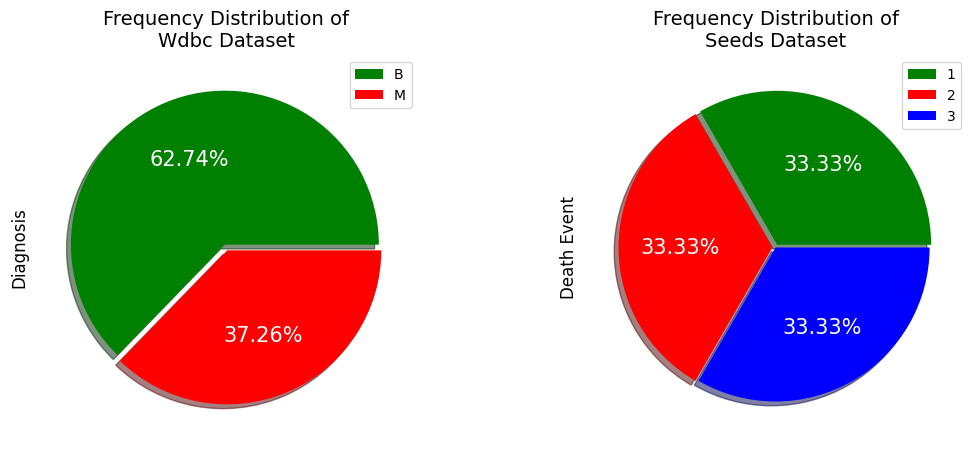
**Unsupervised Learning Algorithms**

1. **DATASETS**

Wheat Seeds Dataset UCI Wheat Seeds database includes 3 labels and 7 features that are well-distributed among the 3 classes. This makes it ideal for investigating algorithms and conducting classification and Clustering tasks.

Breast Cancer Wisconsin (Diagnostic) Dataset from UCI. The Breast Cancer Wisconsin (Diagnostic) dataset is used for classifying breast tumors as malignant or benign based on features extracted from cell nuclei in digitized images. It offers a diverse set of real-valued attributes, making it suitable for clustering to uncover potential subgroups or patterns within the data.



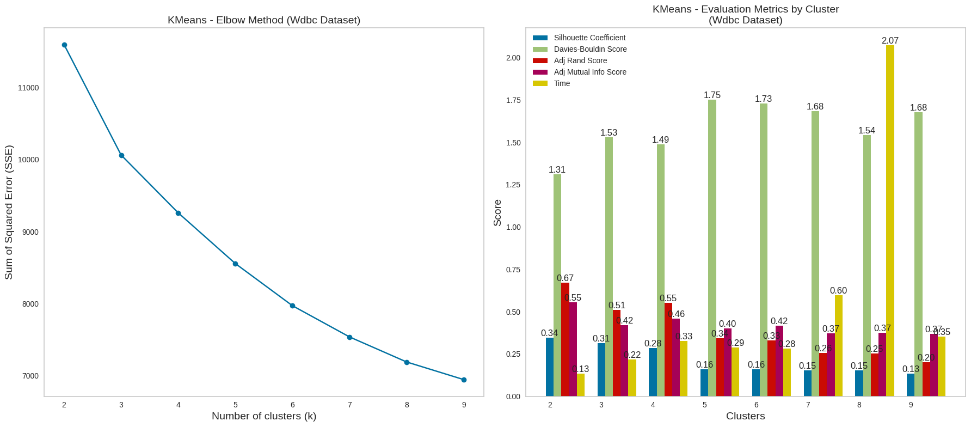
1. **CLUSTERING**

K-Means clustering group data into K clusters by iteratively updating cluster centroids until convergence, making it efficient for well-behaved data. In this case, the KMeans algorithm from scikit-learn uses the Euclidean distance by default as the distance metric to measure the distance between data points and cluster centroids.

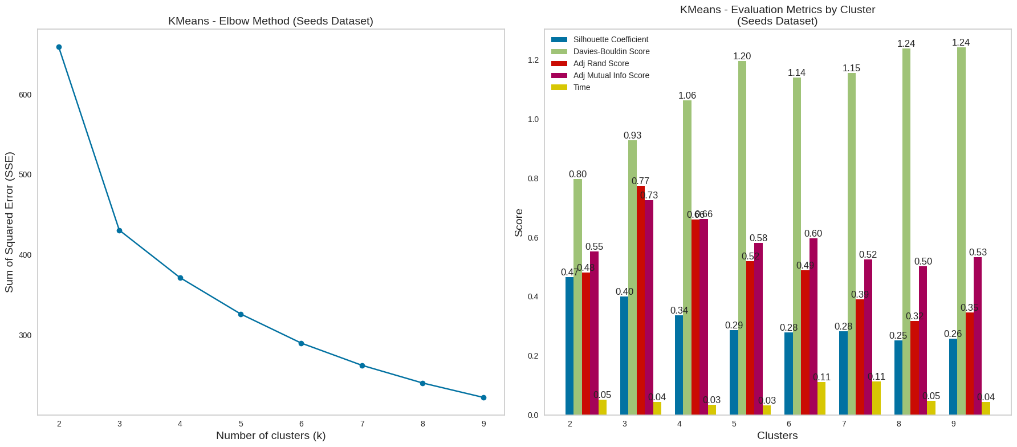
I have utilized a range of evaluation metrics to evaluate the clustering algorithms. The Silhouette Coefficient quantifies the similarity between data points within and between clusters, offering insights into cluster quality and cohesion. The Davies-Bouldin Score assesses cluster separation, measuring how distinct clusters are from each other. The Elbow Method helps pinpoint the ideal cluster count by analyzing the within-cluster sum of squares. In addition, the Adjusted Rand Score (ARS) and Adjusted Mutual Information Score (AMIS) provide valuable information on the agreement between cluster assignments and actual labels.

To determine the optimal number of clusters (K), metrics like the Silhouette Coefficient, Davies-Bouldin Score, Adjusted Rand Score, and Adjusted Mutual Information Score are useful, with higher values signifying better clustering; however, the Elbow Method provides a visual trade-off insight (1).

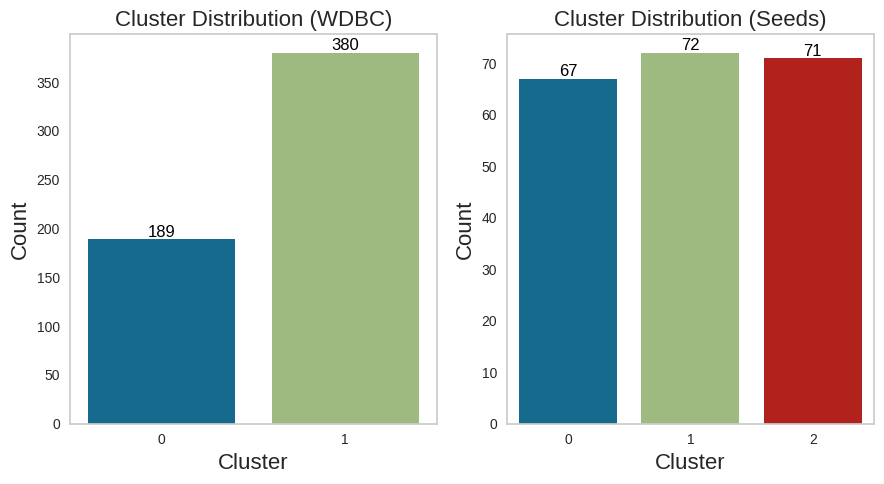
For the Wisconsin Breast Cancer Dataset, the optimal number of clusters is 2, as indicated by a high Adjusted Rand Score, Silhouette Score, Adjusted Mutual Information Score, and a low Davies-Bouldin Score. The silhouette sample plot visually supports this choice, reaffirming the clear and well-distributed clustering structure. Overall, a 2-cluster solution best aligns with the dataset's inherent structure and performs exceptionally well across multiple evaluation metrics, making it the preferred choice.



For the Seeds dataset, which originally had three ground truth labels, the evaluation metrics suggest that the optimal number of clusters is either 3 or 4. However, choosing 3 clusters aligns with the true class count and exhibits strong performance in terms of Silhouette, Davies-Bouldin, Adjusted Rand, and Adjusted Mutual Info scores, making it the most appropriate choice. This configuration provides a balance between cluster quality and adherence to the dataset's intrinsic structure.



Cluster distribution refers to the arrangement of data points into groups based on their similarities, which can vary from well-defined, distinct clusters to overlapping or sparse distributions. Cluster distribution for both Datasets is illustrated below.



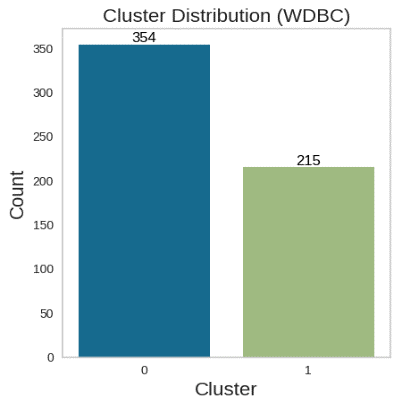
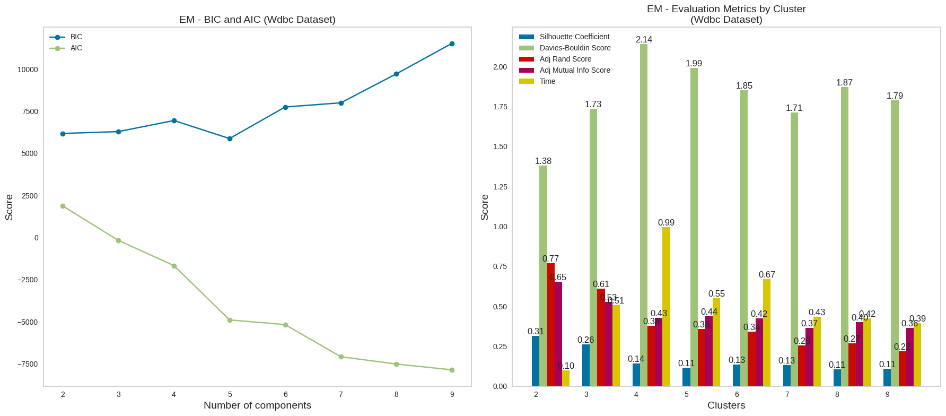
* 1. **Expectation Maximization**

The Expectation Maximization algorithm computes the probabilities of data points belonging to specific clusters, essentially maximizing the likelihood that the data points are associated with those clusters. I used the Gaussian Mixture Model (GMM). GMM is a probabilistic model that models data as a mixture of multiple Gaussian distributions, allowing for the soft assignment of data points to clusters and capturing complex cluster shapes.

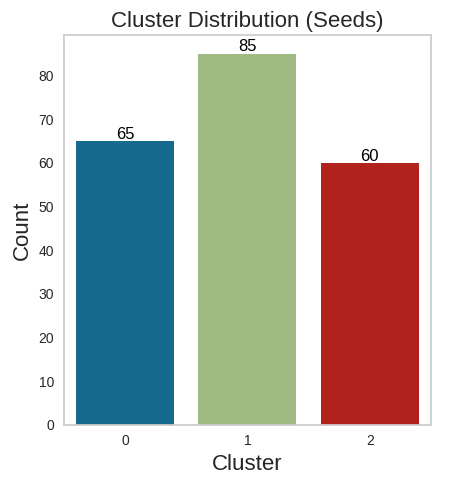
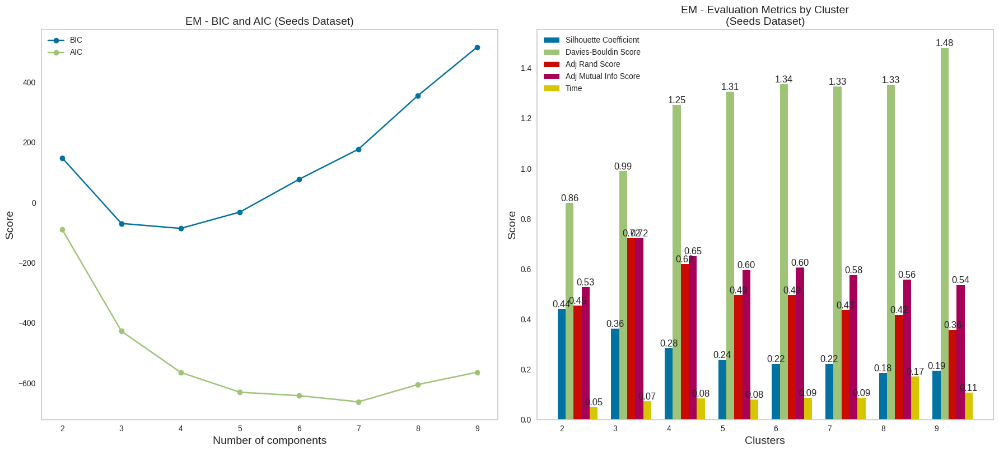
We calculate and plot the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) to choose the value of k. Lower BIC and AIC values indicate better models. Look for the K value where BIC and AIC values start to level off or decrease more slowly, similar to the Elbow Method in K-Means. This point can be a good estimate of the optimal number of components.

In determining the optimal number of components for Gaussian Mixture Model (GMM) clustering, we rely on the Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC), which are meant to be minimized. After analyzing the BIC and AIC charts, the most appropriate cluster numbers are selected based on the low points in the charts.

