.

- **Accuracy**: Accuracy measures the proportion of correct predictions made by the model, calculated as the ratio of true positives and true negatives to the total number of predictions. Our model achieved an accuracy score of **88%**. This indicates that the model correctly classifies a large portion of the instances.
- **Precision**: Precision is the ratio of true positives to the sum of true positives and false positives. High precision indicates that when the model predicts heart disease, it is likely correct. Our model achieved a precision of **86%** for negative diagnoses (0) and **90%** for positive diagnoses (1), indicating strong reliability.
- **Recall (Sensitivity)**: Recall is the ratio of true positives to the sum of true positives and false negatives. High recall ensures that the model correctly identifies most cases of heart disease, minimizing the risk of missed diagnoses. Our model achieved a recall of **88%** for both negative and positive diagnoses.

C. Logistic Regression Model - *Analysis of Model*

The logistic regression model showed the following results:

Accuracy: The model achieved an accuracy of 88%, indicating that it correctly classified X% of the instances. This high accuracy suggests that the model performs well overall, though it should be considered alongside other metrics. To gauge, this level is comparable to many prior established clinical diagnostic tools, showing the viability of the model.

Precision and Recall: With such high precision and recall results, the model demonstrates a balanced ability to accurately predict heart disease cases while also minimizing false negatives. High recall is particularly important in this scenario, as correctly identifying patients with heart disease is critical for timely intervention.

From our visualization showing the logistic regression coefficients, we can also identify which features are positively and negatively correlated with heart disease. For instance, bigger values for *age* and *max heart rate* are typically correlated with heart disease.

D. Logistic Regression Model - *Next Steps*

To further improve the predictive performance of the model and enhance its applicability, one major next step comes to mind: cross-validation. Applying cross-validation to evaluate model robustness and minimize overfitting, providing a more reliable estimate of model performance across different subsets of data.

Another step that may provide utility would be feature interaction analysis. Interactions between symptoms may have combined effects on heart disease risk, and understanding this interplay could improve our model's predictive ability

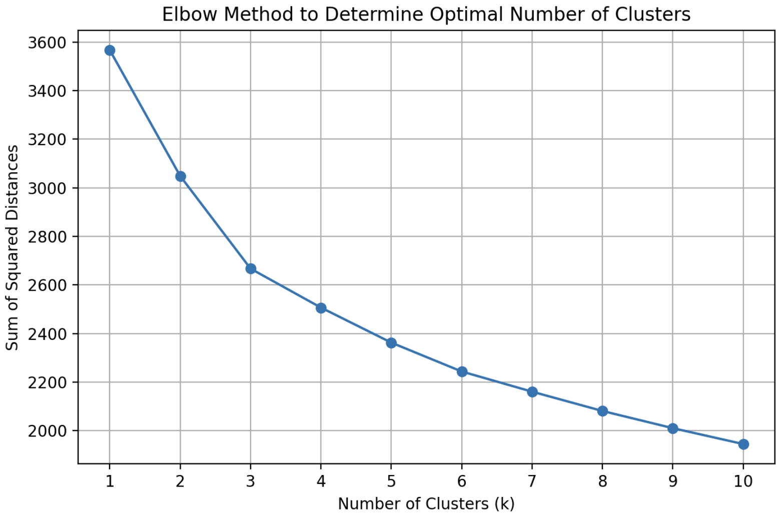
2.2 K-Means Clustering

K-Means clustering is an unsupervised machine learning algorithm used to partition data into distinct clusters by minimizing the variance within each cluster. We applied this model to extend our analysis beyond simply predicting heart disease. By implementing K-Means, we sought to uncover natural groupings within the data and align these clusters with known heart disease subtypes: Sudden Cardiac Arrest (SCA), Myocardial Infarction (MI), Heart Failure, and Coronary Artery Disease (CAD).

Our analysis focused solely on the subset of data representing patients who had confirmed heart disease. This filtering step was essential to ensure that our clustering efforts would yield meaningful and medically relevant insights, as grouping both diseased and non-diseased cases would dilute the analysis and reduce the interpretability of clusters in the context of disease subtypes.

