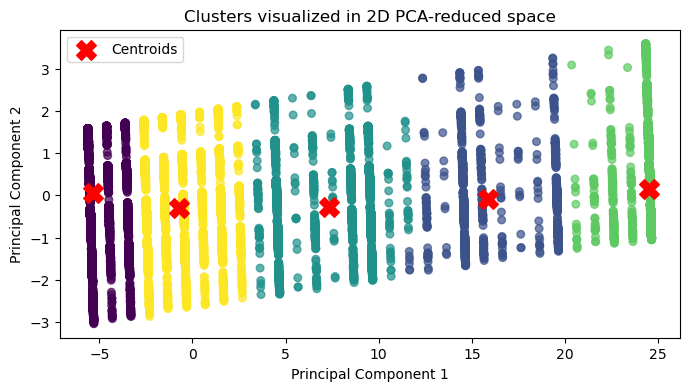
**Clustering**

Our dataset contains 20 features, which makes visualizing and interpreting clusters in higher dimensions challenging.

PCA (Principal Component Analysis) was used to reduce the dataset to 2 dimensions while preserving as much variance as possible. This allows us to visualize the clusters effectively in a 2D space.

The primary motivations for using PCA were:

1. **Dimensionality Reduction for Visualization:** PCA reduced the high-dimensional dataset to two principal components, enabling us to visualize and interpret the clusters effectively in 2D.
2. **Improving Clustering Performance:** By reducing the dimensionality, PCA helped minimize the curse of dimensionality, where distance metrics become less meaningful in higher dimensions, potentially degrading clustering quality.
3. **Reducing Redundancy:** PCA transformed the dataset into uncorrelated components, removing redundancy from highly correlated features and focusing on the most significant directions of variance.
4. **Noise Reduction:** By projecting the dataset onto the top principal components, PCA helped remove noise from less important dimensions, allowing K-Means to focus on the most relevant features.
5. **Efficiency:** Lower-dimensional data is computationally more efficient for K-Means clustering and avoids issues associated with working in a high-dimensional space.



**Cluster Separation**:

* Each colour represents a different cluster as assigned by K-Means.
* We can observe that the clusters are well-separated in the PCA-reduced space, which suggests that K-Means was able to group similar data points effectively, even though we've reduced the dimensionality.

**Centroids**:

* The red 'X' markers represent the centroids of each cluster. These centroids were calculated in the original high-dimensional space, then transformed to 2D PCA space for visualization.
* The placement of centroids in the middle of each cluster indicates that the algorithm has effectively identified central points that minimize the distances within each cluster.

**Insights on Data Structure**:

* From the spread and positioning of clusters, it appears that the dataset has some natural grouping tendencies, which K-Means was able to capture.
* The distribution along the principal components could suggest some underlying patterns or variations in the data that PCA managed to extract.

**Limitations**:

* Since this is a dimensionality-reduced representation, there might be information loss, and we may not see the full complexity of clusters in the original high-dimensional space.
* Overlapping clusters in PCA space might not be as distinct, especially if the data's original structure has overlapping clusters.
* **Silhouette Method** focuses on the quality of the clustering structure, measuring how well-separated clusters are.
* **Elbow Method** emphasizes the cost (inertia) and diminishing returns as you increase the number of clusters.