In the case of the Wisconsin dataset, the analysis of Bayesian Information Criterion (BIC) and Akaike Information Criterion (AIC) charts reveals a clear optimal choice of 2 clusters, as indicated by the lowest BIC score. This decision aligns well with the true nature of the data. Additionally, other evaluation metrics, including the Silhouette Coefficient, Davies-Bouldin Score, Adjusted Rand Score, and Adjusted Mutual Information Score, all consistently point towards 2 clusters as the optimal configuration. This collective agreement among metrics highlights that 2 clusters provide an effective balance between data clustering and model simplicity for the Wisconsin dataset.



Conversely, the analysis of the Seeds dataset using BIC and AIC charts demonstrates an inverted BIC graph, reflecting the presence of well-defined 3 labels in the data. This reinforces the selection of 3 clusters as the most appropriate choice, in line with the true class labels. The same confirmation is observed when considering additional evaluation metrics, such as the Silhouette Coefficient, Davies-Bouldin Score, Adjusted Rand Score, and Adjusted Mutual Information Score, all of which consistently advocate for 3 clusters as the optimal configuration for the Seeds dataset. This collective evidence underscores the strong suitability of 3 clusters for effectively capturing the dataset's inherent structure.

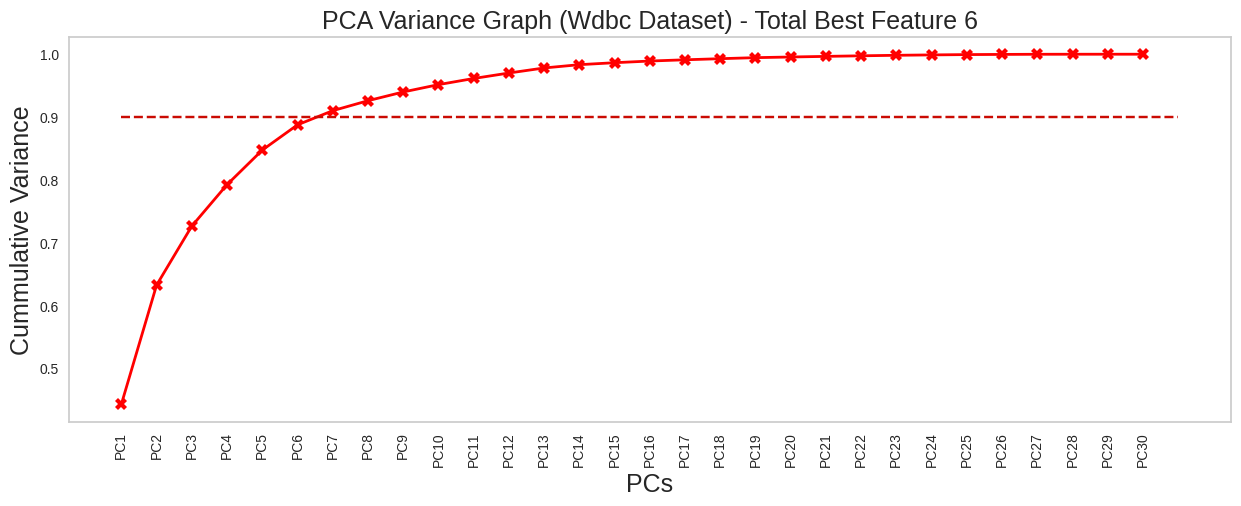


1. **DIMENSIONALITY REDUCTION**

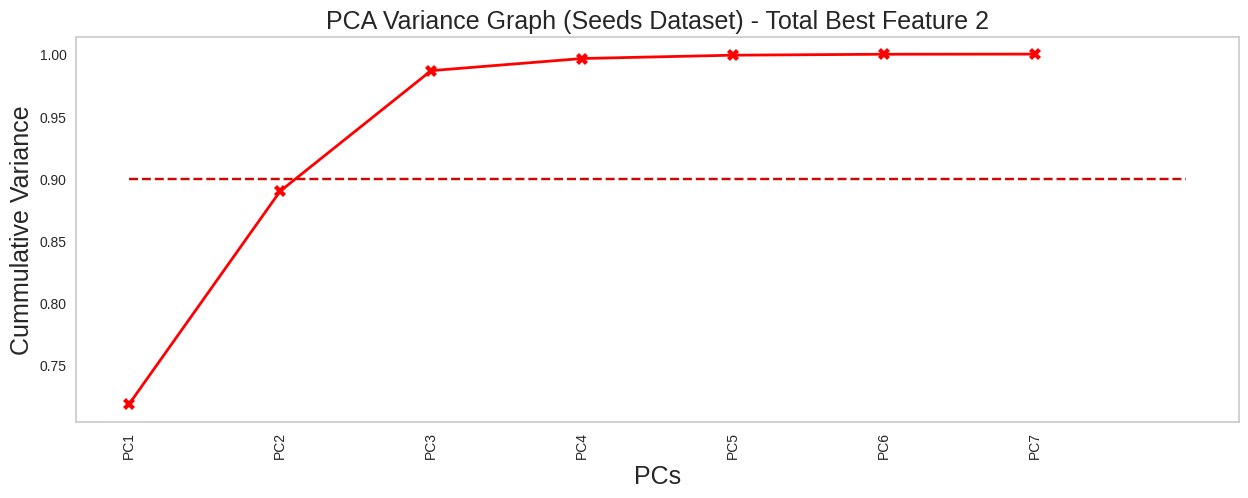
**3.1. Principal Component Analysis (PCA)**

Principal Component Analysis (PCA) is a dimensionality reduction technique that helps to reduce the number of features (dimensions) in a dataset while preserving the most critical information. I've created plots to illustrate the cumulative variance explained by increasing numbers of principal components and the individual variance explained by each component. I've selected the number of components on the 90% threshold of the cumulative variance for further analysis.

In the context of the Wisconsin dataset, it's remarkable that a significant dimensionality reduction can be achieved by selecting just 6 principal components, capturing 90% of the dataset's variance. It’s interesting to note that only 6 out of 30 are below the threshold due to the high proportion of variance that can be explained by a smaller subset of features.

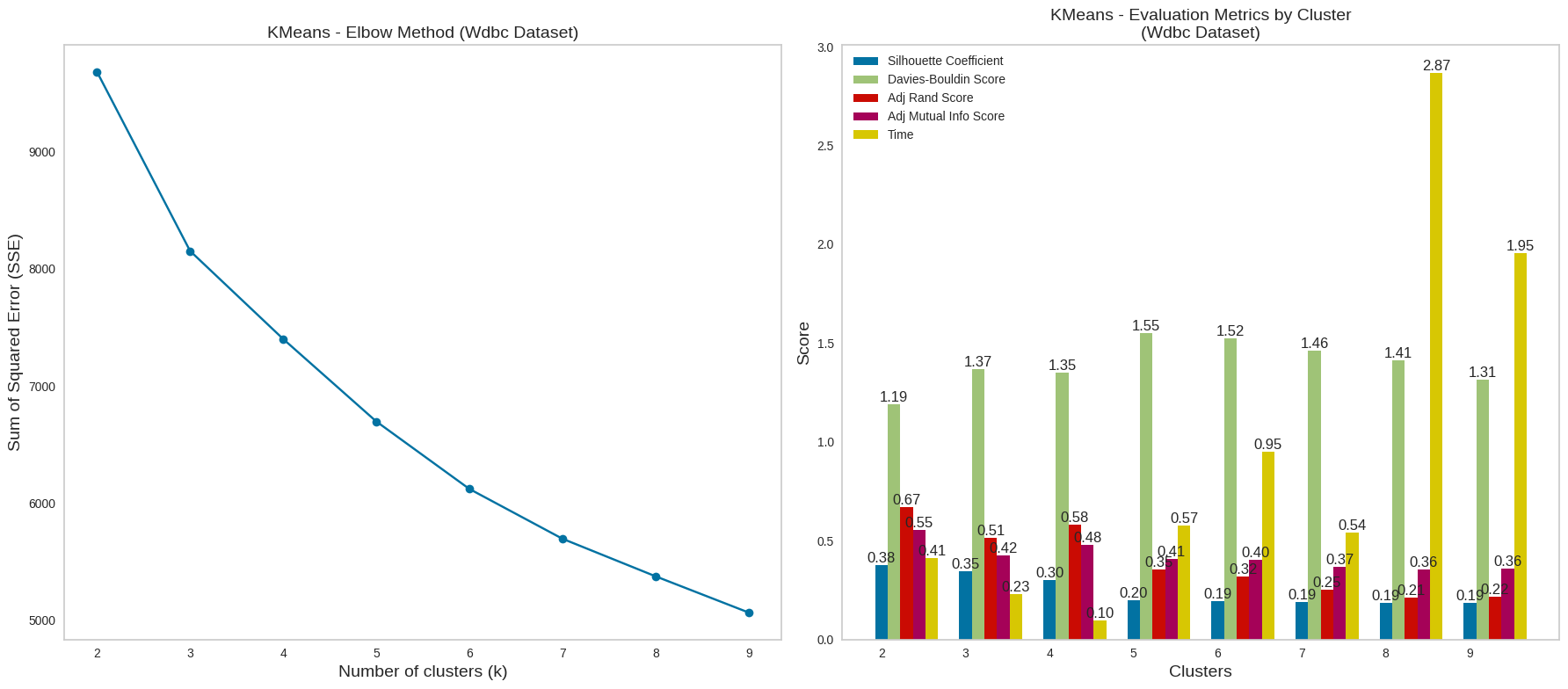


When applied to the Seeds dataset, it's striking that a mere 2 principal components suffice to capture 90% of the dataset's variance. This finding highlights the inherent redundancy in the Seeds dataset and PCA's ability to efficiently represent its underlying structure with a minimal number of features.

