

# **DATA MINING AND MACHINE LEARNING (EBUS537)**

## **Individual Assignment 2 – Development of a novel data mining application**

**Set by Dr Eric Leung Submission**

**Deadline: 12th January 2024, 12:00 noon**

**Contribution to the Final Mark of Module: 50% Maximum Word Length: 1800 words  
(excluding references and appendices)**

### **1. Suggest an applicable area where a tool can be applied –**

#### **a. Brainstorm and suggest a real-life scenario ...**

Ans: Real Life Scenario : Applying machine learning tools like K-Means Clustering to predict heart disease risk in individuals based on relevant health indicators.

#### **b. Discuss why this tool is selected for this scenario :**

Ans: K-Means clustering, a part of unsupervised learning , is a versatile tool known for its ability to group data points based on similarity. In the context of heart failure prediction, it can be applied to identify distinct patient clusters with similar health profiles. This information can assist healthcare professionals in tailoring treatment plans and preventive measures for different patient groups. The technical features of K-means clustering make it a suitable choice for this scenario:

#### **1. Unsupervised Learning:**

One of the key reasons for choosing K-Means clustering is its status as an unsupervised learning algorithm. In the context of heart disease prediction, where the exact risk level may not be known for each individual, an unsupervised approach is advantageous. Unlike supervised learning algorithms that require labeled data for training, K-Means can operate without such labels, making it well-suited for scenarios where the outcome or risk level is not precisely defined for each patient.

#### **2. Data Exploration and Pattern Recognition:**

K-Means is known for its effectiveness in **data exploration and pattern recognition**. In the case of the Heart Disease Prediction dataset, which is likely to be diverse with various health indicators contributing to overall risk, K-Means can unveil hidden patterns within the data. By identifying groups of individuals with similar characteristics, the algorithm allows for a more nuanced understanding of risk factors, providing healthcare professionals with valuable insights for tailored interventions.

### 3. Clustering for Risk Stratification:

Another crucial aspect that makes K-Means a suitable choice is its capability for **risk stratification**. Heart disease risk is multifaceted, involving various factors such as age, cholesterol levels, and blood pressure. K-Means can effectively group individuals based on these factors, creating clusters that represent different risk profiles. This clustering approach facilitates targeted interventions and enables the development of personalized healthcare plans, as healthcare professionals can address the specific needs of each identified cluster.

### 4. Scalability and Efficiency:

The **scalability and efficiency of K-Means** further contribute to its selection for this scenario. Given the potentially large number of individuals in the Heart Disease Prediction dataset, the algorithm's computational efficiency and scalability make it well-suited for processing extensive datasets. This characteristic ensures that the algorithm remains efficient even when dealing with a considerable volume of observations, enhancing its practical applicability in real-world healthcare settings.

### 5. Centroid-Based Representation:

Moreover, the **centroid-based representation** of clusters in K-Means adds to its interpretability. The algorithm represents clusters using centroids, which serve as prototypes for each risk group. This centroid-based representation provides a clear understanding of the average health profile within each cluster, aiding healthcare professionals in comprehending the distinctive characteristics of different patient groups identified by the algorithm.

K-means clustering is chosen for heart disease risk prediction due to its unsupervised nature, ability to reveal hidden patterns, efficiency with large datasets, and suitability for dynamic risk stratification. Its interpretability and quantifiable evaluation metrics further enhance its applicability in the healthcare domain.

## 2. Identify and discuss a dataset –

### a. Introduce the dataset:

Ans: The UCI Machine Learning Repository's Index of **Heart Disease Data Sets** contains all of the used datasets on the following link:

<https://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/>

After accounting for 31% of all fatalities worldwide, cardiovascular diseases (CVDs) are the leading cause of death, killing an estimated 17.9 million lives annually. Heart attacks and strokes account for four out of every five CVD deaths, with persons under the age of 70 accounting for one-third of these untimely deaths. 12 characteristics in this dataset can be used to predict the likelihood of a heart disease, as heart failure is a typical event brought on by CVDs. A machine learning model can be very helpful in the early detection and management of cardiovascular disease and high cardiovascular risk individuals (due to the existence of one or more risk factors, such as hypertension, diabetes, hyperlipidemia, or already established disease).