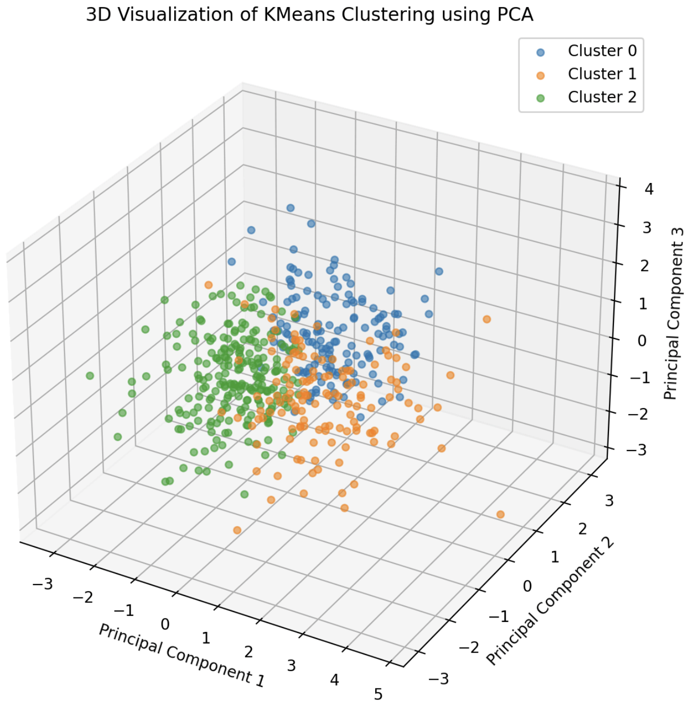
A. K-Means - *Visualizations*

The **Elbow Method** is a technique used to determine the optimal number of clusters for K-Means. It involves plotting the sum of squared distances (SSD) from data points to their assigned cluster centers for a range of potential cluster numbers (k). The "elbow" point, where the rate of decrease sharply slows, indicates the best k value. As shown in our Elbow Method graph, the optimal number of clusters was 3.



Graph 1: Elbow Method to Determine Optimal Number of Clusters Our analysis revealed that k=3 was the most suitable choice for clustering. We then implemented K-Means with k=3 and undid the standardization of features to interpret the results meaningfully.

To visualize the clusters in three dimensions, we used **Principal Component Analysis (PCA)**, which reduces the dataset's dimensionality while retaining the most significant variance. The 3D scatter plot highlights how patients are grouped based on the most influential features.



Graph 2: 3D Visualization of K-Means Clustering using PCA This visualization offers a clear depiction of how the data points are distributed across the three clusters, making it easier to understand the underlying patterns.

B. K-Means Clustering - *Quantitative Metrics*

After running the K-Means model, we obtained the following cluster centers in the original scale, which we mapped to specific heart disease subtypes:

<u>Cluster 0 Results:</u>

- Key Features:
  - Age = 59.28
  - MaxHR = 112.1 (low)
  - ST_Slope_Flat = 0.76
  - ChestPainType_ASY = 0.79
  - ExerciseAngina_Y = 0.58
- Mapped Subtypes:
  - Coronary Artery Disease (CAD): Due to asymptomatic chest pain and exercise-induced angina.
  - Heart Failure: Suggested by the lower MaxHR and flat ST slope, which indicate heart muscle weakness.

Reasoning: Despite the low cholesterol, the combination of older age, lower MaxHR, and ischemic indicators suggests CAD and potential heart failure risk.

Cluster 1 Results:

- Key Features:
  - Age = 47.81 (younger)
  - Cholesterol = 190.57
  - MaxHR = 149.28 (high)
  - ChestPainType_ASY = 0.69
  - ExerciseAngina_N = 0.67
- Mapped Subtypes:
  - Mixed Indicators (Potential Early CAD or Low Risk for MI): Asymptomatic chest pain is concerning, but the overall lower risk factors point to early or less severe conditions.

Reasoning: This cluster represents a lower-risk profile, possibly indicating early signs of heart issues but no immediate threat of severe conditions.

Cluster 2 Results:

- Key Features:
  - Age = 58.53
  - Cholesterol = 257.71 (high)
  - RestingBP = 142.15 (elevated)
  - Oldpeak = 1.92 (high)
  - ST_Slope_Flat = 0.81
  - ExerciseAngina_Y = 0.82
- Mapped Subtypes:
  - Myocardial Infarction (MI): Indicated by high cholesterol, hypertension, and significant ST depression.
  - Coronary Artery Disease (CAD): High cholesterol and exercise-induced angina make CAD highly likely.

Reasoning: The presence of multiple high-risk factors strongly suggests MI and CAD.