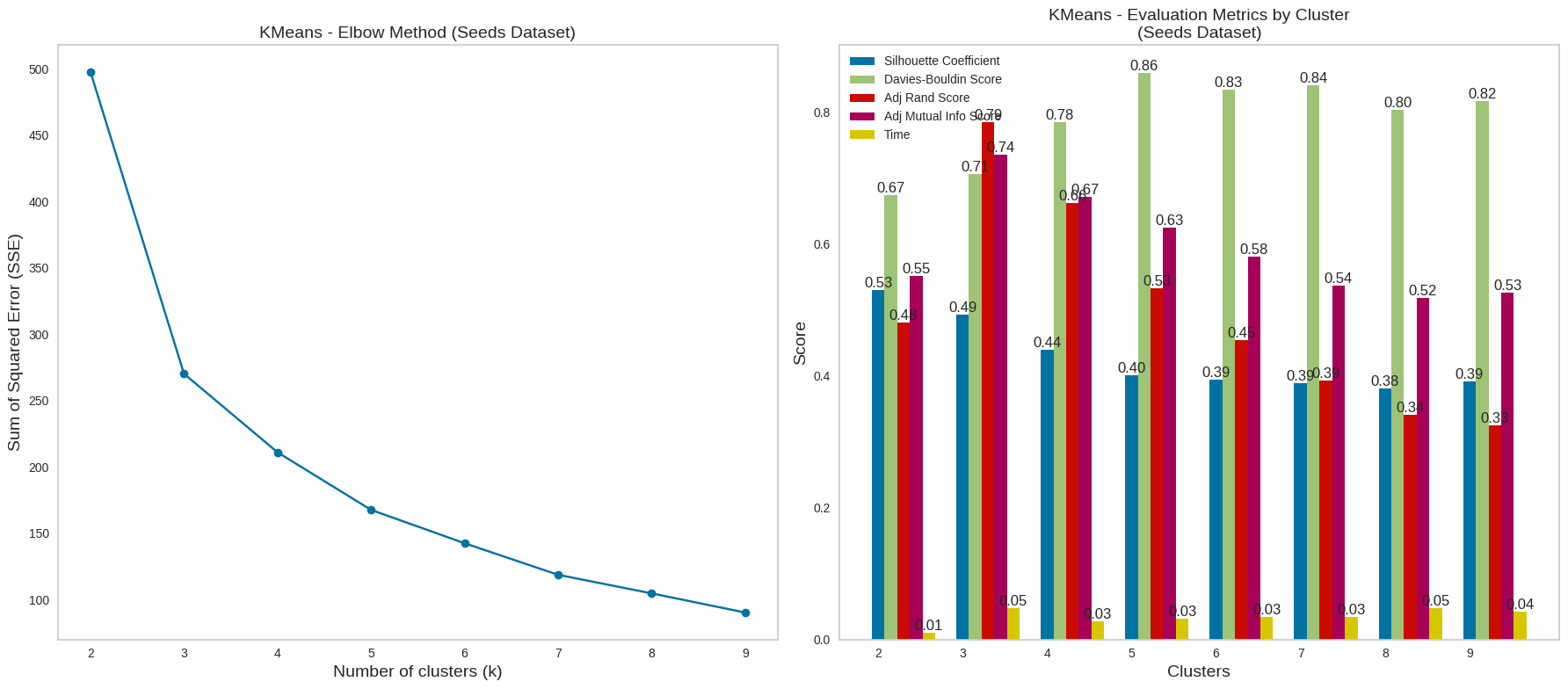


**PCA with K means**

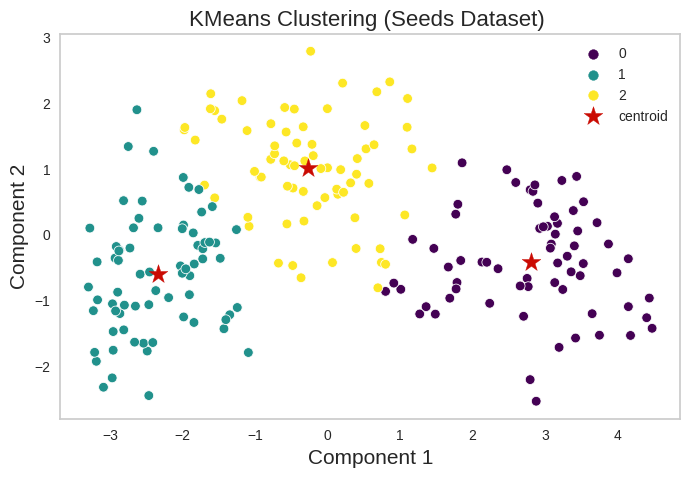
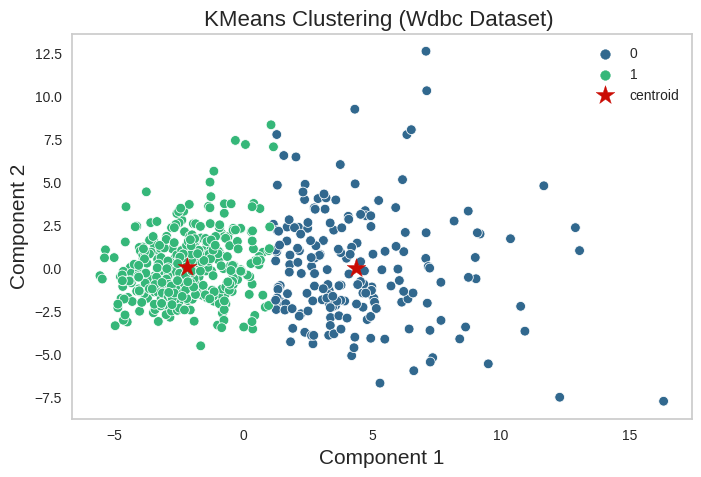
For the Wdbc dataset, it’s observed that number of optimal clusters K=2 still performing well because of its highest silhouette score of 0.38, Adj Rand Score of 0.67, and Adj Mutual Info Score of 0.55 as compared to others as illustrated below. It also has the lowest Davies-Bouldin Score which is better when it is lower.



For the seeds dataset, the optimal number of clusters is still K=3 because of its value of silhouette score of 0.49 that is also good but Rand and Mutual information score are best and lowest value of Davies-Bouldin which indicates PCA performing efficiently.

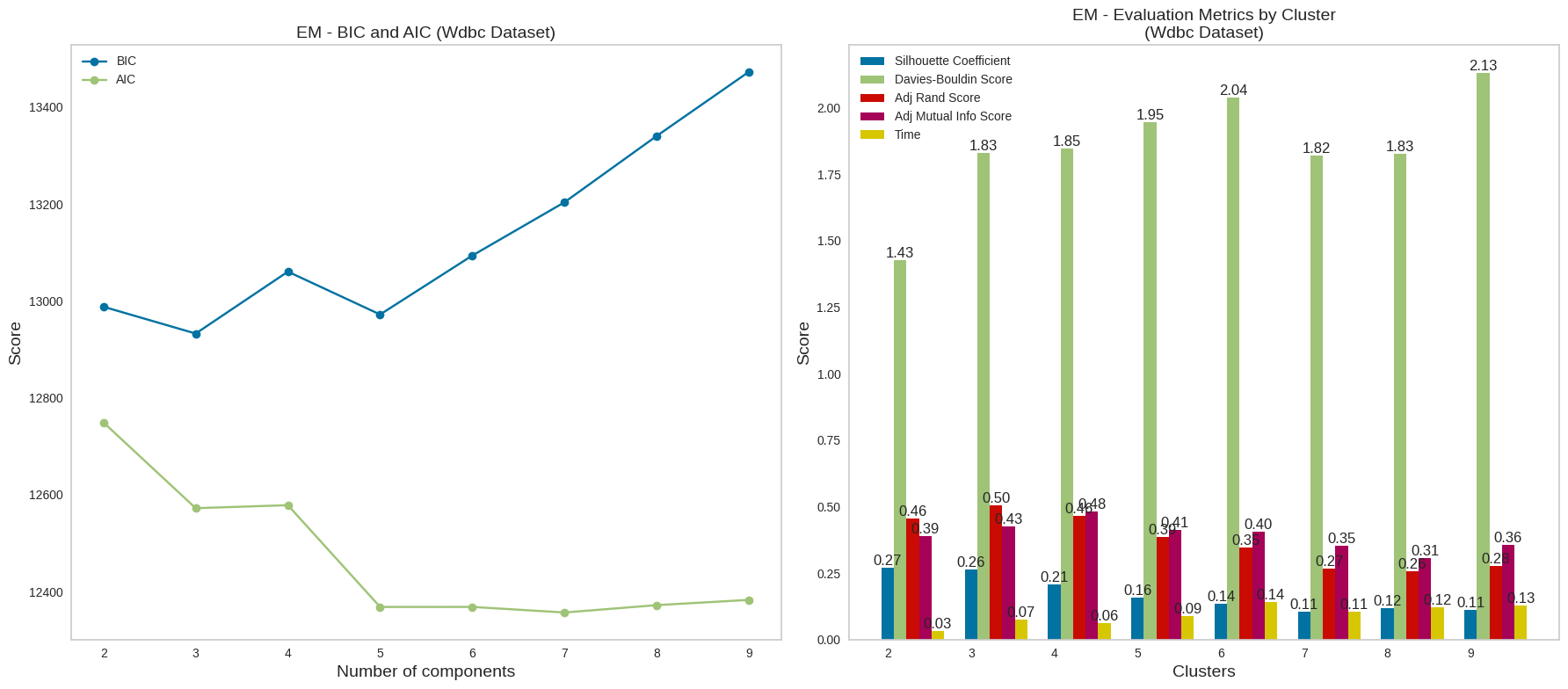


The scatter plot for both data sets reveals a notably limited presence of outliers. This means that there are only a few data points that deviate significantly from the general pattern or trend observed in the data.

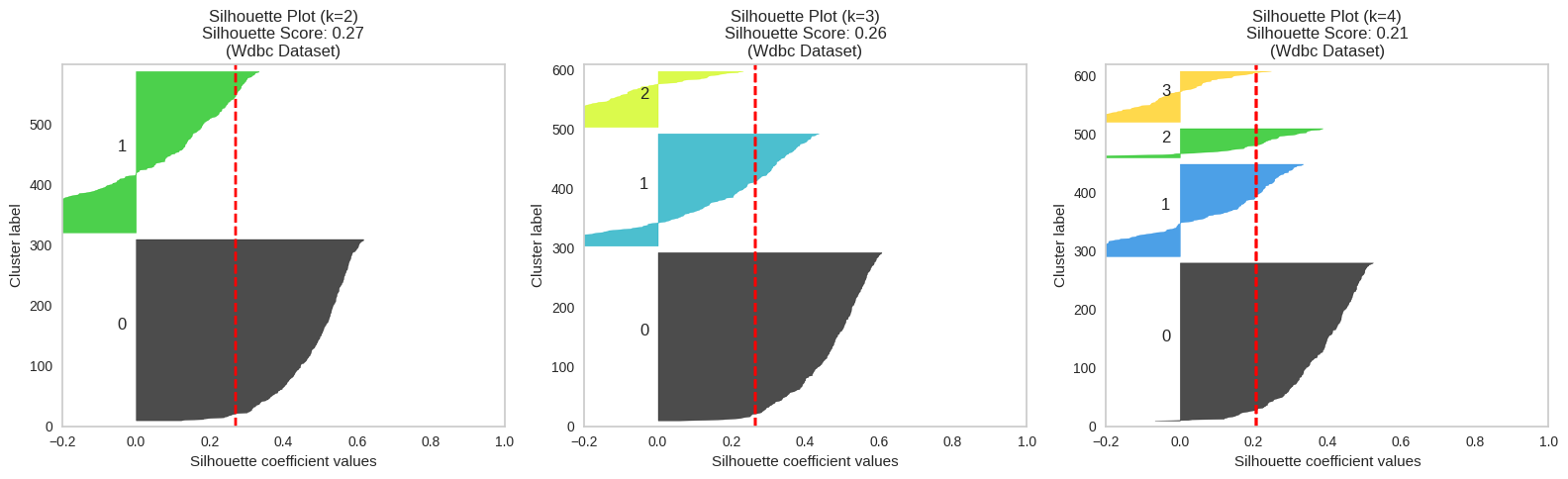


**PCA with EM**

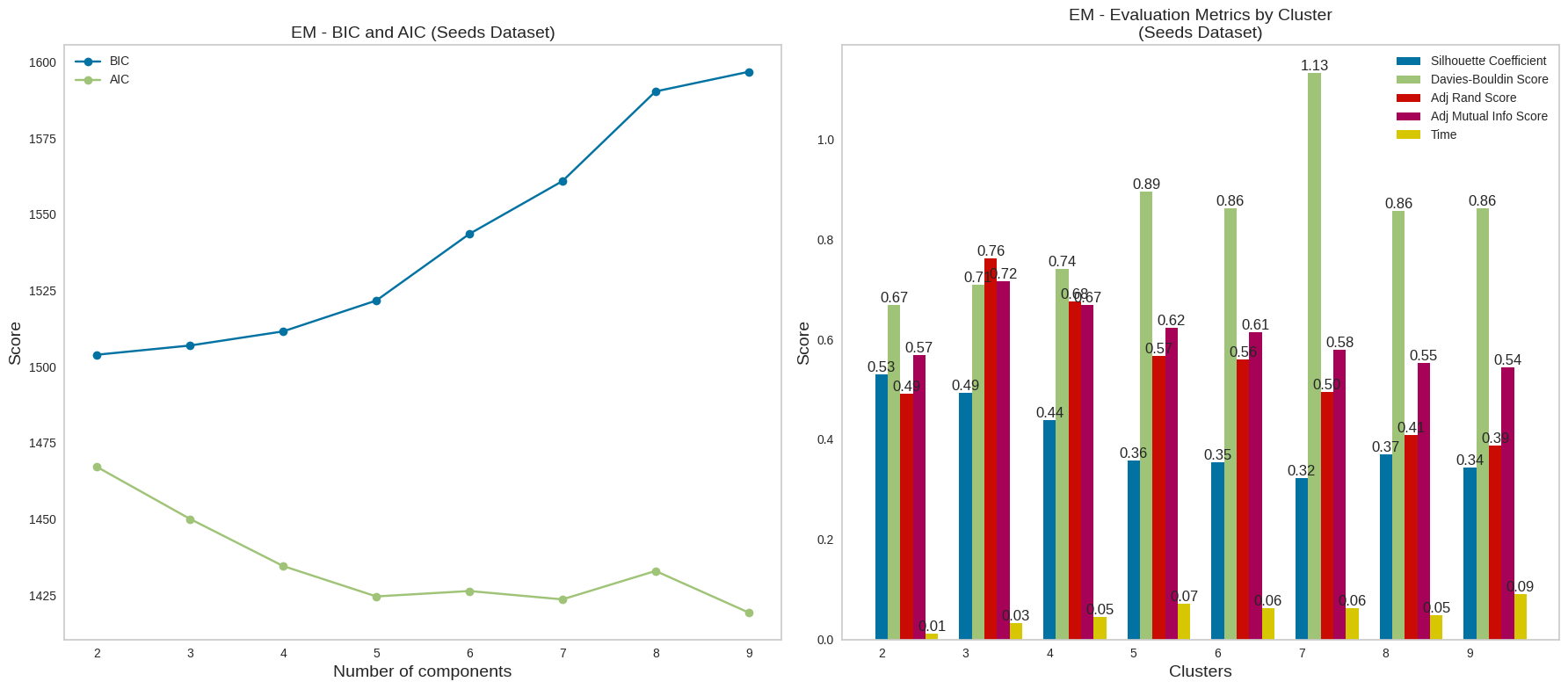
In the case of the 'Wdbc' dataset, we assessed the clustering quality for various values of K in which the silhouette score for the value of K=2 value is 0.27 which is the highest as compared to others and its lowest Davies-Bouldin score of 1.43 justified that number of cluster K=2 in reduced dimensionally as shown below.

