Jared Frazier | CSCI 4350 | OLA 4 | Unsupervised Learning with K-Means | Due 2021/11/30

**Introduction**

In pattern recognition, image processing, and many other problems, the task of finding groups of data points that have similar features is known as clustering [1], [2]. More broadly, such a task is referred to as unsupervised learning because there are no explicit labels associated with the features of a dataset [2]. To elucidate this concept, consider several features that are relevant to the identification of a particular flower: petal length, petal width, stem length, and stem diameter. Different types of flowers will naturally have different features (e.g., a rose is quite different than a sunflower). If the only information that one possesses are the aforementioned measurements, the objective of a clustering algorithm for such a dataset is to determine a point or series of points that captures features for a particular type of flower. Such a trained clustering algorithm could be used to classify different flowers based only on flower features, though the kind of flower (e.g., whether it’s a rose or sunflower) will not be known. Intuitively, this problem is unsupervised because features are given for different categories (or labels) of flowers, but the categories themselves may not be given.

**Methods**

One such method of clustering data is the (naive) k-means algorithm (aka Lloyd’s algorithm) [3], [4]. This algorithm assigns N data points into K clusters. A cluster is composed typically of two parts, but a third part is included based on the specifications of OLA 4: (1) the centroid; (2) the set of data points that are closest to the centroid; and (3) a label that describes the relationship of the centroid to some category in the dataset. The centroid is simply a vector containing a single continuous value per feature in the dataset. The centroid represents mean of the set of data points that are determined during the iterations of the k-means algorithm. The addition of the label is unique to this OLA, and it is used to measure how well the clustering algorithm can performs as a supervised learning problem. The computation of the centroids is the objective of the k-means algorithm, and such a computation proceeds as below [5]:

k\_random\_centroids = random\_sample\_without\_replacement(dataset)

cur\_centroids = k\_random\_centroids

converged = False

while not converged:

prev\_iter\_centroids = cur\_iter\_centroids

cur\_iter\_centroids = assign\_and\_update\_centroids(centroids, dataset)

if prev\_iter\_centroids == cur\_iter\_centroids:

converged = True

The assignment and update step of the k-means algorithm occurs by calculating the straight line (or Euclidean distance) for all data points to each of the K clusters. Data points are then assigned to a cluster and the centroid of a given cluster is subsequently updated by calculating the mean along the row axis of a N x M matrix that represents the dataset of N examples and M features. The algorithm converges when the centroids of each cluster no longer change as a result of subsequent iterations. Hence, the clusters for the n-1 and ­n iteration of the algorithm are tracked in order to check whether the previous iteration’s clusters and the current iteration’s clusters are the same.

To validate the performance of the k-means algorithm, leave-10 repeated random subsampling cross validation was performed. The repeated random subsampling size was 100—meaning 100 different random seeds were used—and for each such subsample, the performance of the clustering algorithm was computed using the iris flower and breast cancer diagnostic dataset using several cluster sizes [6]. The cluster size range for the iris dataset was 1 to 140, while the cluster size range for the cancer dataset was 1 to 95. The ranges were determined by determining the total number of examples N for each dataset and subtracting 10 because 10 such examples are used to test the performance of the algorithm.

