Python Final Project

Problem Statement

The objective of this project is to implement the k-means clustering algorithm on the Wisconsin Breast Cancer dataset and to present and interpret the data using built-in and custom made processes in Python. As breast cancer is a grave concern, Detecting it's presence and commencing treatment can, quite literally save lives. Through this analysis and reporting pipeline, cancer data can be used to classify patients' cancer as benign or malignant, using k-means clustering, to guide prognosis and treatment.

Methods

Phase 1: Data Preparation and Exploration

In the initial phase of the project, the program organizes and prepares the Wisconsin Breast Cancer dataset. The process begins by loading the dataset into Python, utilizing the pandas library. The dataset initially includes some missing values denoted by "?", which are cleaned by replacing these placeholders temporarily with NaN values. Column names are also assigned to augment clarity. The missing values of column A7 are then addressed by inputting the column mean for a complete dataset that attempts to preserve accuracy as much as possible in subsequent analysis. While this method is not perfect, it introduces as little skew as possible.

Following the cleaning, the program calculates fundamental statistical measures for attributes A2 through A10, including the mean, median, variance, and standard deviation of each attribute. From these, histograms are plotted for each of the attributes in A2 through A10 to provide visual representation of the frequency distribution of the data points. Often this makes the identification of trends and outliers within the dataset more apparent than simple tables of data.

Phase 2: K-Means Clustering

Phase 2 implements the k-means clustering algorithm to classify the breast cancer data into two distinct clusters: benign and malignant. The clustering process begins with the selection of two initial centroids, chosen randomly from the dataset using the random method from the numpy package. These centroids, representing the starting points for the clustering, are designated as $\mu 2$ and $\mu 4$.

Each data point is assigned to the nearest centroid based on Euclidean distance which effectively groups the data into two clusters. After assigning all data points, the centroids are recalculated as the mean of the data points within each cluster. This recomputation step is performed iteratively, updating the centroids until either they stabilize or the maximum number of 50 iterations is reached.

The final centroids, representing the centers of the clusters, are then output. Following this, each data point is updated with a new column, "Predicted Class," indicating its assigned cluster.

Phase 3: Error Rate Calculation and Final Report

In the final phase, the performance of the k-means clustering algorithm is assessed by calculating error rates based on the predicted cluster assignments compared to the actual class labels. The program computes the error rates for both benign and malignant clusters, as well as the total error rate for the entire dataset.

The error rates are derived from comparing the predicted cluster assignments with the true class labels. Specifically, the program calculates the number of misclassified data points for each cluster, resulting in “error B” (error rate for benign cells), “error M” (error rate for malignant cells), and “error T” (total error rate). If the total error rate exceeds 50%, indicating a potential issue with cluster assignments, the predicted clusters are adjusted by swapping labels 2 and 4. The error rates are then recalculated to reflect this adjustment.

The following includes a comprehensive overview of the clustering results, detailing the calculated error rates and any adjustments made to the predicted clusters hopefully providing insights into the accuracy of the k-means algorithm and the effectiveness of the clustering process.

Results

Phase 1 Results

As mentioned above, Phase 1 analyzed the dataset by computing various statistical measures and visualizing the data distribution through histograms. The statistics for attributes A2 through A10 were computed as follows:

Attribute A2 ----------
Mean: 4.4
Median: 4.0
Variance: 7.9
Standard Deviation: 2.8

Attribute A4 ----------
Mean: 3.2
Median: 1.0
Variance: 8.8
Standard Deviation: 3.0

Attribute A3 ----------
Mean: 3.1
Median: 1.0
Variance: 9.3
Standard Deviation: 3.0

Attribute A5 ----------
Mean: 2.8
Median: 1.0
Variance: 8.1
Standard Deviation: 2.9