C. K-Means Clustering - *Analysis of Model*

The silhouette score for the K-Means clustering was 0.1474, reflecting moderate clustering performance. This value suggests some overlap between clusters but still allows for meaningful differentiation among groups. Despite the modest score, the identified clusters align with real-world subtypes of heart disease, such as CAD, MI, and Heart Failure. For instance, patterns such as high cholesterol and significant ST slope values were strongly associated with MI, while lower maximum heart rates and asymptomatic chest pain were linked to CAD. This mapping back to clinical contexts enhances the interpretability of the clusters, providing valuable insights despite the limitations in separation. However, the relatively low silhouette score indicates room for improvement, either through feature engineering to better represent relationships or by exploring alternative clustering methods that may improve group distinctiveness and cohesion. These results demonstrate the potential of K-Means to reveal patterns in complex medical data, even with moderate clustering performance.

2.3 Support Vector Machine (SVM)

Support Vector Machines (SVM) are a powerful supervised learning algorithm used for classification and regression tasks. The primary objective of SVM is to find a hyperplane that best separates data points of different classes in a high-dimensional space. In order to do this, SVM uses a **margin**, which is the distance between the hyperplane and the nearest data points from each class (support vectors). SVM also uses **kernels**, which enable SVM to handle non-linearly separable data by mapping it into a higher-dimensional space. In our implementation, we tested different kernels to compare their accuracy.

We decided to pursue this model for two key reasons:

- **High Dimensionality**: The dataset includes an abundance of numerical and categorical features. SVM efficiently handles datasets with multiple features by creating hyperplanes in high-dimensional spaces.
- **Model Complexity**: Using the RBF kernel, for instance, SVM can capture intricate, non-linear relationships between features. This differs from the logistic regression model, which models linear relationships.

A. Support Vector Machine - *Quantitative Metrics*

We implemented SVMs with both linear and non-linear kernels (RBF) to evaluate their performance on the Heart Failure Prediction Dataset. The following metrics were used to assess the model's performance: accuracy, precision, recall, F1-score, and confusion matrix.

A.1 Linear Kernel Results

- **Accuracy**: Accuracy represents the percentage of all the correct predictions (both true positives and true negatives) out of the total number of predictions. The linear kernel achieved an accuracy score of 83.70%, indicating that out of 100 people, 84 were classified correctly (whether they had heart disease or not).
- **Precision**: Precision focuses on when the model predicts someone has heart disease (**1**). It tells us how often the model's prediction of "heart disease" is actually correct. From

our dataset, a precision of 89% for class **1** means that when the model says someone has heart disease, it's right 89% of the time. On the other hand, for class **0** (no heart disease), a precision of 78% means that when the model predicts "no heart disease," it's right 78% of the time.

- **Recall**: Recall measures how good the model is at finding all the people who truly have heart disease (**1**). For example, a recall of 82% for class **1** means that out of 100 people who actually have heart disease, the model correctly identified 82 of them. Similarly, for class **0**, a recall of 86% means the model correctly identified 86 out of every 100 people without heart disease
- **F1-Score**: The balanced F1-scores (0.81 for class 0 and 0.85 for class 1) confirm a strong balance between precision and recall across classes.

A.2 RBM Kernel Results

Using a similar logic as the linear kernel:

- **Accuracy:** The RBF kernel achieved an accuracy score of 86.41%, indicating the model classified approximately 84 out of 100 people correctly.
- **Precision:** For class 1 (heart disease), the precision was 89%, and for class 0 (no heart disease), it was 82%. This reflects the model's reliability in correctly identifying positive and negative predictions.
- **Recall:** The recall was 87% for class 1, meaning the model identified 87 out of 100 people who actually had heart disease. For class 0, the recall was 86%, showing it successfully found 86 out of 100 people without heart disease.
- **F1-Score:** The F1-scores were 0.88 for class 1 and 0.84 for class 0, maintaining a strong balance between precision and recall.

B. Support Vector Machine - *Analysis of Mode*l

Accuracy: Both the linear and RBF kernels achieved high accuracy, with the RBF kernel slightly outperforming the linear kernel (83.96% vs. 83.70%). This small improvement suggests that the RBF kernel better captures the non-linear patterns in the dataset. From the confusion matrix, the RBF kernel correctly classified 159 out of 184 samples, whereas the linear kernel correctly classified 154 out of 184 samples.

Precision vs. Recall: The RBF kernel achieved a precision of 89% and a recall of 87% for identifying heart disease (class 1), reflecting its ability to reliably predict and detect true cases. The confusion matrix highlights this balance, with fewer false positives (11) and false negatives (14).