**Results and Discussion**

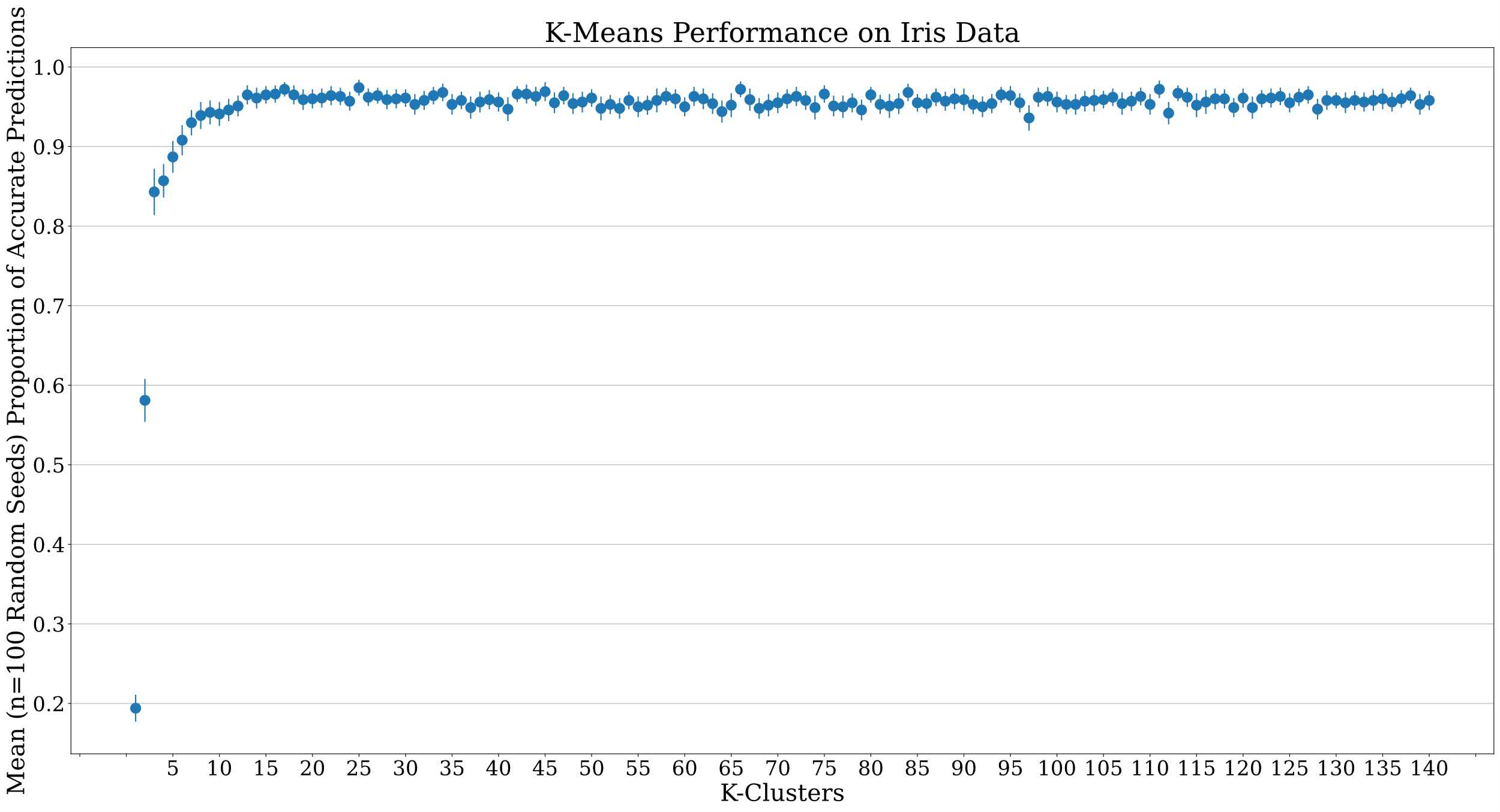
****

Figure 1. Performance on iris dataset. Error bars determined by 1.96 \* standard error (95% confidence interval).