****

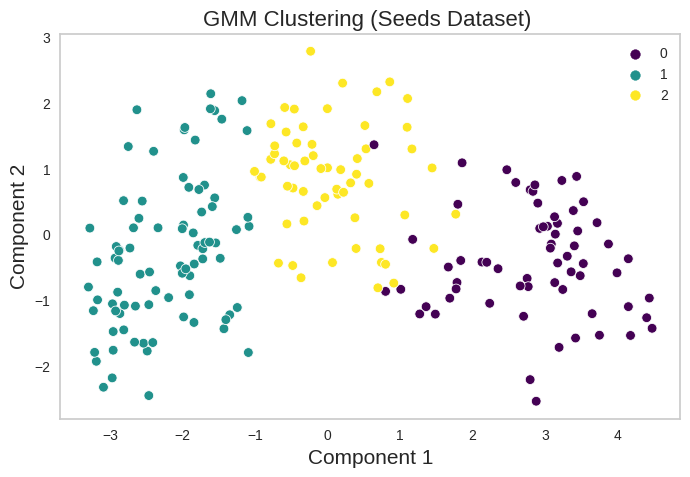
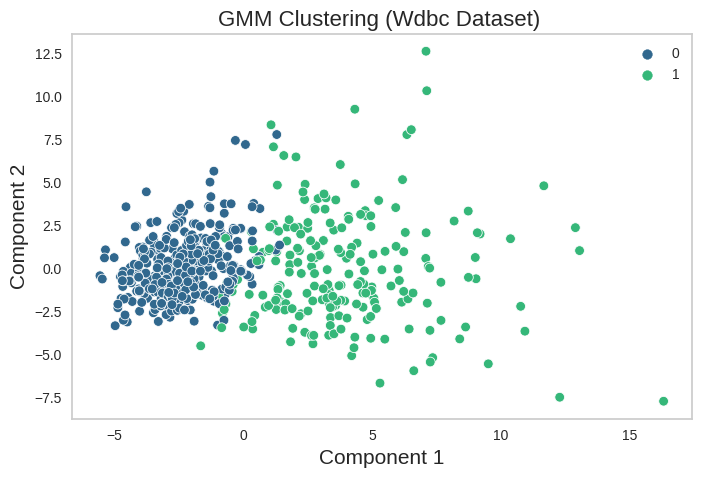
We observe variations in the silhouette score as K takes on different values. Specifically, when K equals 2, the silhouette score is highest at 0.27, whereas for K equal to 5, the score is 0.21, as depicted below.

****

On the Seeds data set, our analysis revealed fluctuations in the silhouette score across varying K values. Notably, the optimal number of K=3 is because of their high Adj Rand and Adj Mutual info score. In that case Davies-Bouldin and are silhouette score extremely close on the graph.



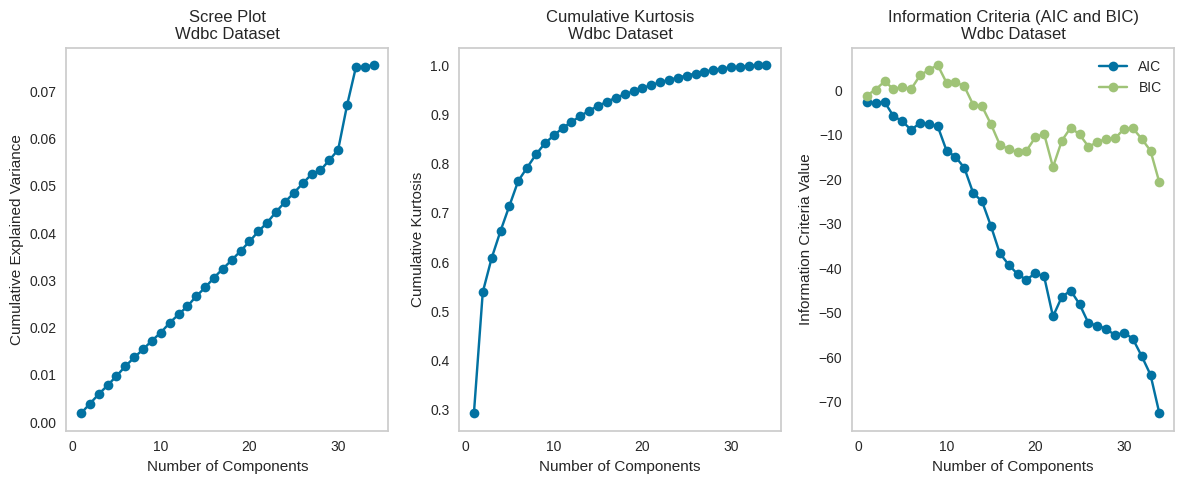
The scatter plot for both data sets not only indicates that the clusters are correctly distributed but also provides strong evidence of the data's inherent structure and the presence of distinct, discernible patterns within the dataset.

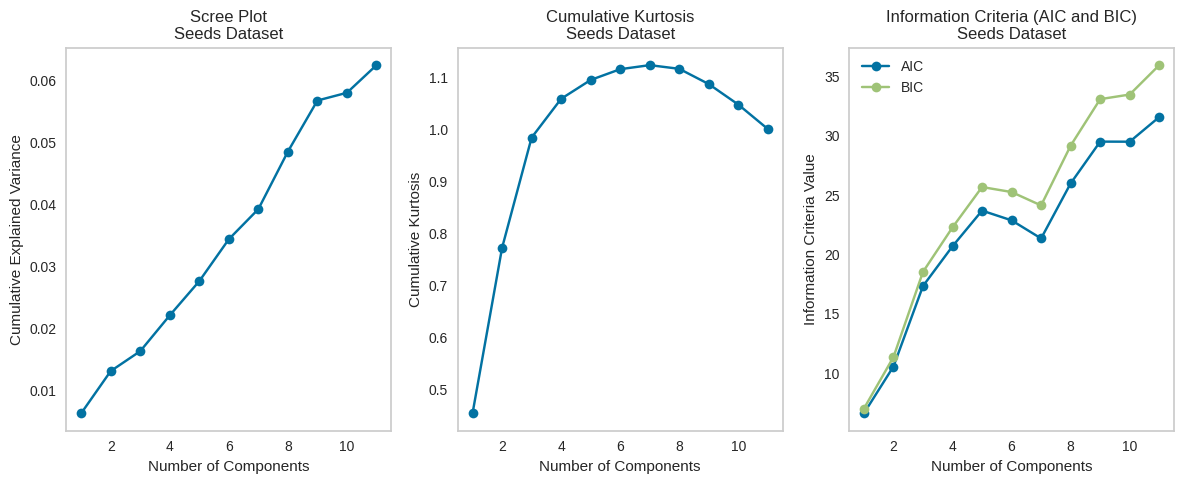


* 1. **Independent Component Analysis (ICA)**

ICA (Independent Component Analysis) decomposes the original features into statistically independent components. The model's evaluation involves assessing non-gaussianity, which is achieved by maximizing kurtosis. In this analysis, the sklearn FastICA algorithm is employed to carry out the decomposition and evaluation.

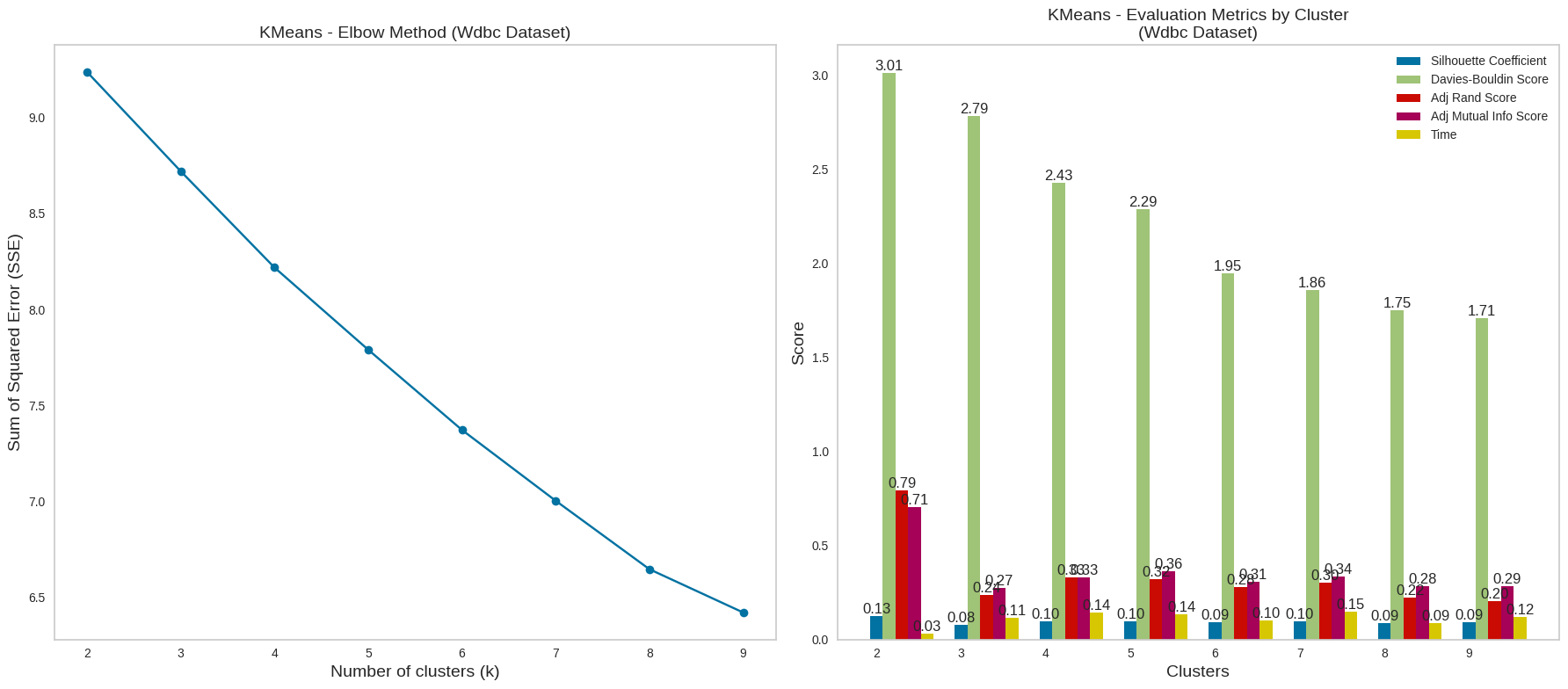
For the Wisconsin dataset, a suggested choice for the number of components is 10 based on kurtosis analysis, as it best encapsulates the non-Gaussian and independent features within the dataset, making it conducive for dimensionality reduction or further analysis. In the case of the Seeds dataset, kurtosis analysis points to 4 as the optimal number of components, effectively capturing the key non-Gaussian and independent attributes of the data, and facilitating efficient feature reduction or analysis.



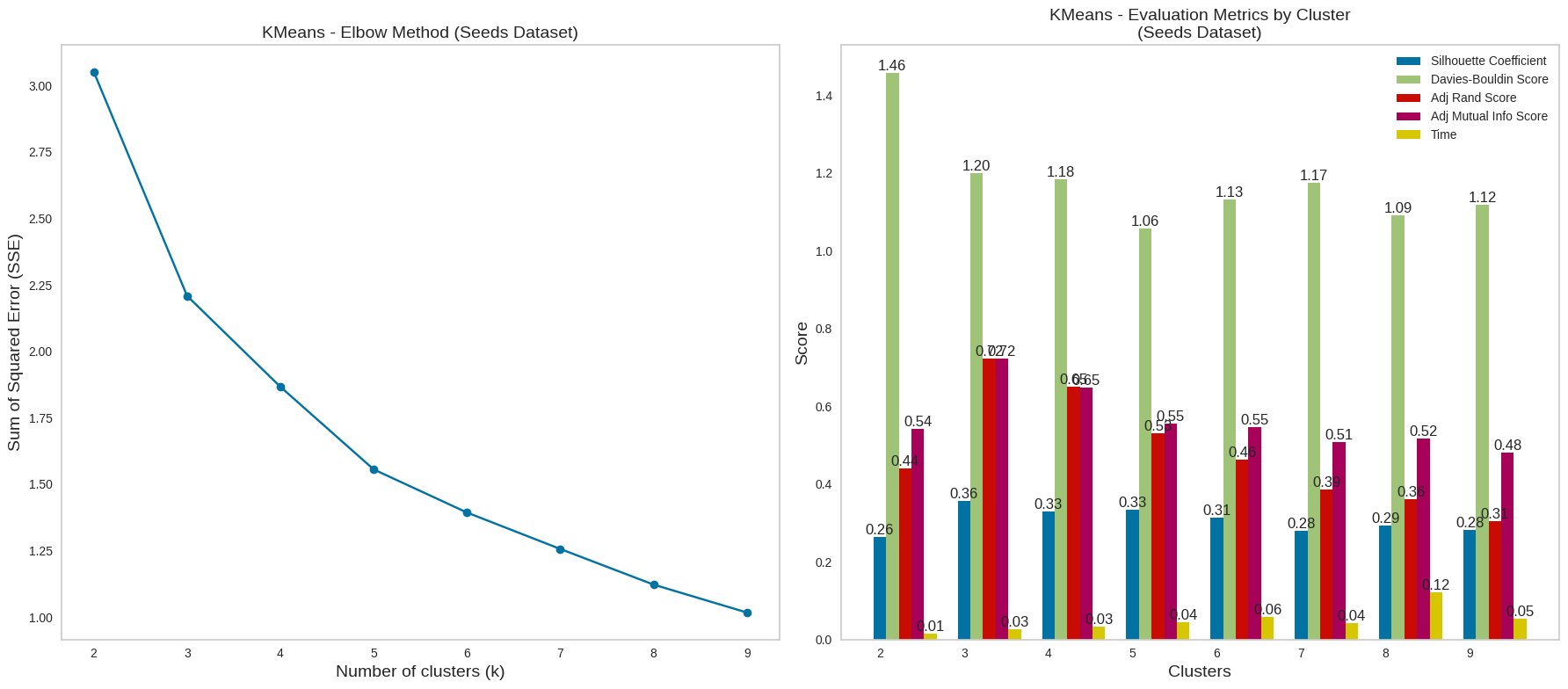


**ICA with K-Means**

In the context of the 'Wdbc' dataset, we examined the quality of clustering across different values of K, representing the number of clusters, using the silhouette score. In our analysis, when K is set to 2 it achieves the highest silhouette score is 0.13, the highest adj rand score of 0.79, and the Adj mutual info score of 0.71 as exemplified in the below given graph below.