Python Final Project

Attribute A6 ----------

Mean: 3.2

Median: 2.0

Variance: 4.9

Standard Deviation: 2.2

Attribute A7 ----------

Mean: 3.5

Median: 1.0

Variance: 13.0

Standard Deviation: 3.6

Attribute A8 ----------

Mean: 3.4

Median: 3.0

Variance: 5.9

Standard Deviation: 2.4

Attribute A9 ----------

Mean: 2.9

Median: 1.0

Variance: 9.3

Standard Deviation: 3.1

Attribute A10 ----------

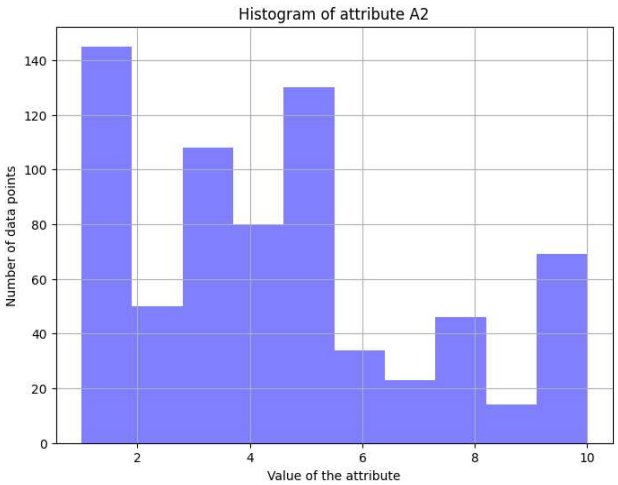
Mean: 1.6

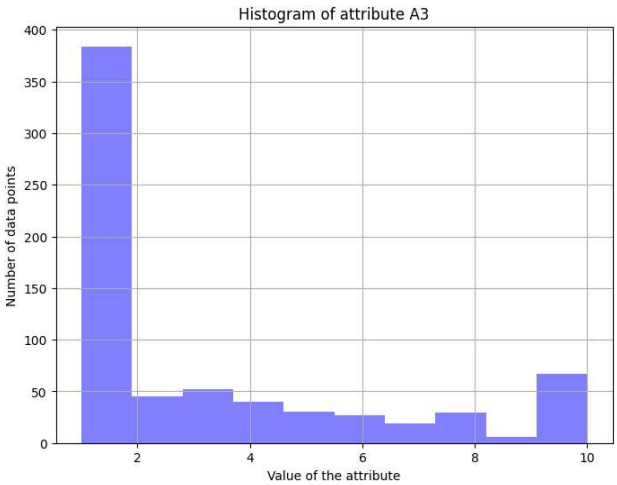
Median: 1.0

Variance: 2.9

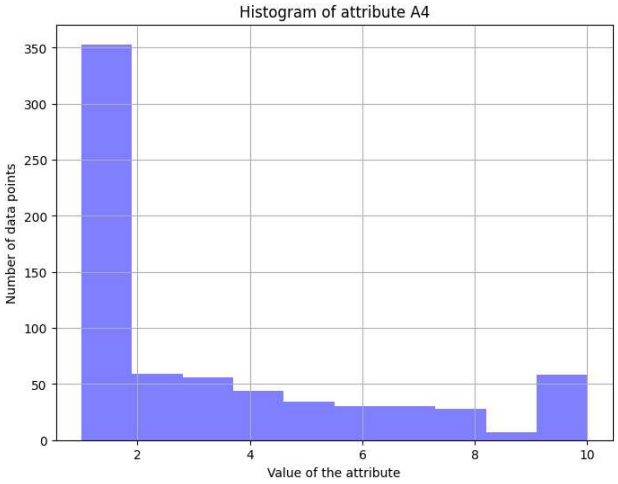
Standard Deviation: 1.7

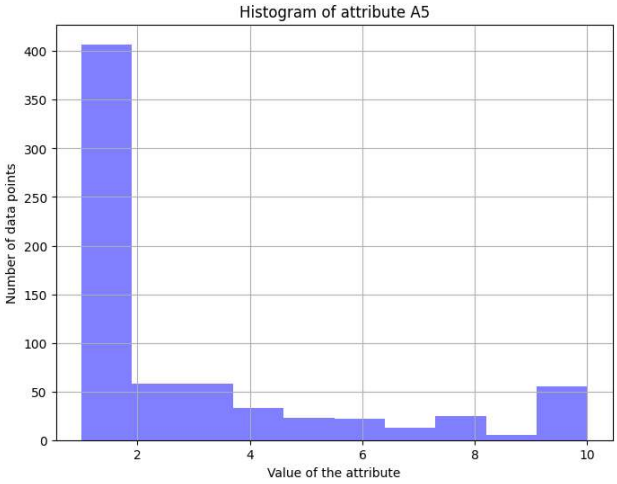
As mentioned, it is often easier to see the distribution and dispersion of the attributes at a glance via figures, thus the histograms were generated to aid in visualization.