This data set dates from 1988 and consists of four databases: Cleveland, Hungary, Switzerland, Long Beach V and Stalog (Heart). It has 76 properties total, including the anticipated attribute, however only a subset of 12 are used in the published studies. The patient's heart disease status is indicated in the "Heart Disease" field. It has two integer values: 0 for normal and 1 for cardiac disease.

It comprises 12 features such as

1. **Age**: age of the patient [years]
2. **Sex**: sex of the patient [M: Male, F: Female]
3. **ChestPainType**: chest pain type [TA: Typical Angina, ATA: Atypical Angina, NAP: Non-Anginal Pain, ASY: Asymptomatic]
4. **RestingBP**: resting blood pressure [mm Hg]
5. **Cholesterol**: serum cholesterol [mm/dl]
6. **FastingBS**: fasting blood sugar [1: if FastingBS > 120 mg/dl, 0: otherwise]
7. **RestingECG**: resting electrocardiogram results [Normal: Normal, ST: having ST-T wave abnormality (T wave inversions and/or ST elevation or depression of > 0.05 mV), LVH: showing probable or definite left ventricular hypertrophy by Estes' criteria]
8. **MaxHR**: maximum heart rate achieved [Numeric value between 60 and 202]
9. **ExerciseAngina**: exercise-induced angina [Y: Yes, N: No]
10. **Oldpeak**: oldpeak = ST [Numeric value measured in depression]
11. **ST\_Slope**: the slope of the peak exercise ST segment [Up: upsloping, Flat: flat, Down: downsloping]
12. **HeartDisease**: output class [1: heart disease, 0: Normal]

## Source

Several datasets that were previously available separately but had never been combined were combined to generate this new dataset. This dataset is the largest heart disease dataset currently

available for research purposes because it combines five different heart datasets with eleven shared features. The following five datasets were used in its curation:

- Cleveland: 303 observations
- Hungarian: 294 observations
- Switzerland: 123 observations
- Long Beach VA: 200 observations
- Stalog (Heart) Data Set: 270 observations

**Total: 1190 observations**

**Duplicated: 272 observations**

**Final dataset: 918 observations**

**b. the potential insights that can be generated through applying the selected tool to mine the data from the selected dataset.**

Ans : By applying K-Means clustering to this dataset, we can identify distinct patient clusters based on their health attributes. Insights may include recognizing specific risk factors or patterns associated with different clusters, aiding in personalized healthcare strategies. Some of Them are elaborated in detail:

### **1. Identification of Distinct Risk Profiles:**

K-Means clustering provides a granular understanding of patient populations by revealing distinct clusters based on health indicators. For example, one cluster might include individuals with high cholesterol, high blood pressure, and advanced age, indicating a high-risk group. This insight is crucial for healthcare professionals as it allows them to categorize patients into specific risk profiles, enabling more targeted and personalized care strategies for each group.

### **2. Personalized Risk Assessment:**

Assigning individuals to specific clusters allows healthcare professionals to conduct personalized risk assessments. Beyond traditional risk factors, the clustering approach considers a combination of health indicators, providing a more comprehensive view of an individual's health. This personalized risk assessment enables tailored interventions and preventive measures based on an individual's cluster membership, ensuring that healthcare strategies are aligned with the unique needs of each patient.

### **3. Targeted Intervention Strategies:**

Different clusters may exhibit varying risk factors and health characteristics. For instance, a cluster with younger individuals might show lifestyle-related risk factors, while an older cluster might demonstrate age-related factors. This diversity allows for the development of targeted intervention strategies. Younger clusters might benefit from lifestyle modifications and education, while older clusters might require more aggressive medical management. Tailoring interventions to specific clusters enhances the precision and effectiveness of healthcare strategies.

### **4. Early Detection of High-Risk Groups:**

K-Means clustering enables the early identification of high-risk groups within the population. This early detection is critical for implementing proactive healthcare measures. For instance, if a specific cluster is identified as high risk, healthcare professionals can initiate regular monitoring, early interventions, and lifestyle changes for individuals within that cluster. Early detection and intervention contribute significantly to mitigating the risk of heart disease and improving long-term outcomes.

### **5. Dynamic Monitoring and Adaptation:**

The clusters generated by K-Means clustering are not static; they can adapt as new data becomes available or additional risk factors are identified. This dynamic nature ensures that the model stays up-to-date with evolving healthcare knowledge. Healthcare professionals can continuously monitor and adapt interventions based on the changing landscape of risk factors. This adaptability enhances the resilience of the model and its ability to provide relevant insights over time.