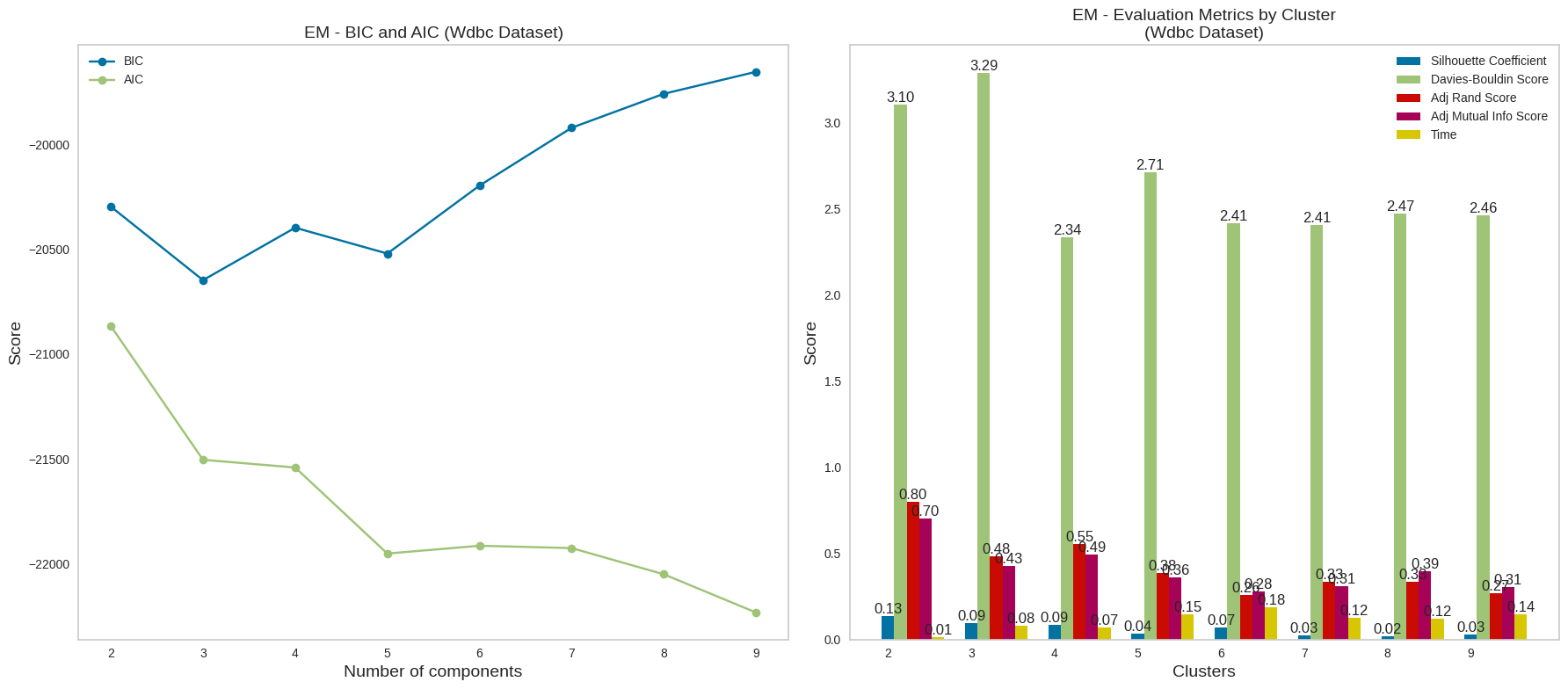


When applied to the Seeds dataset, our examination unveiled variations in the silhouette score when K=3, the silhouette score is 0.36, the Adj rand score is 0.72, and the Adj Mutual info score also 0.72 which is highest among others as illustrated below. These are the reasons for selecting the K=3 in the seeds dataset.

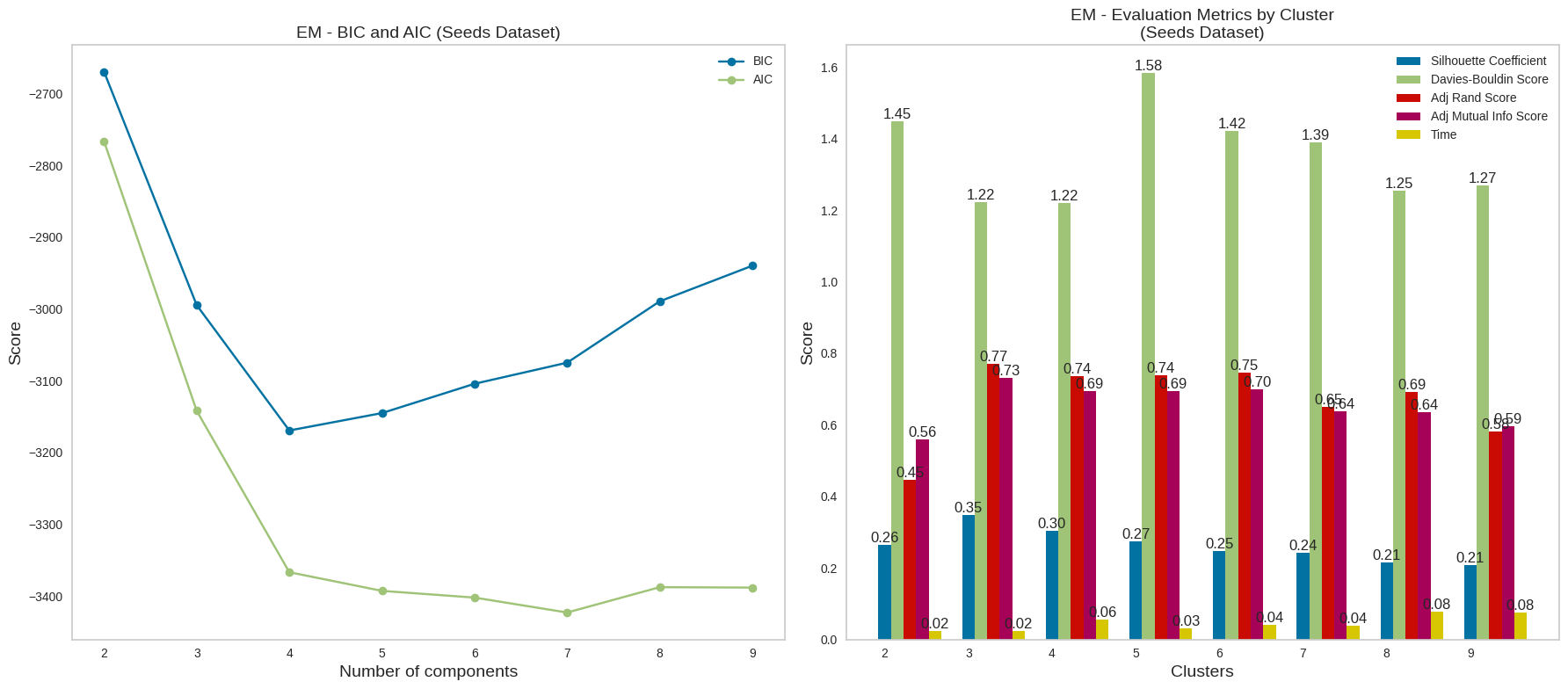


**ICA with EM**

In the case of the Wdbc dataset, the silhouette score is notably high when K=2 which is 0.13. Adj rand score is 0.80 and Adj Mutual info is 0.70 are also the highest among others which also justify K=2.



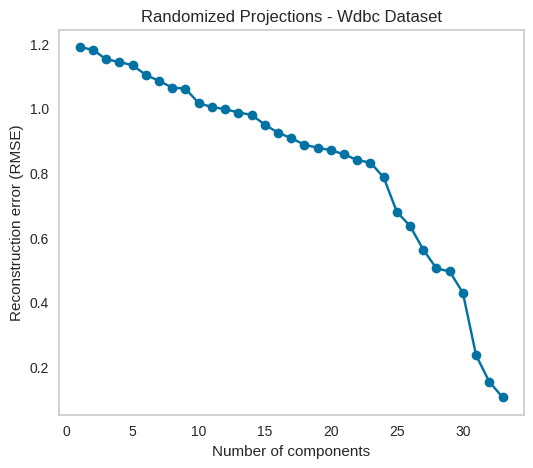
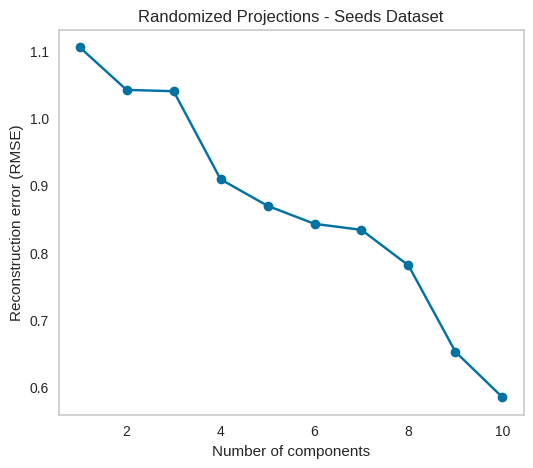
Regarding the Seeds dataset, the silhouette score is high at 0.35 when K=3. However, as the number of clusters (K) is increased the silhouette score falls. The score Adj rand score and Adj Mutual info are also high.



* 1. **Randomized Projections (RP)**

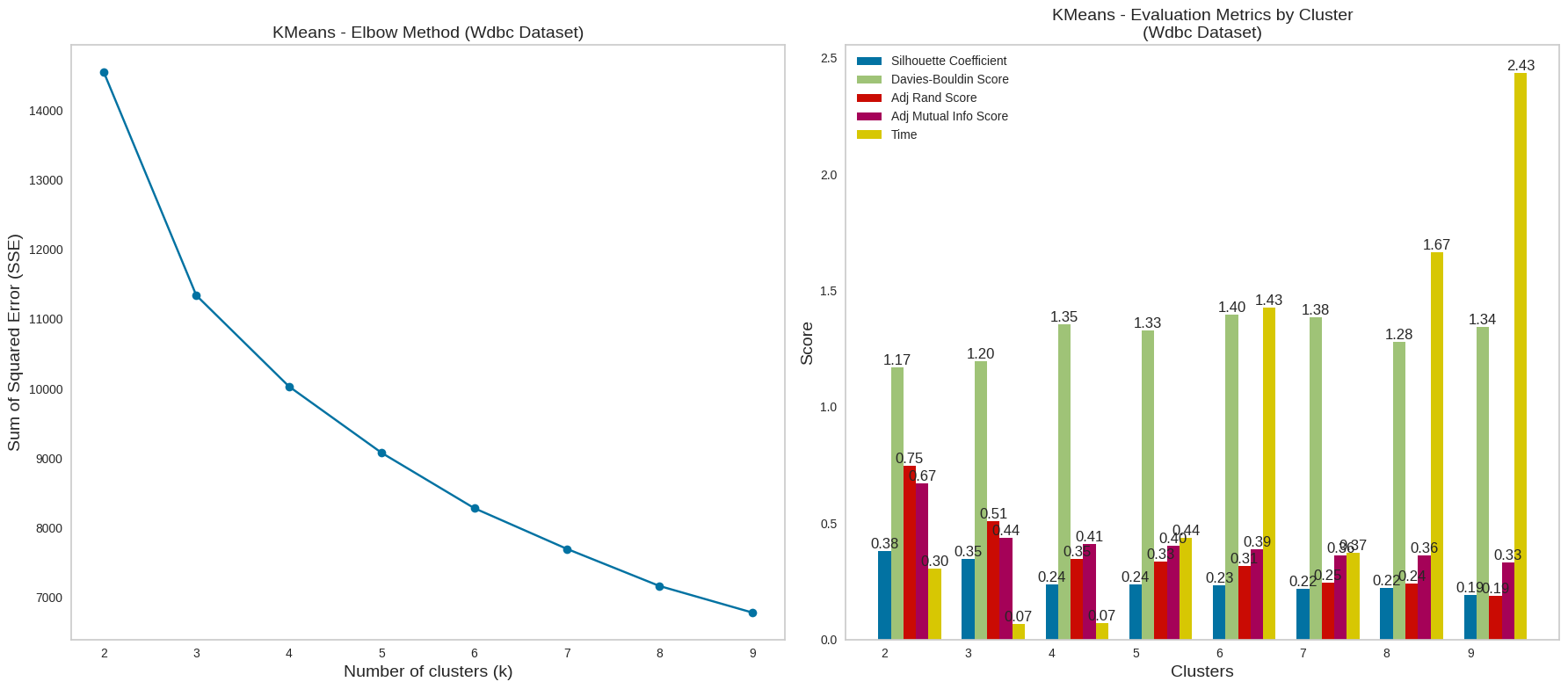
Randomized Projections (RP) is a dimensionality reduction technique used in machine learning and data analysis. It is primarily employed for reducing the dimensionality of high-dimensional data while preserving important information and structure.

For the Wisconsin dataset, around 10 components provide effective dimensionality reduction, as the reconstruction error levels off towards zero at 30 components. In the case of the Seeds dataset, about 3 components suffice for dimensionality reduction, with the reconstruction error towards zero at 10 number of components.