**Silhouette score Analysis:**

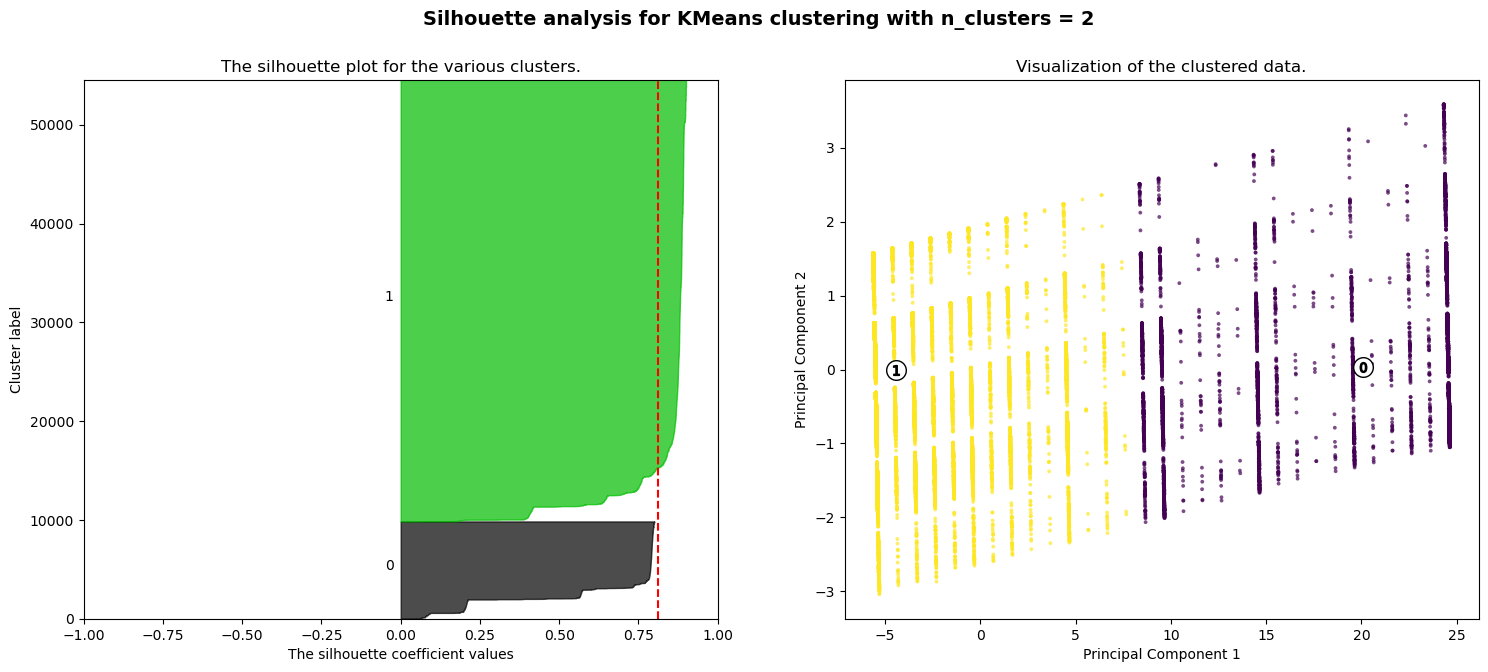
* For nclusters=2n\_{\text{clusters}} = 2nclusters​=2: The silhouette score is **0.811**, which is relatively high, suggesting that two clusters could be a reasonable choice for this dataset.
* For nclusters=3n\_{\text{clusters}} = 3nclusters​=3: The score drops to **0.751**, indicating that while three clusters might work, it doesn't perform as well as two clusters.
* For nclusters=4n\_{\text{clusters}} = 4nclusters​=4: The score is **0.617**, showing a further decrease, which suggests that increasing the number of clusters to four may not improve the clustering quality.
* For nclusters=5n\_{\text{clusters}} = 5nclusters​=5: The score drops to **0.531**, indicating even weaker clustering cohesion with five clusters.
* For nclusters=6n\_{\text{clusters}} = 6nclusters​=6: The score is **0.296**, which is quite low, suggesting that six clusters do not represent the structure in the data effectively.

**Optimal Number of Clusters**:

* Based on the silhouette scores, the optimal number of clusters appears to be **2**, as it has the highest silhouette score (0.811). This suggests that dividing the data into two clusters provides the most cohesive grouping, with each cluster having clear separation and compactness.

**Comparison with Higher Cluster Counts**:

* The decline in silhouette scores as the number of clusters increases suggests that splitting the data into more clusters causes a loss of cohesion within each cluster. This often happens if the data doesn't naturally divide into that many distinct groups, leading to fragmented clusters.