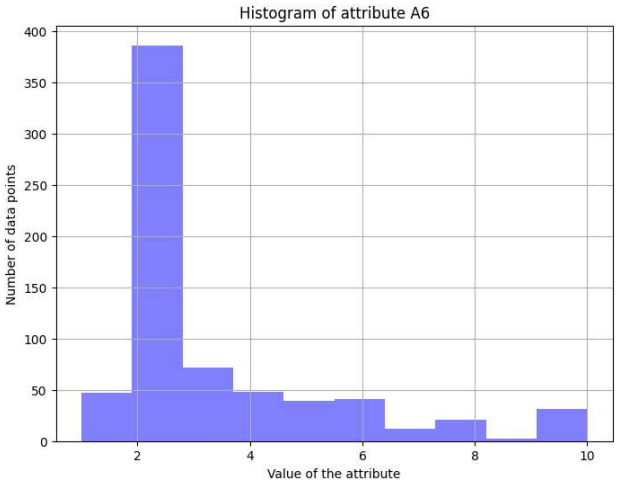


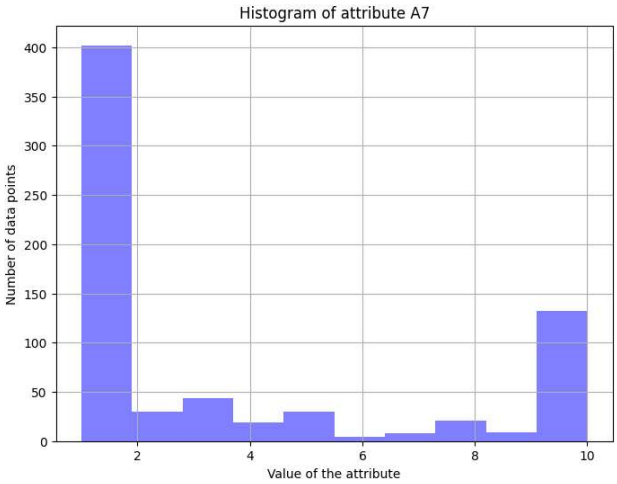


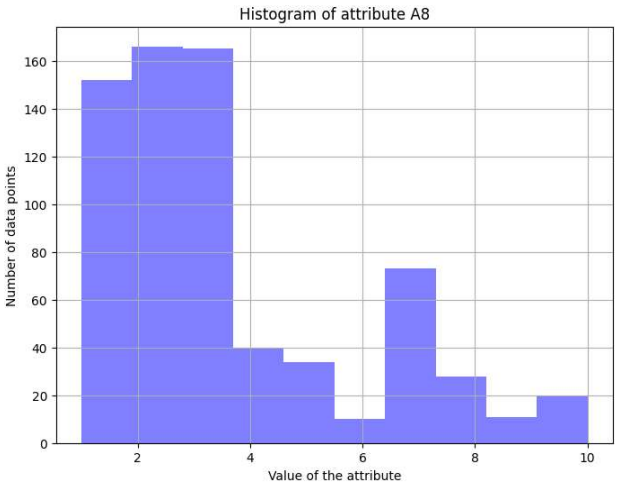
Python Final Project

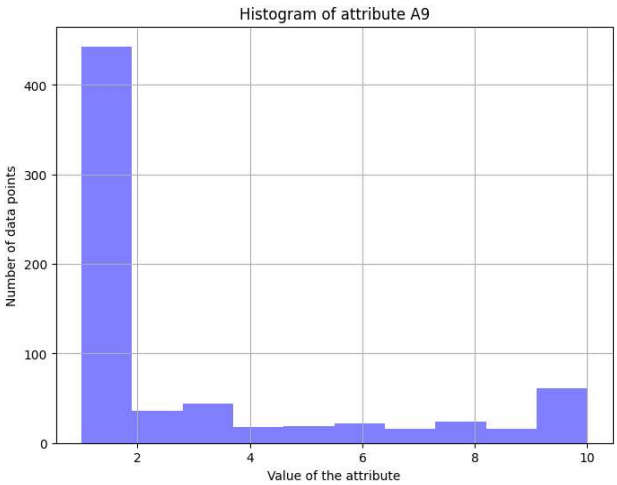


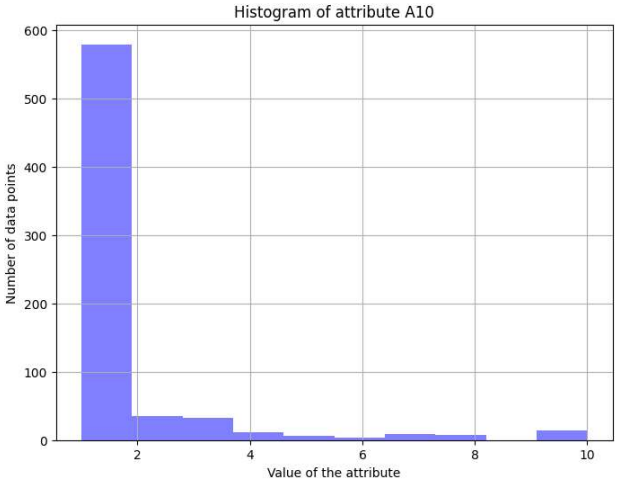












Phase 2 Results

In Phase 2, the k-means clustering algorithm was implemented to classify the data into two clusters based on randomly selected initial centroids and subsequent iterations. After 5 iterations, the final centroids were determined as:

Final centroid μ2:		Final centroid mu_4:	
A2	3.032328	A2	7.153191
A3	1.295259	A3	6.765957
A4	1.435345	A4	6.706383
A5	1.338362	A5	5.706383
A6	2.088362	A6	5.442553
A7	1.363224	A7	7.851824
A8	2.092672	A8	6.093617
A9	1.247845	A9	6.063830
A10	1.109914	A10	2.536170

The final cluster assignments were reviewed, and it was determined that the clusters were not swapped. The initial error rates for benign (class 2) and malign (class 4) cells, as well as the total error rate, were calculated as follows:

- Error rate for benign cells (class 2): 3.7%
- Error rate for malign cells (class 4): 4.7%
- Total error rate: 4.0%

The final cluster assignment for the first 20 data points was:

	Scn	Class	Predicted_Class
0	1000025	2	2
1	1002945	2	4
2	1015425	2	2
3	1016277	2	4
4	1017023	2	2
5	1017122	4	4
6	1018099	2	2
7	1018561	2	2
8	1033078	2	2
9	1033078	2	2
10	1035283	2	2
11	1036172	2	2
12	1041801	4	2
13	1043999	2	2
14	1044572	4	4
15	1047630	4	2
16	1048672	2	2
17	1049815	2	2
18	1050670	4	4
19	1050718	2	2

Phase 3 Results

In Phase 3, the error rates were computed to assess the performance of the k-means clustering. The error rates were calculated as follows:

- Error rate for benign cells (class 2): 3.7%
- Error rate for malign cells (class 4): 4.7%
- Total error rate: 4.0%

Since the total error rate was below 50%, it was confirmed that the clusters were not swapped.