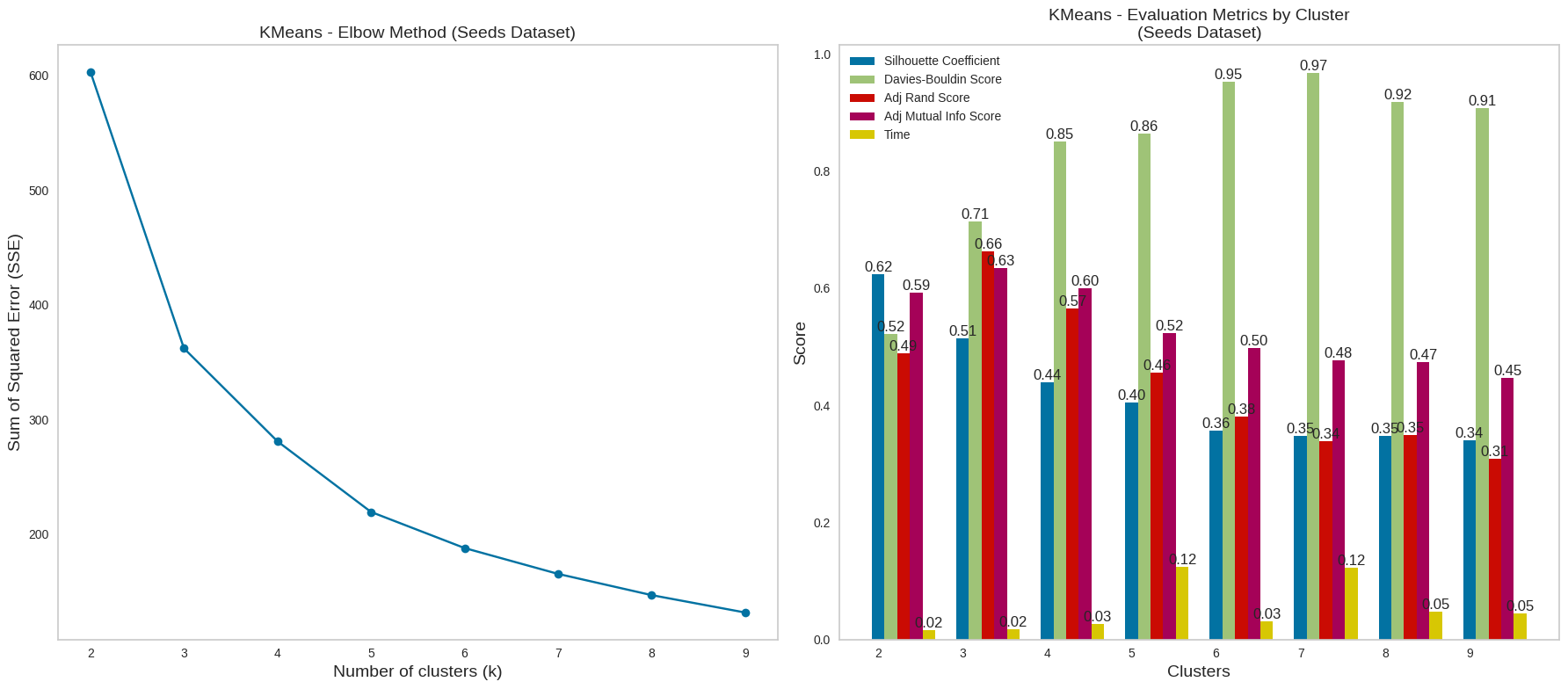


**RP with K-Means**

In the case of the Wdbc dataset, the silhouette score for the specified values of K is as follows: When K=2, the silhouette score is 0.38 which is the highest. The same is the case of the Adj rand score of 0.67 and adj mutual info score of 0.75 score which is the highest. It also has the lowest Davies-Bouldin score of 1.17.

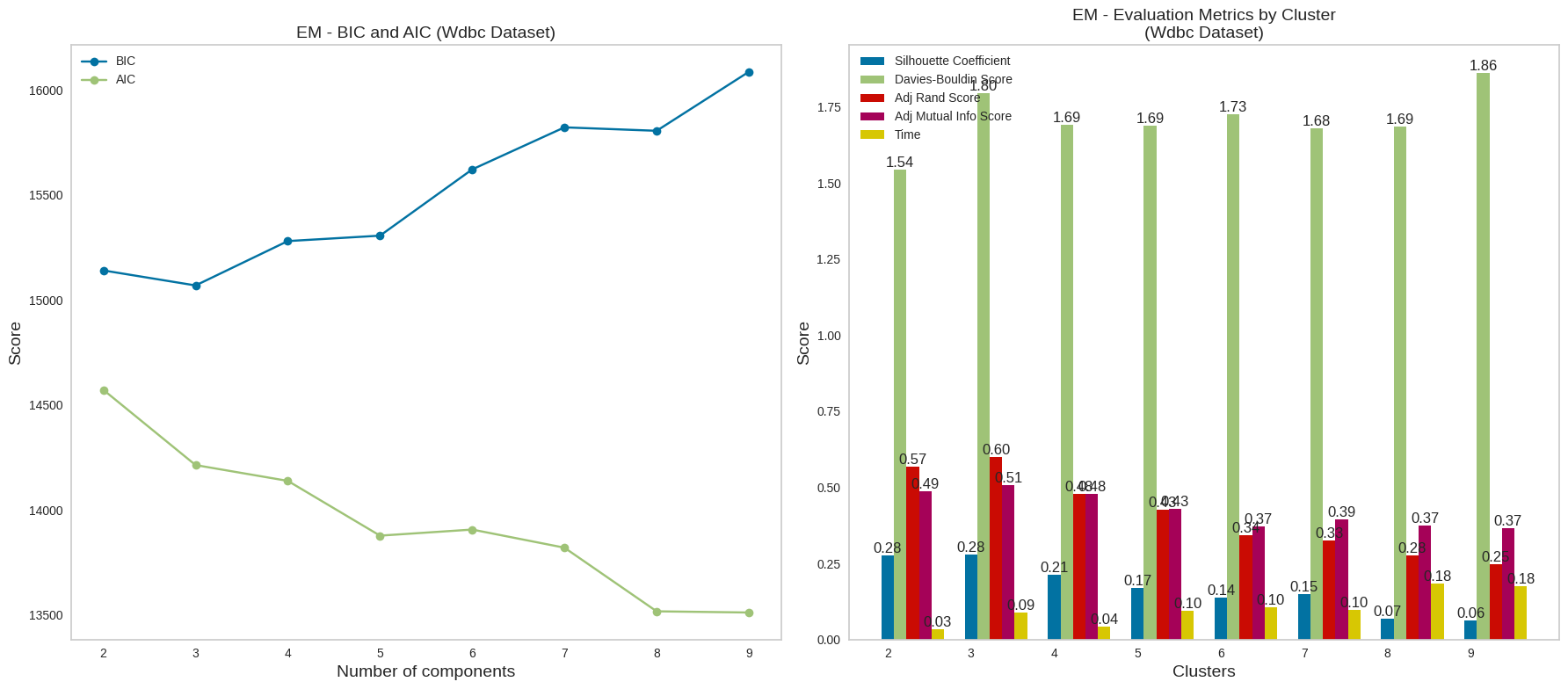
****

In the seeds dataset, with K=3, the silhouette score is 0.51. Adj rand score is 0.66 and Adj Mutual info is 0.63 which is the highest among others.

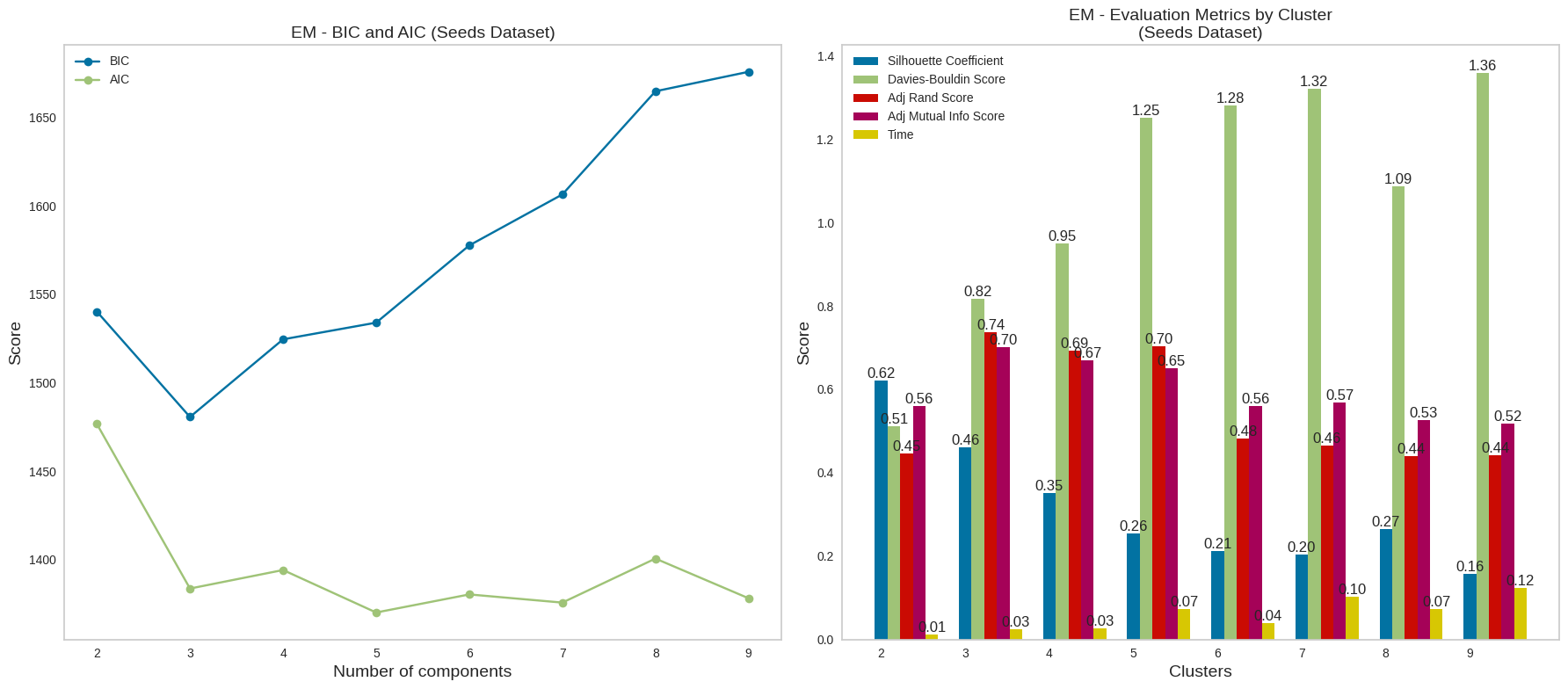


**RP with EM**

In the EM case, in wbdc dataset case show when K=2 the silhouette score is 0.28 which is the highest among other as illustrated below.

****

In seeds K=3, the Adj and score are 0.74, and Adj Mutual info is 0.70 which is higher than the other. So in this case we select the value of K.

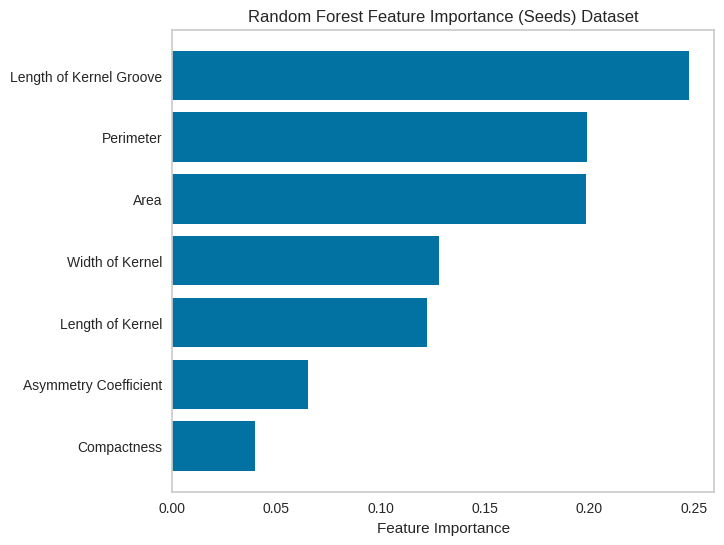
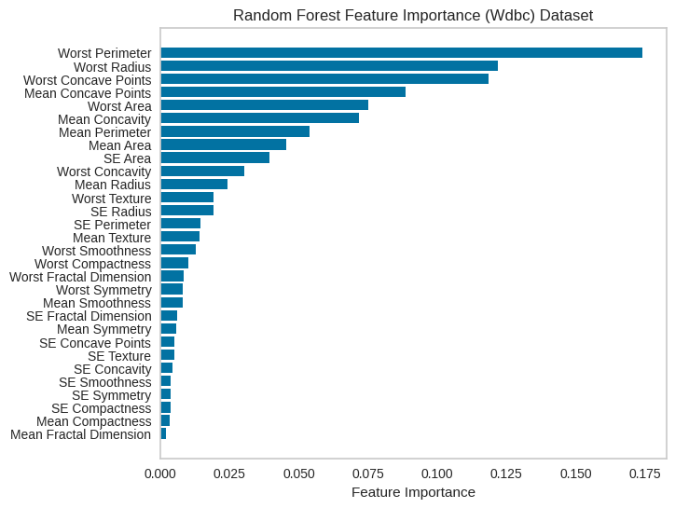
****

1. **Random Forest Feature Selection (RF)**

Random Forest Feature Selection is a process of determining the most important features in a dataset by utilizing a Random Forest model. It helps identify and retain the most influential variables for better predictive performance and data analysis. I've employed the sklearn library to perform feature selection based on feature importance scores exceeding a threshold, typically set at the mean value of these scores.