Kernel Comparison: The RBF kernel demonstrated a slight edge over the linear kernel, correctly classifying more samples (159 vs. 154) and effectively handling the dataset's non-linear patterns. This improvement aligns with the complex relationships often seen in clinical features.

Clinical Relevance: The high recall for the positive class (87% for the RBF kernel) is particularly significant in the context of healthcare, where failing to identify patients with heart disease (false negatives) can have serious consequences. Furthermore, the model's ability to maintain a low false positive rate (as seen in the confusion matrix) reduces unnecessary follow-ups or treatments, enhancing its practical applicability as a diagnostic tool.

	Predicted: 0	**Predicted 1**
Actual: 0	66	11
Actual: 1	19	88

Table 1. Confusion Matrix Linear Kernel

	Predicted: 0	**Predicted 1**
Actual: 0	66	11
Actual: 1	14	93

Table 2. Confusion Matrix RBF

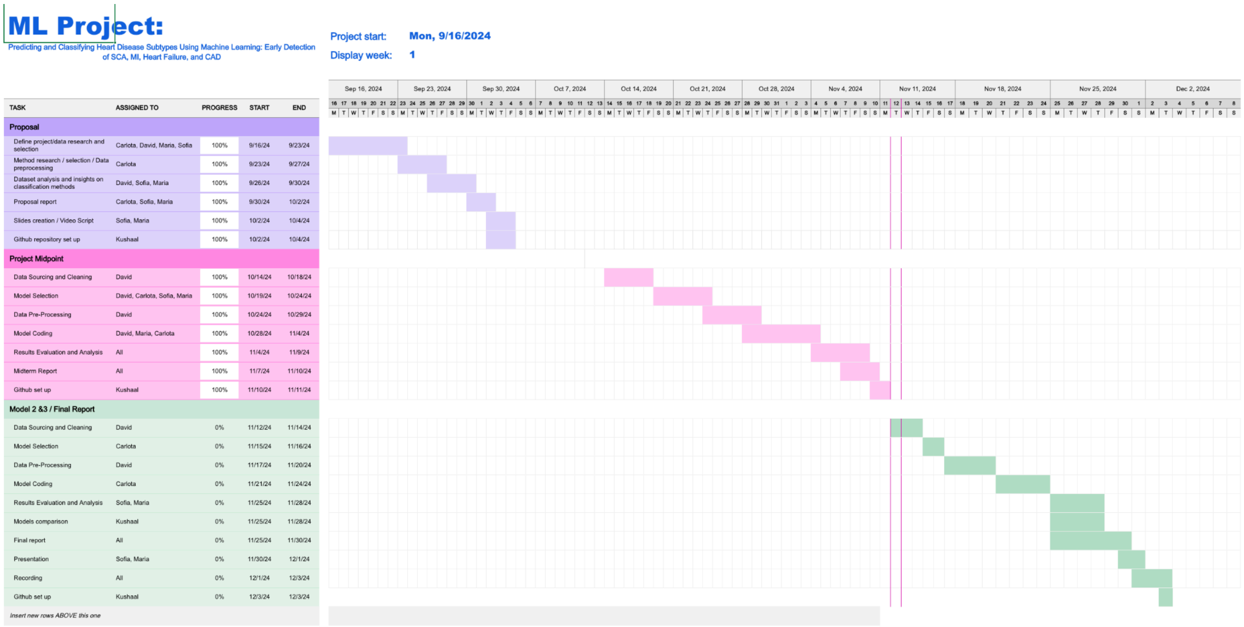
3. Comparison of Models

In total, we implemented three different machine learning algorithms to analyze cardiovascular disease. Our supervised methods include logistic regression and SVM, while the unsupervised method we selected is k-means clustering.

The contrast between both of our supervised methods lies predominantly on the type of relationship they capture. Logistic regression is designed for binary or multi-class classification problems. It is straightforward to implement, interpretable, and performs well when the relationship between features and the target is approximately linear. SVM, on the other hand, is more versatile and powerful, especially for complex datasets. By employing kernels, SVM can model non-linear relationships and create robust decision boundaries even in high-dimensional spaces, making it suitable for cases where logistic regression may underperform. Judging by their quantitative performance, logistic regression and SVM achieved an accuracy of 88% and 84% respectively, meaning logistic regression performed better.

Since K-means is our only unsupervised method and does not rely on labeled data, it cannot be directly compared to the supervised methods with a benchmark; therefore, we exclude it from this comparison.