The details of the error data points are summarized in the following tables:

Error data points for Predicted Class 2:

	Scn	Class	Predicted_Class
12	1041801	4	2
15	1047630	4	2
50	1108370	4	2
51	1108449	4	2
57	1113038	4	2
59	1113906	4	2
63	1116132	4	2
65	1116998	4	2
101	1167439	4	2
103	1168359	4	2
105	1169049	4	2
222	1226012	4	2
273	428903	4	2
348	832226	4	2
356	859164	4	2
455	1246562	4	2
489	1084139	4	2

Error data points for Predicted Class 4:

	Scn	Class	Predicted_Class
1	1002945	2	4
3	1016277	2	4
40	1096800	2	4
196	1213375	2	4
252	1017023	2	4
259	242970	2	4
296	616240	2	4
315	704168	2	4
319	721482	2	4
352	846832	2	4
434	1293439	2	4

Conclusion

The final project aimed to apply the k-means clustering algorithm to the Wisconsin Breast Cancer dataset to classify patients into benign and malign groups. This exercise provided practical experience in implementing clustering algorithms and evaluating their performance using real-world data.

In Phase 1, an initial statistical analysis of the dataset attributes was performed, providing a comprehensive overview of the data's central tendencies and variances. The histograms revealed the distribution patterns of attributes that are necessary for understanding the dataset's characteristics.

Phase 2 applied the k-means clustering algorithm which identified two clusters with initial and final centroids demonstrating the clustering process's convergence. The results

indicated a reasonable separation between the benign and malign groups, with initial and final centroids evolving to better represent the clusters.

Phase 3 evaluated the clustering results by calculating error rates. The initial results showed that the clusters were correctly assigned, as evidenced by error rates well below the 50% threshold, confirming the accuracy of the clustering process and ensuring that no cluster swapping occurred.