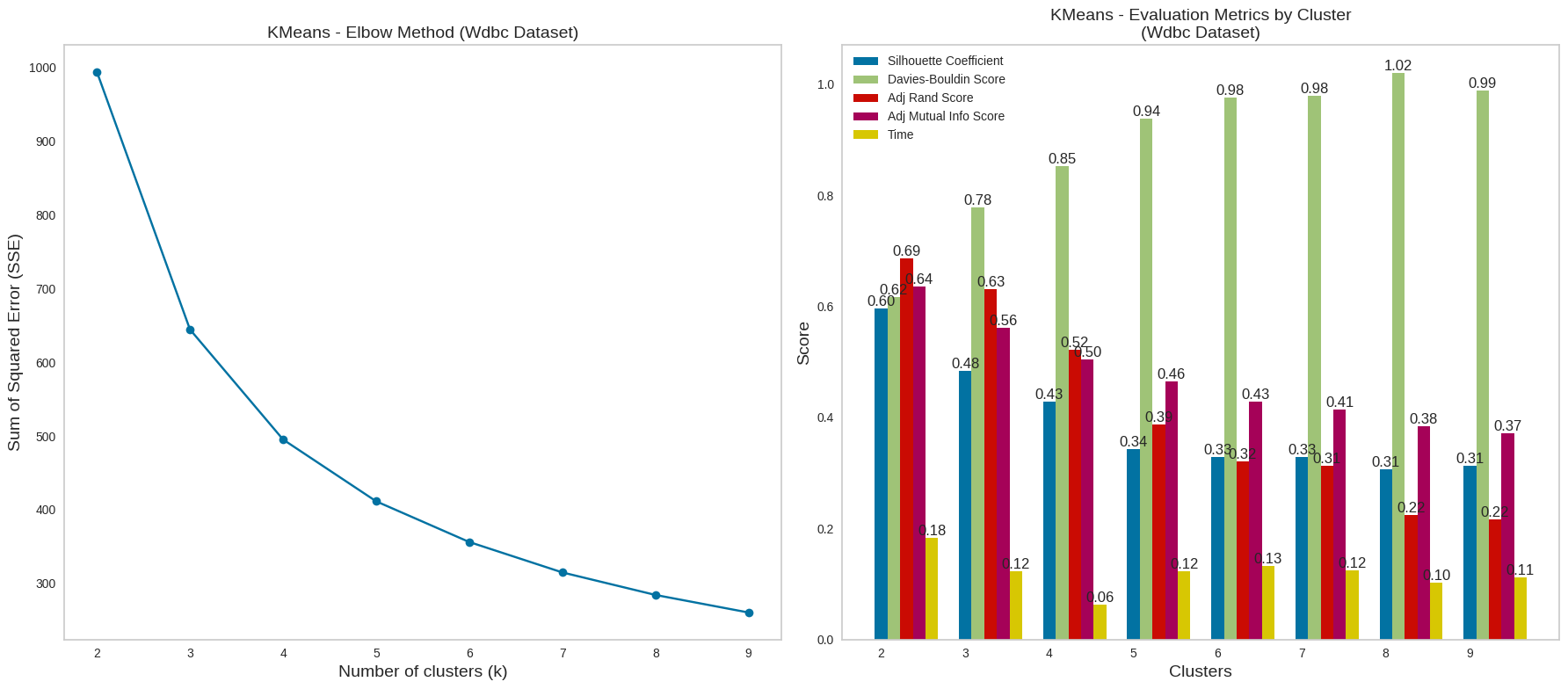
In the case of the Wisconsin dataset, the top 5 features with the highest feature importance are chosen, including 'Worst Perimeter,' 'Worst Radius,' 'Worst Concave Points,' 'Mean Concave Points,' and 'Worst Area.'

Regarding the Seeds dataset, the feature selection process identifies the first 3 features with the highest feature importance, namely 'Length of Kernel Groove,' 'Perimeter,' and 'Area.'

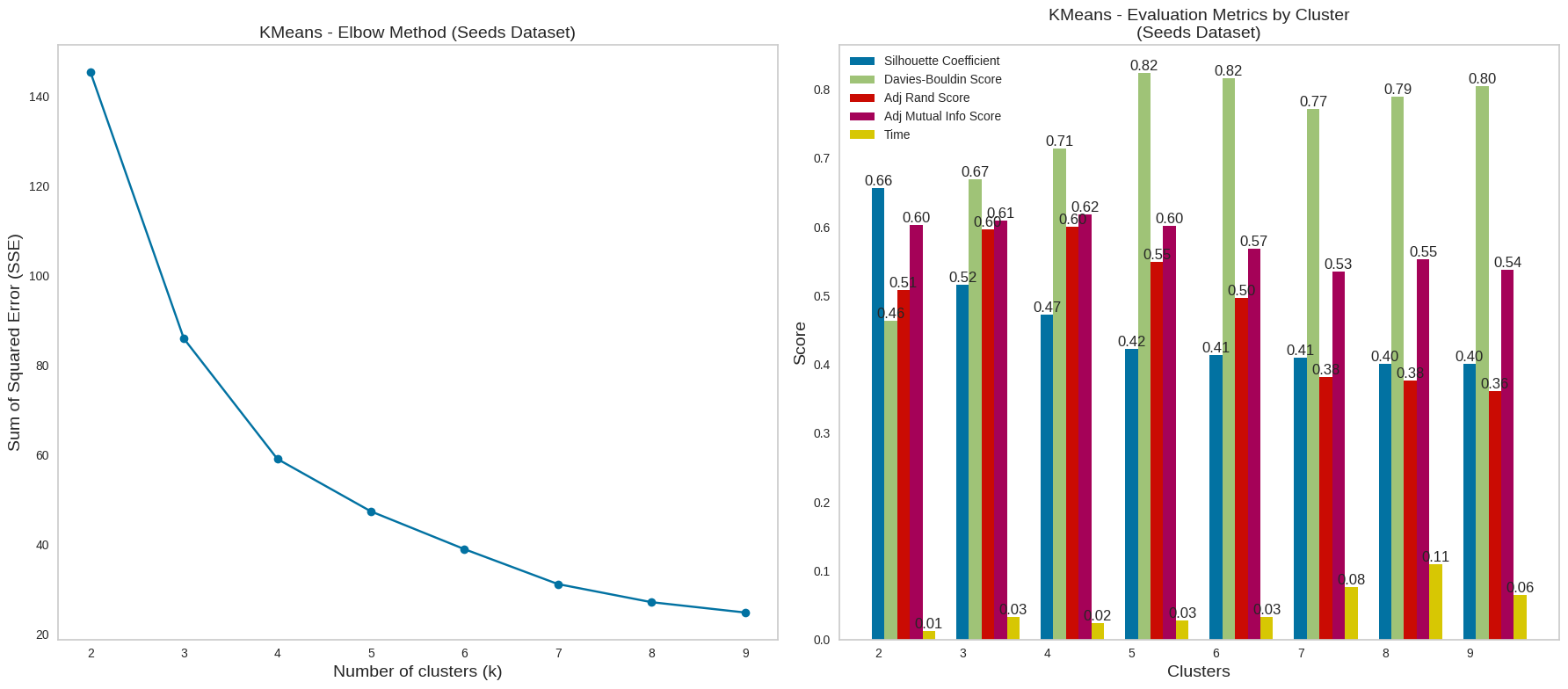


**RF With K-Means**

For the wdbc dataset, K=2, its highest silhouette score is 0.60, the highest Adj Rand score is 0.69, and the Adj mutual info score. Its lowest Davies-Bouldin score of 0.62 justified the number of cluster K=2 in reduced dimensionally as shown below

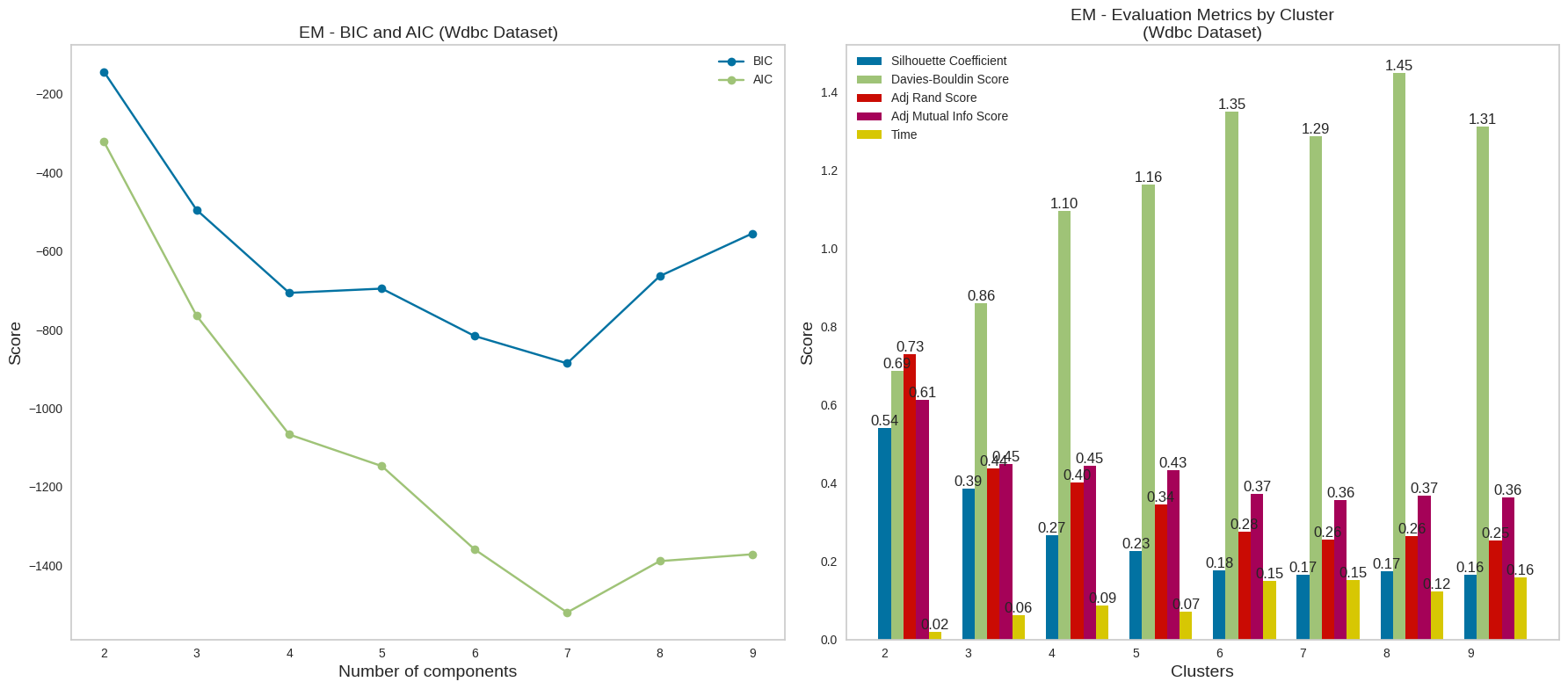
****

In the seeds dataset, When K=3, the silhouette score is 0.52, indicating strong cluster separation. However, it also high Adj Rand score which is 0.60, and an Adj mutual info score of 0.61.