IV. Gantt Chart



ML Project:

Predicting and Classifying Heart Disease Subtypes Using Machine Learning: Early Detection of SCA, MI, Heart Failure, and CAD

Project start: Mon, 9/16/2024

Display week: 1

TASK	ASSIGNED TO	PROGRESS	START	END
Proposal				
Define project/data research and selection	Carlota, David, Maria, Sofia	100%	9/16/24	9/23/24
Method research / selection / Data preprocessing	Carlota	100%	9/23/24	9/27/24
Dataset analysis and insights on classification methods	David, Sofia, Maria	100%	9/26/24	9/30/24
Proposal report	Carlota, Sofia, Maria	100%	9/30/24	10/2/24
Slides creation / Video Script	Sofia, Maria	100%	10/2/24	10/4/24
Github repository set up	Kushaal	100%	10/2/24	10/4/24
Project Midpoint				
Data Sourcing and Cleaning	David	100%	10/14/24	10/18/24
Model Selection	David, Carlota, Sofia, Maria	100%	10/19/24	10/24/24
Data Pre-Processing	David	100%	10/24/24	10/29/24
Model Coding	David, Maria, Carlota	100%	10/28/24	11/4/24
Results Evaluation and Analysis	All	100%	11/4/24	11/9/24
Midterm Report	All	100%	11/7/24	11/10/24
Github set up	Kushaal	100%	11/10/24	11/11/24
Model 2 &3 / Final Report				
Data Sourcing and Cleaning	David	0%	11/12/24	11/14/24
Model Selection	Carlota	0%	11/15/24	11/16/24
Data Pre-Processing	David	0%	11/17/24	11/20/24
Model Coding	Carlota	0%	11/21/24	11/24/24
Results Evaluation and Analysis	Sofia, Maria	0%	11/25/24	11/28/24
Models comparison	Kushaal	0%	11/25/24	11/28/24
Final report	All	0%	11/25/24	11/30/24
Presentation	Sofia, Maria	0%	11/30/24	12/1/24
Recording	All	0%	12/1/24	12/3/24
Github set up	Kushaal	0%	12/3/24	12/3/24
Insert new rows ABOVE this one				

V. Contribution Table

Name	Responsability
David Silvera	● Data Preprocessing ● Model Implementation ● Visualization ● Quantitative metrics ● Video presentation
Maria Jose Jimenez	● Introduction ● Problem Definition ● Presentation slides ● Analysis of Model ● Video presentation
Carlota Huertas	● Data Preprocessing ● Model Implementation ● Visualization ● Analysis of Model ● Video presentation
Sofia Fraija	● Quantitative Metrics ● Next Steps ● Presentation slides ● Github Oversight
Kushaal Palasamudrum	● Model analysis ● Github Oversight

VI. References

Kaggle Heart Failure Prediction Dataset (2021). Retrieved from [https://www.kaggle.com/fedesoriano/heart-failure-prediction].

Scikit-learn Model Evaluation Documentation (n.d.). Retrieved from [https://scikit-learn.org/stable/modules/model_evaluation.html].

Scikit-learn User Guide: Clustering (n.d.). Retrieved from [https://scikit-learn.org/stable/modules/clustering.html].

American Heart Association. (Feb. 16, 2024). *What your cholesterol levels mean*. https://www.heart.org/en/health-topics/cholesterol/about-cholesterol/what-your-cholesterol-levels-mean

Mayo Clinic. (Aug. 13, 2024). *Heart disease: Symptoms and causes*. https://www.mayoclinic.org/diseases-conditions/heart-disease/symptoms-causes/syc-20353118

Mayo Clinic. (Jan. 19, 2023). *Sudden Cardiac Arrest*. Retrieved from https://www.mayoclinic.org/diseases-conditions/sudden-cardiac-arrest/symptoms-causes/syc-20350634

American Heart Association. (Feb. 15, 2024). *What Is Heart Failure?* Retrieved from https://www.heart.org/en/health-topics/heart-failure/what-is-heart-failure

Cleveland Clinic. (April 20, 2023.). *Heart Attack (Myocardial Infarction)*. Retrieved from https://my.clevelandclinic.org/health/diseases/16818-heart-attack-myocardial-infarction

Mayo Clinic. (June 14, 2024.). *Heart Failure*. Retrieved from https://www.mayoclinic.org/diseases-conditions/heart-failure/symptoms-causes/syc-20373142