**For n\_clusters = 2**

* **Silhouette Score**: The highest average silhouette score, approximately 0.81, suggesting a strong separation between clusters.
* **Cluster Structure**: Two distinct clusters with good cohesion, as seen by the large silhouette widths.
* **Conclusion**: The data is well-separated into two clusters, which might be an optimal number of clusters. However, it’s quite simplistic, especially if there are complex relationships within the dataset.

A close-up of a graph

Description automatically generated  
**For n\_clusters = 3**

* **Silhouette Score**: The average silhouette score drops to around 0.75, but it still shows reasonable separation between clusters.
* **Cluster Structure**: Three clusters appear, but some regions in the plot show narrow silhouette values, indicating that some points are closer to other clusters.
* **Conclusion**: Adding a third cluster decreases cluster cohesion slightly, but this configuration still captures some additional structure within the data that may be beneficial for more nuanced segmentation.

A close-up of a chart

Description automatically generated  
**For n\_clusters = 4**

* **Silhouette Score**: The score decreases to around 0.62, indicating a moderate clustering structure.
* **Cluster Structure**: The four clusters have noticeable variations in silhouette widths, with some regions showing narrower values. This suggests that certain clusters overlap more, and some points are closer to the boundaries.
* **Conclusion**: While it captures additional complexity, the increased number of clusters also reduces cohesion, leading to less distinct clustering compared to two or three clusters.

A close-up of a chart

Description automatically generated  
**For n\_clusters = 5**

* **Silhouette Score**: The silhouette score continues to decrease to approximately 0.53, reflecting increased overlapping between clusters.
* **Cluster Structure**: The plot shows several clusters with lower silhouette values and more overlap, meaning the clusters are less well-separated.
* **Conclusion**: Five clusters may be too many for this dataset, as it results in less-defined clusters, increasing ambiguity in cluster assignments.

A close-up of a chart

Description automatically generated  
**For n\_clusters = 5**

* **Silhouette Score**: The silhouette score drops significantly to around 0.30, indicating poor separation and structure in clustering.
* **Cluster Structure**: Six clusters are shown, but with many points close to zero on the silhouette plot, suggesting considerable overlap and poor cohesion.
* **Conclusion**: Six clusters over-partition the data, resulting in poor clustering quality.

**Conclusion:**  
**Optimal Number of Clusters**: Based on the silhouette analysis, **two or three clusters** appear to be the most appropriate, with two clusters yielding the highest silhouette score.

**Comparison**: As the number of clusters increases, silhouette scores decrease, suggesting that additional clusters lead to overlapping or poorly defined clusters.

A graph with a line

Description automatically generated

We used the Elbow Method to verify the optimal number of clusters suggested by the Silhouette Analysis.

* This plot represents the inertia, or the total distance between points and their respective cluster centers, for different numbers of clusters.
* The elbow point, visible at k=2, indicates that 2 clusters strike a good balance between clustering quality and simplicity.

**Conclusion:**

Both the Silhouette Score and the Elbow Method point to k=2 as the most suitable choice for our dataset. Beyond 2 clusters, we see diminishing returns in terms of inertia reduction.

With this analysis, we can conclude that the data is best clustered into two groups, at least according to K-Means, the silhouette and elbow metrics.

When testing other clustering algorithms, such as Gaussian Mixture Models (GMM), we can see if similar results emerge or if a soft clustering method identifies more subtle substructures.

**Evaluating Clustering Results**

We ran the K-Means algorithm with two clusters, aligning with the binary nature of the target variable and based on our silhouette and elbow metrics results. The algorithm assigned each instance to one of the two clusters (Cluster A or Cluster B). To evaluate the clustering performance, we compared the predicted cluster assignments (y\_pred) with the true labels (y\_true) using a confusion matrix. This allowed us to see how many points in each cluster matched their true class.

**Results**

Here’s what we found with two clusters:

* Cluster A primarily represented the 'No heart disease' class, with a purity of 90%.
* Cluster B, however, had a much lower purity for the 'Yes heart disease' class at 26%.
* Overall, the clustering accuracy was 82%, meaning that 82% of all instances were correctly grouped into clusters that aligned with their true labels.

This analysis suggests that while Cluster A aligns well with one class, Cluster B has significant overlap, likely due to similarities in data points across the two classes.