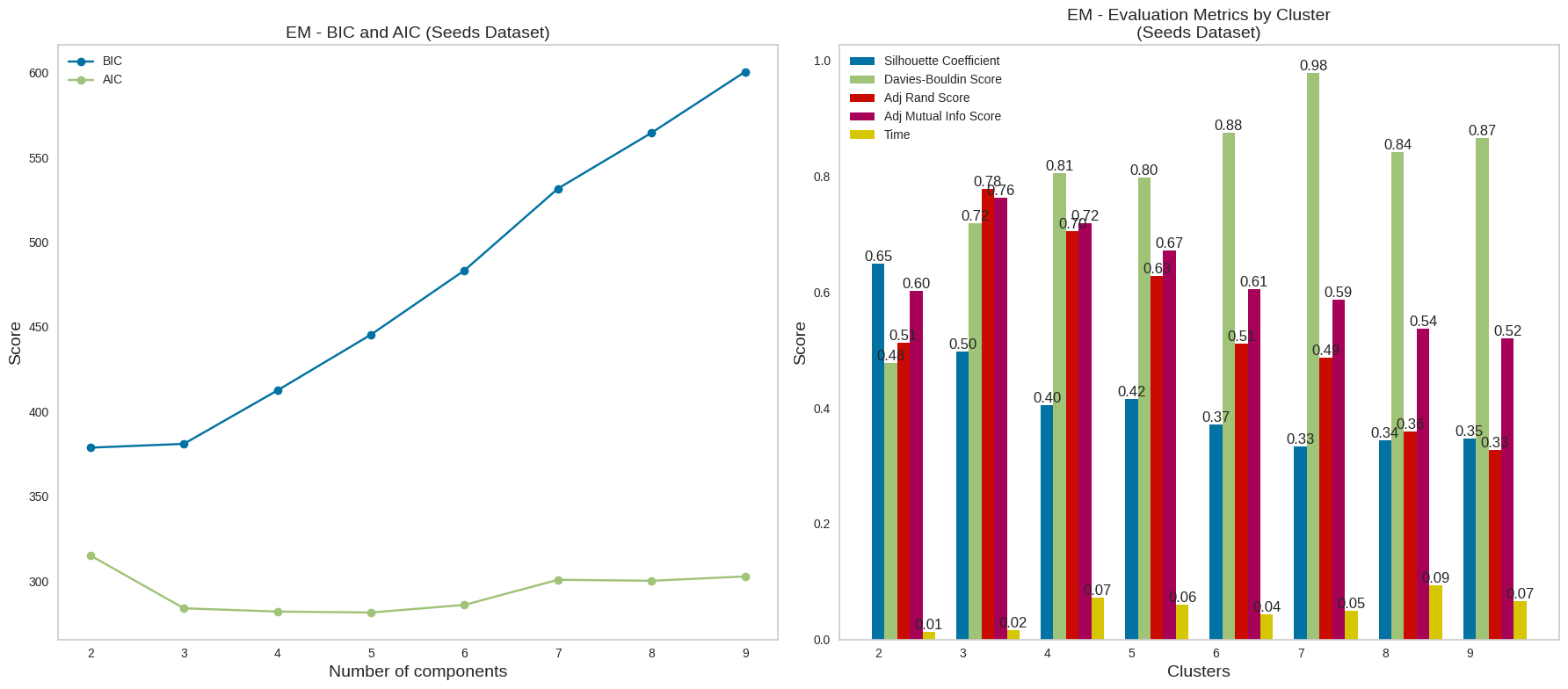
****

**RF With EM**

For wdbc data sets the value of the silhouette score number is the highest among others as illustrated below. It’s also Davies-Bouldin's low score which is an indicator of a good cluster chosen.

****

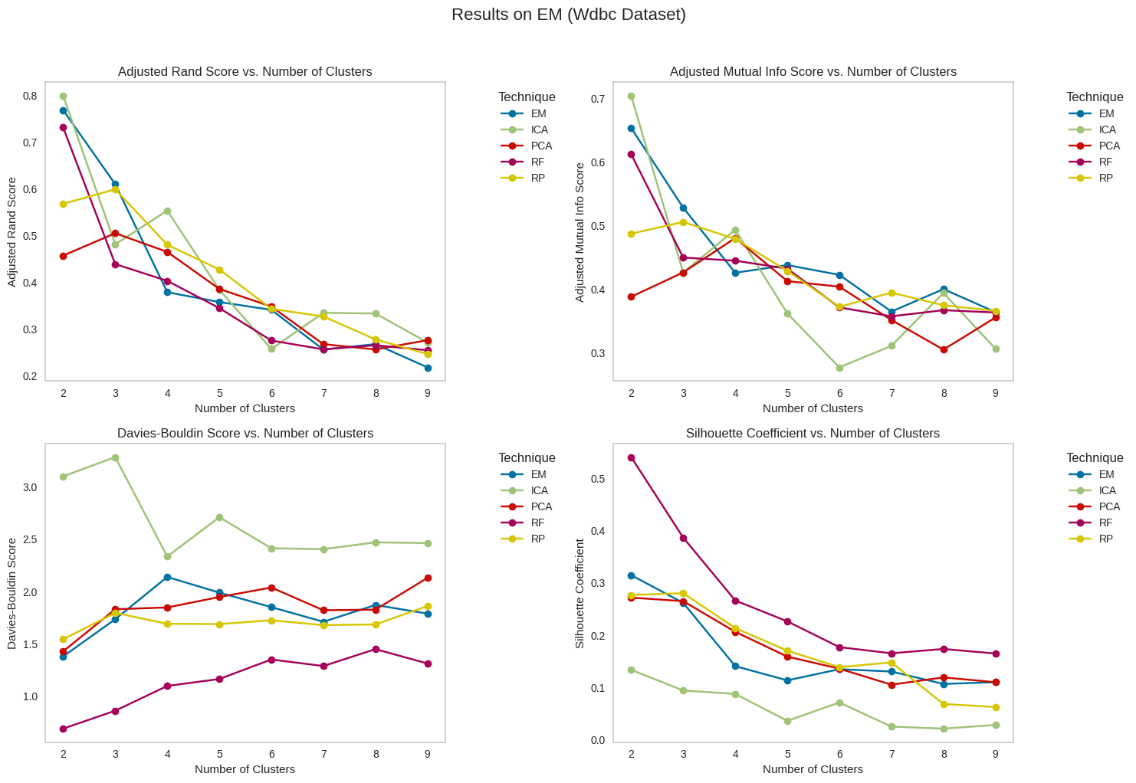
When we applied this method to the seeds dataset using K values of 3, we achieved a remarkable Adjusted Rand score of 0.78 and an Adjacent Mutual Information score of 0.76. These exceptional performance metrics underscore the rationale behind selecting K=3 as the optimal value.

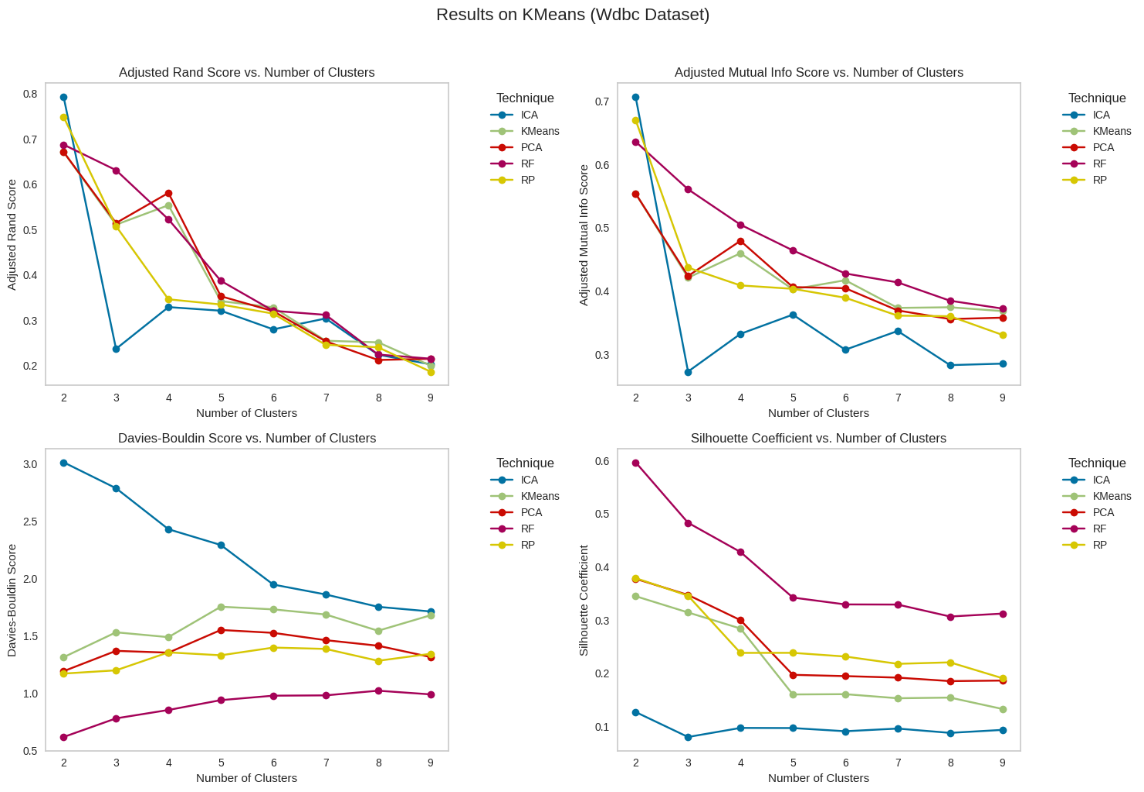
****

1. **All Reduction Technique Analysis**

The analysis reveals that Independent Component Analysis (ICA) and Random Forest (RF) have demonstrated notable performance when clustering the Wisconsin dataset into two distinct groups, closely aligning with the original KMeans labels. This suggests that ICA and RF are effective techniques for this specific clustering task, underlining their suitability for similar applications.

In the context of the Seeds dataset with three clusters, both Independent Component Analysis (ICA) and Principal Component Analysis (PCA) have exhibited strong performance. These techniques have proven to be particularly effective in clustering the dataset into its intended three distinct groups, emphasizing their utility in handling the dataset's complex structure and facilitating meaningful categorization.

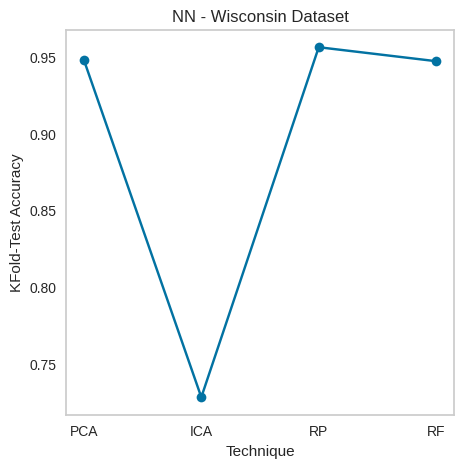




1. **Neural Network With Dimensionality Reduction**

For these experiments, I utilized both the Seeds and Wdbc datasets. I employed the MLPClassifier from the sklearn library, assessed the performance using accuracy scores, and divided the data into an 80-20 split for training and testing, recording both accuracy scores.

Different dimensionality reduction techniques and their associated accuracy scores on both the training (KFold-Train Acc) and testing (KFold-Test Acc) datasets. Notably, PCA and Random Projections (RP) achieved high accuracy scores on both the training and testing sets, with values around 0.99, indicating strong performance in preserving data information. Independent Component Analysis (ICA) and Random Forest (RF) also produced respectable results, though slightly lower, with accuracy scores around 0.76 to 0.95. These findings suggest that PCA and RP may be particularly effective in maintaining data integrity during dimensionality reduction, while ICA and RF also offer competitive performance.



1. **Neural Network with Dimensionality Reduction and Clustering Labels**

In neural networks with dimensionality reduction and Clustering, Random Projections (RP) showed the highest accuracy on the training data (100%) and excellent performance on the testing data (99.13%). PCA achieved high accuracy on the testing data (96.52%). Independent Component Analysis (ICA) and Random Forest (RF) delivered slightly lower accuracy on both training and testing data but still maintained competitive performance, around 75.87% and 75.57% accuracy on the testing data, respectively.

**K-Mean**

Our evaluation of clustering quality across different datasets, such as 'Wdbc' and 'Seeds,' using the silhouette score revealed important trends. In general, as the number of clusters (K) increases, the silhouette score tends to decrease. For instance, in the 'Wdbc' dataset, when K is set to 2, a relatively high silhouette score of 0.38 is obtained, indicating well-defined clusters. However, as K increases to 3 or 5, the silhouette score decreases, suggesting less distinct clustering.

Similar patterns are observed in the 'Seeds' dataset. These findings underscore the importance of choosing an appropriate number of clusters to optimize clustering quality, as excessively dividing the data can result in less meaningful and distinct cluster assignments.

Random Projections achieve high accuracy as compared to others.

**EM**

Our evaluation of clustering quality across the 'Wdbc' and 'Seeds' datasets, using the silhouette score and varying values of K to represent the number of clusters, revealed consistent trends. In both datasets, a lower silhouette score was observed when K was initially set to 2, suggesting that the clusters were less well-defined. However, as K increased to 4, there was a general improvement in the silhouette score, indicating more distinct clusters. These results emphasize the importance of selecting the right number of clusters, as an excessive subdivision of data into smaller clusters may hinder overall clustering quality. In contrast, striking the correct balance can result in more meaningful and distinct cluster assignments.

1. **References**

* Sharma, Pulkit. “The Most Comprehensive Guide to K-Means Clustering You’ll Ever Need.” Analytics Vidhya, 19 Aug. 2019, [www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/](http://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/).
* Brownlee, Jason. “A Gentle Introduction to Expectation-Maximization (EM Algorithm).” Machine Learning Mastery, 31 Oct. 2019, machinelearningmastery.com/expectation-maximization-em-algorithm/.
* Jaadi, Zakaria. “A Step by Step Explanation of Principal Component Analysis.” Built In, 4 Sept. 2019, builtin.com/data-science/step-step-explanation-principal-component-analysis.
* Stone, James V. “Independent Component Analysis: An Introduction.” Trends in Cognitive Sciences, vol. 6, no. 2, Feb. 2002, pp. 59–64, <https://doi.org/10.1016/s1364-6613(00)01813-1>.
* Random Projections. Lecture Notes 5 2016.
* IBM. “What Is Random Forest? | IBM.” Www.ibm.com, 2023, [www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly](http://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly).
* Khalfe, Afreen. “Unsupervised Machine Learning: Clustering, Dimensionality Reduction, and Anomaly Detection Techniques.” The Talent500 Blog, 4 Aug. 2023, talent500.co/blog/unsupervised-machine-learning-clustering-dimensionality-reduction-and-anomaly-detection-techniques/. Accessed 22 Oct. 2023.