### **6. Identification of Underlying Health Patterns:**

Clusters may reveal underlying health patterns that are not immediately apparent when considering individual risk factors in isolation. For example, a cluster with seemingly normal cholesterol levels but high stress indicators might highlight the importance of considering mental health in heart disease risk assessment. Uncovering these underlying patterns contributes to a more holistic understanding of health and enables healthcare professionals to address comprehensive well-being.

## **7. Optimization of Resource Allocation:**

Understanding the distribution of risk across clusters allows for optimized resource allocation in healthcare settings. Healthcare resources can be directed more efficiently to the high-risk clusters, ensuring that interventions are focused where they are most needed. This targeted approach not only maximizes the impact of interventions but also optimizes resource utilization, contributing to a more efficient and effective healthcare system.

Applying K-Means clustering to the Heart Disease Prediction dataset thus goes beyond traditional risk assessment, providing a multi-faceted approach to personalized risk evaluation, targeted interventions, and dynamic adaptation to evolving healthcare knowledge. These insights collectively enhance the effectiveness of heart disease prevention and management strategies, ultimately improving patient outcomes.

## **3. Apply the selected tool on the dataset you picked/created in Step 2 –**

### **a. Discuss and interpret the data mining results after applying the tool:**

### **Implementation Steps:**

- **Data Preprocessing:** Handle missing values, encode categorical variables, and scale numerical features.
- **Split the dataset** into training and testing sets.
- Train a **K-mean Clustering model** on the training set.
- **Evaluate the model** on the testing set and interpret feature importance.

### **1. Data Preprocessing :**

Data Preprocessing is a crucial stage in data analysis and machine learning that involves transforming raw, often messy data into a clean, organized, and usable format suitable for further analysis, modeling, or other downstream tasks. It's often considered the foundation for successful data mining and machine learning projects.

### **Key Steps and Processes in Data Preprocessing:**

## 1. Identify the Null or Unknown Values :

Null or unknown values refer to the absence or lack of information in a data set. These values typically represent missing or undefined data points. In databases or spreadsheets, null values indicate that a specific piece of information is not available or has not been recorded. Dealing with null values is crucial in data analysis and management, requiring careful handling to avoid skewed results or inaccuracies in computations. Understanding and addressing null values is essential for ensuring the integrity and reliability of data-driven processes. According Fig : 1, each column in heart Disease Dataset has not null or unknown values.

```
175] #checking the null values

df.isnull().sum()

Age          0
Sex          0
ChestPainType 0
RestingBP    0
Cholesterol  0
FastingBS    0
RestingECG   0
MaxHR        0
ExerciseAngina 0
Oldpeak      0
ST_Slope     0
HeartDisease 0
dtype: int64
```

*Fig-1: Identify Null or unknown values*

## 2. Info of each Column :

The `.info()` method provides information on the data including the DataFrame object, the number of entries, data types for each column, and number of non-null values. There are 918 entries and 12 columns. The data has a mixture of data types including int64 (integers), object, and float64

```
7] df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 918 entries, 0 to 917
Data columns (total 12 columns):
 #   Column                Non-Null Count  Dtype  
---  -
 0   Age                   918 non-null   int64  
 1   Sex                    918 non-null   object  
 2   ChestPainType          918 non-null   object  
 3   RestingBP              918 non-null   int64  
 4   Cholesterol             918 non-null   int64  
 5   FastingBS              918 non-null   int64  
 6   RestingECG             918 non-null   object  
 7   MaxHR                  918 non-null   int64  
 8   ExerciseAngina         918 non-null   object  
 9   Oldpeak                918 non-null   float64 
10   ST_Slope               918 non-null   object  
11   HeartDisease           918 non-null   int64  
dtypes: float64(1), int64(6), object(5)
memory usage: 86.2+ KB
```