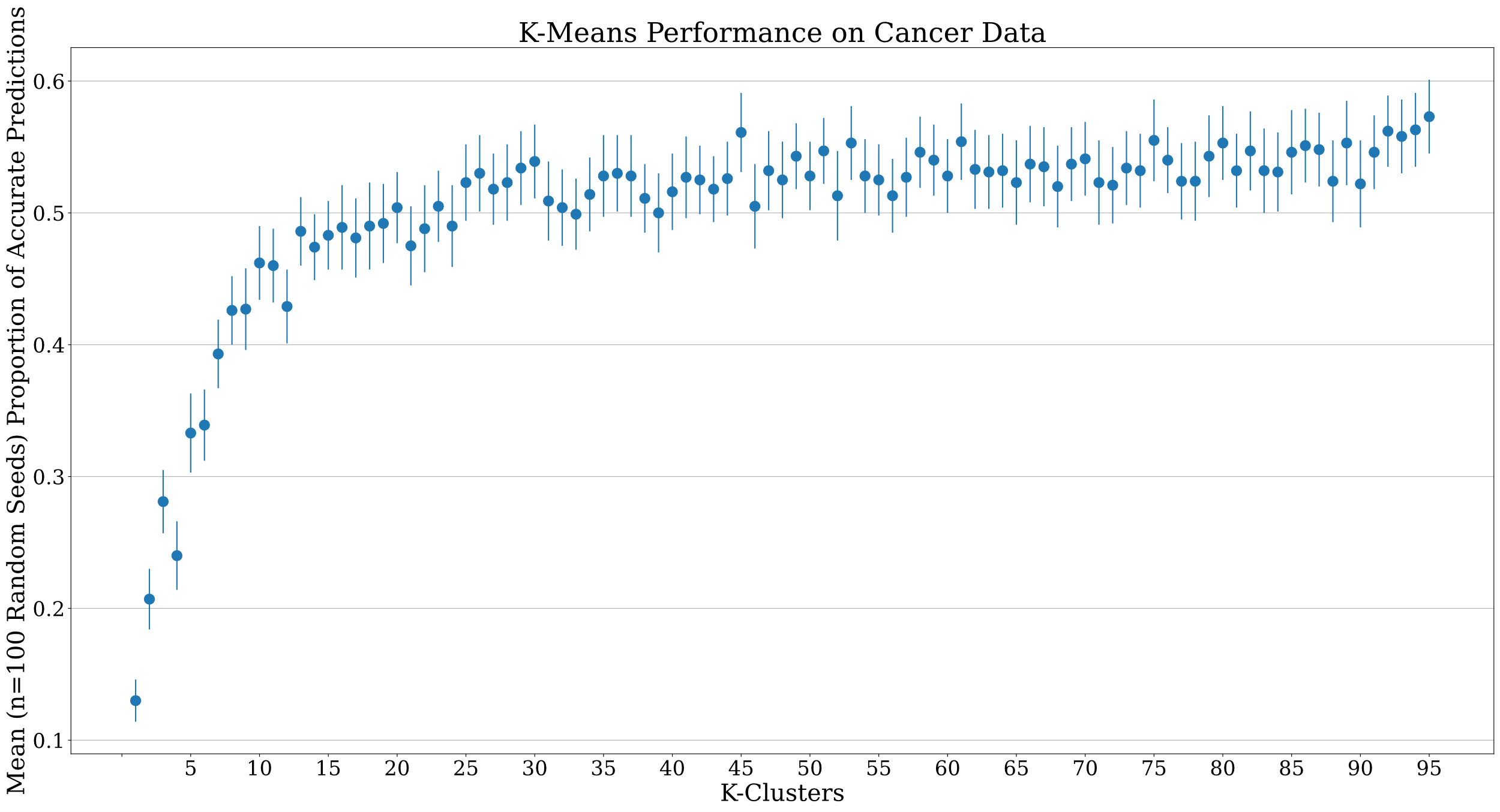
****

Figure 2. Performance on cancer dataset. Error bars determined by 1.96 \* standard error (95% confidence interval).

(Cluster, Mean Accuracy, Stderror)

Best Performer Iris**:** 25.0 0.974 0.01

Best Performer Cancer: 95.0 0.573 0.028

**References**

[1] A. Likas, N. Vlassis, and J. J. Verbeek, “The global k-means clustering algorithm,” Pattern Recognition, vol. 36, no. 2, pp. 451–461, 2003, doi: https://doi.org/10.1016/S0031-3203(02)00060-2.

[2] Wikipedia contributors, “Unsupervised learning — Wikipedia, The Free Encyclopedia.” 2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Unsupervised\_learning&oldid=1055916874

[3] D. Natingga, “Data Science Algorithms in a Week: Top 7 Algorithms for Computing, Data Analysis, and Machine Learning,” in Chapter 5: Clustering into K Clusters, Packt Publishing, 2017, pp. 102–108.

[4] Wikipedia contributors, “K-means clustering — Wikipedia, The Free Encyclopedia.” 2021. [Online]. Available: https://en.wikipedia.org/w/index.php?title=K-means\_clustering&oldid=1056937903

[5] S. Shukla and S. Naganna, “A review on K-means data clustering approach,” International Journal of Information and Computation Technology, vol. 4, no. 17, pp. 1847–1860, 2014.

[6] D. Dua and C. Graff, “UCI Machine Learning Repository.” 2017. [Online]. Available: http://archive.ics.uci.edu/ml