*Fig-2: Info of each column*

## 3. Description of each column:

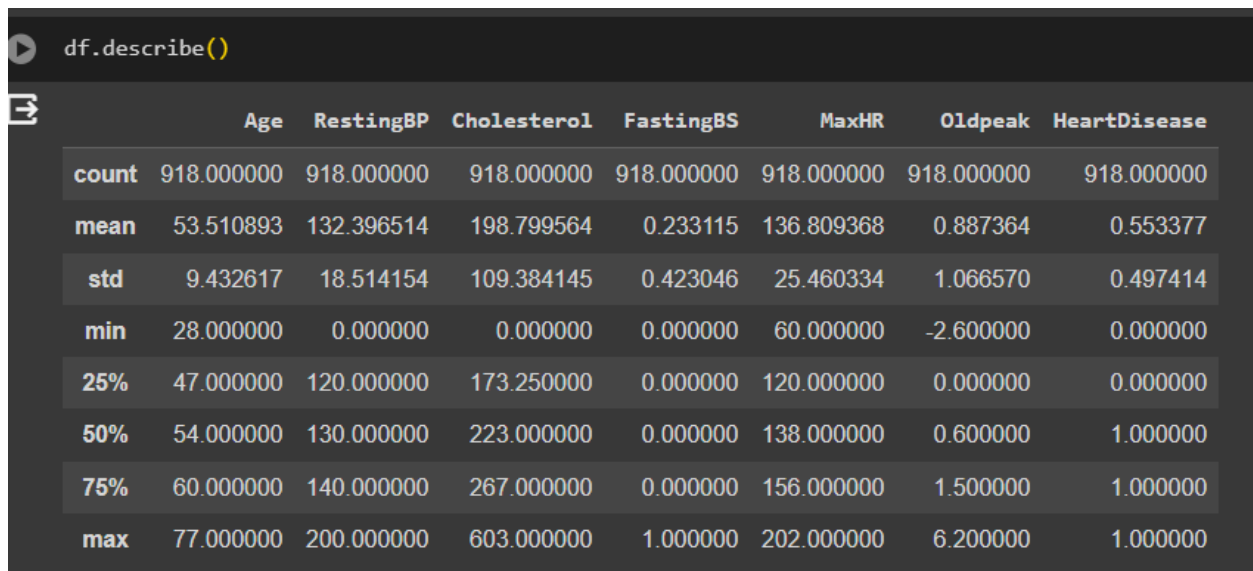
The `.describe()` method is used to provide some summary statistics of the data, including the mean, standard deviation, minimum, maximum, quartiles, etc., for each numerical column.



Here are some of the observations from the data:

- There are 918 entries in the dataset.
- The average age is 53.5 years.
- The average resting blood pressure is 132.4 mmHg.
- The average cholesterol level is 199 mg/dL.
- There are two entries with a fasting blood sugar of -2.6, which may be errors.
- The average maximum heart rate is 136.8 beats per minute.
- The average Oldpeak value is 0.89.

About 55% of the patients have heart disease



```
df.describe()
```

	Age	RestingBP	Cholesterol	FastingBS	MaxHR	Oldpeak	HeartDisease
count	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000	918.000000
mean	53.510893	132.396514	198.799564	0.233115	136.809368	0.887364	0.553377
std	9.432617	18.514154	109.384145	0.423046	25.460334	1.066570	0.497414
min	28.000000	0.000000	0.000000	0.000000	60.000000	-2.600000	0.000000
25%	47.000000	120.000000	173.250000	0.000000	120.000000	0.000000	0.000000
50%	54.000000	130.000000	223.000000	0.000000	138.000000	0.600000	1.000000
75%	60.000000	140.000000	267.000000	0.000000	156.000000	1.500000	1.000000
max	77.000000	200.000000	603.000000	1.000000	202.000000	6.200000	1.000000

Fig-3: Describe of each column

#### 4. Label Encoding :

Label encoding is a technique used to convert categorical variables into numerical values for machine learning algorithms.

In the code, different features like "Sex", "ChestPainType", "RestingECG", and "ST\_Slope" are transformed using the "LabelEncoder" function. This function assigns a unique integer to each category within the feature. For example, in the "Sex" feature, "1" might represent male and "0" female.

By converting categorical features into numerical values, the data can be prepared for use in machine learning models that only understand numerical data.

```
[178] from sklearn.preprocessing import LabelEncoder
      # Create a LabelEncoder instance
      label_encoder = LabelEncoder()

[179] # Label Encoding encoding
      df['Sex']=label_encoder.fit_transform(df['Sex'])
      df['ChestPainType']=label_encoder.fit_transform(df['ChestPainType'])
      df['RestingECG']=label_encoder.fit_transform(df['RestingECG'])
      df['ExerciseAngina']=label_encoder.fit_transform(df['ExerciseAngina'])
      # df['ExercerciseAngina']=label_encoder.fit_transform(df['ExercerciseAngina'])
      df['ST_Slope']=label_encoder.fit_transform(df['ST_Slope'])

[180] df.head()
```

	Age	Sex	ChestPainType	RestingBP	Cholesterol	FastingBS	RestingECG	MaxHR	ExerciseAngina	Oldpeak	ST_Slope	HeartDisease
0	40	1	1	140	289	0	1	172	0	0.0	2	0
1	49	0	2	160	180	0	1	156	0	1.0	1	1
2	37	1	1	130	283	0	2	98	0	0.0	2	0
3	48	0	0	138	214	0	1	108	1	1.5	1	1
4	54	1	2	150	195	0	1	122	0	0.0	2	0

Fig-4: Label Encoding

## 2. Split the dataset and Standardize :

The code uses the `train_test_split` function from the `scikit-learn` library to split the data into two sets: a training set that will be used to train the machine learning model, and a testing set that will be used to evaluate the model's performance.

The code specifies that 20% of the data will be used for testing, and the remaining 80% will be used for training. The `random_state` parameter is set to 42 to ensure that the split is reproducible.

This is an important step in the machine learning process, as it helps to prevent overfitting, which is when the model learns the training data too well and doesn't generalize well to new data.

```
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2, random_state=42)

# Standardize the features
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Fig-5: Split dataset into training and testing

### 3. Build and Train K-mean Clustering model

The code snippet shows K-Means clustering being performed on a scaled training dataset (`X_train_scaled`). Here's a breakdown:

- **K-Means object creation:** A `KMeans` object (`kmeans`) is created with `n_clusters=4` and `random_state=42`. This sets the number of clusters to find (4) and ensures repeatable results.
- **Model fitting:** The `fit` method on the `kmeans` object fits the K-Means model to the training data. This finds the centroids (central points) for each cluster.
- **Predictions:** The `predict` method is used to predict the cluster assignments for the testing data (`X_test_scaled`). Each data point is assigned to the closest cluster based on its features.

```
186] # Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(X, Y, test_size=0.2, random_state=42)

# Standardize the features
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Apply KMeans clustering
kmeans = KMeans(n_clusters=4, random_state=42)
kmeans.fit(X_train_scaled)

# Predict the clusters on the test set
test_predictions = kmeans.predict(X_test_scaled)

# Evaluate the clustering performance
silhouette_score = metrics.silhouette_score(X_test_scaled, test_predictions)
inertia = kmeans.inertia_

# Pseudolabel accuracy
cluster_to_class_mapping = {cluster: np.argmax(y_test[test_predictions == cluster]) for cluster in range(4)}
predicted_labels = np.array([cluster_to_class_mapping[cluster] for cluster in test_predictions])
pseudolabel_accuracy = np.mean(predicted_labels == y_test)

print(f"Silhouette Score: {silhouette_score}")
print(f"Inertia: {inertia}")
print(f"Pseudolabel Accuracy: {pseudolabel_accuracy * 100}%")

Silhouette Score: 0.13549477998849555
Inertia: 5376.1455862608875
Pseudolabel Accuracy: 83.15217391304348
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of 'n_init' will change from 10 to 'auto' in 1.4. Set the value of 'n_init' explicitly to suppress th
warnings.warn(
```

*Fig-6: Build and train K means Clustering*

### 4. Evaluation :

The code then calculates two metrics to evaluate the clustering performance:

- **Silhouette score:** This measures how well data points are assigned to their clusters. A higher score indicates better clustering.
- **Inertia:** This measures the distance of each data point to its assigned cluster's centroid. A lower inertia suggests tighter clusters.

```

# Evaluate the clustering performance
silhouette_score = metrics.silhouette_score(X_test_scaled, test_predictions)
inertia = kmeans.inertia_

# Pseudolabel accuracy
cluster_to_class_mapping = {cluster: np.argmax(y_test[test_predictions == cluster]) for cluster in range(4)}
predicted_labels = np.array([cluster_to_class_mapping[cluster] for cluster in test_predictions])
pseudolabel_accuracy = np.mean(predicted_labels == y_test)

print(f"Silhouette Score: {silhouette_score}")
print(f"Inertia: {inertia}")
print(f"Pseudolabel Accuracy: {pseudolabel_accuracy * 100}%")

Silhouette Score: 0.13549477998849555
Inertia: 5376.145962608875
Pseudolabel Accuracy: 83.15217391304348
/usr/local/lib/python3.10/dist-packages/sklearn/cluster/_kmeans.py:870: FutureWarning: The default value of 'n_init' will change from 10 to 'auto' in 1.4. Set the value of 'n_init' explicitly to suppress t
warnings.warn(

```

*Fig-7: Evaluation*

## **b. Discuss the novelty and significance of this application:**

### **1. Novelty in Utilization:**

The application of K-Means clustering in predicting heart failure brings a novel approach to patient stratification. Unlike traditional heart failure risk assessments that rely on predefined risk factors, K-Means clustering introduces an innovative method to uncover hidden patterns within the dataset. This novel perspective enhances our understanding of the heterogeneity present within the heart failure population, allowing for a more nuanced and precise risk assessment.

The novelty lies in the algorithm's ability to autonomously identify patterns that might not be immediately evident. This innovative approach challenges conventional methods and provides healthcare professionals with a fresh lens through which to view and address heart failure risks.

Additionally, the utilization of unsupervised learning aligns with advancements in machine learning, showcasing the adaptability of cutting-edge technologies in healthcare.

**Contribution to Personalized Medicine:** The novelty lies in the utilization of K-Means clustering for patient stratification, paving the way for personalized medicine. By recognizing distinct patient groups, healthcare practitioners can tailor interventions to specific needs, potentially improving treatment effectiveness. This move towards individualized care aligns with the evolving paradigm of healthcare.

The significance of this application extends to the realm of personalized medicine. By leveraging K-Means clustering for patient stratification, healthcare practitioners can move beyond generic risk assessments and tailor interventions to the specific needs of distinct patient groups. This individualized approach has the potential to significantly improve treatment effectiveness and patient outcomes.

Recognizing and understanding distinct patient clusters enables healthcare professionals to develop targeted strategies that consider the unique combination of risk factors within each group. This move towards individualized care is in harmony with the evolving paradigm of healthcare, emphasizing the importance of tailored interventions.

## **2. Significance for Treatment Outcomes:**

The significance of this application is evident in its potential to enhance treatment outcomes. Healthcare professionals can design targeted interventions based on the identified clusters, optimizing the allocation of resources and improving overall patient care.

**Optimized Resource Allocation:** The ability to allocate resources more efficiently is crucial in healthcare settings. Knowing which clusters represent high-risk patients allows for targeted monitoring and interventions, potentially reducing hospitalizations and improving long-term outcomes.

**Early Intervention Strategies:** Identifying clusters associated with early signs of heart failure allows for proactive intervention. Early treatment or lifestyle modifications can significantly impact the progression of the disease, demonstrating the potential for improved patient outcomes.

## **3. Insights for Public Health Planning:**

The application goes beyond individual patient care, providing valuable insights for public health planning. Understanding the distribution of clusters within the population allows public health planners to design targeted interventions at a broader level. If a particular cluster indicates a high prevalence of modifiable risk factors, public health campaigns can be tailored to address those specific issues within the community.

This macro-level application of insights contributes to more effective public health initiatives, aiming to reduce the overall burden of heart failure within a given population.

## **4. Continuous Improvement in Healthcare Protocols:**

The application of K-Means clustering offers a dynamic framework for continuous improvement in healthcare protocols. As more data becomes available and the model adapts to evolving knowledge, healthcare professionals can refine and update intervention strategies based on the changing landscape of heart disease risk factors.

This adaptability ensures that healthcare protocols remain relevant and effective in the face of emerging health trends and evolving patient demographics. The continuous improvement facilitated by K-Means clustering aligns with the principles of evidence-based medicine, allowing healthcare providers to stay at the forefront of advancements in heart failure prevention and management.

## **5. Integration with Electronic Health Records (EHR):**

Integrating K-Means clustering results with electronic health records (EHR) takes a significant step towards a holistic and data-driven approach to patient care. By incorporating cluster-based risk assessments into patient records, healthcare providers gain a comprehensive view of each individual's unique risk profile.

The seamless integration with EHR enhances the accessibility of cluster-based insights, allowing healthcare professionals to make informed decisions based on the most up-to-date information. This integration supports a patient-centered approach to healthcare, where interventions are tailored based on both historical and real-time health data.

In summary, the application of K-Means clustering in predicting heart failure introduces a novel and significant approach to personalized medicine, public health planning, continuous improvement in healthcare protocols, and integration with electronic health records. This multifaceted impact underscores the potential of advanced data mining techniques in transforming healthcare delivery and outcomes.