*Isabella Reeser*

*DSCI 393A*

*Exploring Seed Varieties Using Clustering Techniques*

In today’s world, a lot of industries use quality control as a critical aspect to ensure their product meets specific standards. One of the earliest cases of quality control has been seen through agriculture. When looking at seeds, quality control is extremely important in determining how we can maintain a reliable output. By analyzing different features and characteristics of seeds, we can identify valuable insight into improving the quality control process.

For my final project, I used the Seeds dataset from the UCI Machine Learning Repository. I proceeded to explore how I could use different clustering techniques we have learned in class to identify meaningful patterns based on the seed features. The dataset itself contains a variety of measurements such as area, perimeter, compactness, and kernel dimensions for three different varieties of wheat seeds specifically. The three different types are Kama, Rosa, and Canadian, however they were not properly identified within the dataset. Therefore, I won’t be able to fully specify which seed belongs to which type. I will only be able to specify the features that allow for the best quality. All measurements were made in centimeters.

My goal in applying clustering algorithms to this data is to find hidden relationships between the seed features and discover groups that could help to improve quality control. Clustering is a powerful unsupervised learning technique that is used to identify natural groups within a dataset.

First, after loading in the dataset, I noticed the data needed to be cleaned. I used colnames() to add back in the column names that were listed in the description of the dataset. After that, I used a mean imputation to replace NA values within the dataset. I did this instead of getting rid of these rows since there were a decent amount of NA values within the dataset.

Next, I needed to find the optimal number of clusters for the Seeds dataset using the elbow method, or plotting an elbow graph.

A graph of a number of clusters

Description automatically generated

Figure 1: Elbow Plot for Identifying Optimal Number of Clusters

Figure one shows the plot created to find the “elbow” point, which represents the optimal number of clusters. The elbow point is the point in the graph where adding more clusters no longer significantly reduces the within-cluster sum of squares. Through visual inspection of this plot, we can determine that the ideal number of clusters is three. In doing this, we are using a data-driven way to select the most appropriate number of groupings for the given dataset.

A graph with red green and blue dots

Description automatically generated

Figure 2: Scatter Plot of Seeds Dataset

Figure 2 shows a scatter plot of the Seeds dataset. Kernel length is the feature plotted on the x-axis and the kernel width feature is plotted on the y-axis. The clusters are relatively distinct, which means the K-means algorithm was able to identify reasonably well-separated groups of seeds based on their physical characteristics. The green cluster appears to be the largest and most compact, while the blue and red clusters are more dispersed.

A screenshot of a computer

Description automatically generated

Figure 3: Cluster 1 Characteristics

Figure 3 shows us the specific characteristics for the first cluster in Figure 2, or the cluster highlighted in red. Looking at area first, the seeds have an average of 14.03, with a minimum of 11.02 and a maximum of 16.19. This suggests the seeds in this cluster tend to have a relatively large area. The average perimeter of the seeds is 14.14, with a minimum of 12.63 and a maximum of 15.16. This indicates the seeds have a moderately high perimeter measurement. The compactness measure has a mean of 0.8813, with a minimum of 0.8392 and a maximum of 0.9183. This shows the seeds have a higher degree of compactness, so they are relatively compact in shape. The seeds in Cluster 1 have an average kernel length of 5.2454, with a minimum of 3.485 and a maximum of 5.822. This suggests the seeds tend to have relatively long kernels. The kernel width for Cluster 1 has a mean of 3.282, with a minimum of 2.879 and a maximum of 5.325. This indicates the seeds have a moderately wide kernel shape. The average asymmetry coefficient for Cluster 1 is 2.6365, with a minimum of 0.7651 and a maximum of 5.709. This relatively high asymmetry coefficient suggests the seeds in this cluster have an asymmetric or irregular shape. The mean kernel groove length for Cluster 1 is 5.036, with a minimum of 3.485 and a maximum of 6.735. This aligns with the longer kernel length observed for this cluster.

A screenshot of a computer

Description automatically generated

Figure 4: Cluster 2 Characteristics

Figure 4 shows the characteristics for Cluster 2 in the graph or the cluster in green. The seeds in Cluster 2 have a relatively large average area of 18.27, with a minimum of 15.38 and a maximum of 21.18. This suggests the seeds are generally larger in size compared to the other clusters. The average perimeter of the seeds in Cluster 2 is 16.11, with a minimum of 14.86 and a maximum of 17.25. This indicates the seeds have a moderately high perimeter measurement. The compactness measure for Cluster 2 has a mean of 0.8831, with a minimum of 0.8452 and a maximum of 0.9108. This shows the seeds in this cluster have a high degree of compactness, like Cluster 1. The seeds in Cluster 2 have an average kernel length of 6.151, with a minimum of 5.718 and a maximum of 6.675. This suggests the seeds tend to have relatively long kernels, even longer than those in Cluster 1. The kernel width for Cluster 2 has a mean of 3.67, with a minimum of 3.231 and a maximum of 4.033. This indicates the seeds have wider kernels compared to Cluster 1. The average asymmetry coefficient for Cluster 2 is 3.599, with a minimum of 1.472 and a maximum of 6.682. This relatively high asymmetry coefficient suggests the seeds in this cluster have an asymmetric or irregular shape, like Cluster 2. The mean kernel groove length for Cluster 2 is 6.011, with a minimum of 5.484 and a maximum of 6.550. This aligns with the longer kernel length observed for this cluster.

A screenshot of a computer

Description automatically generated

Figure 5: Cluster 3 Characteristics

Figure 5 shows the characteristics for Cluster 3 in the graph or the cluster in blue. The seeds in Cluster 3 have a relatively small average area of 10.80, with a minimum of 1.00 and a maximum of 14.49. This suggests these seeds are generally smaller in size compared to the other clusters. The average perimeter of the seeds in Cluster 3 is 13.12, with a minimum of 1.00 and a maximum of 14.61. This indicates the seeds have a moderately low perimeter measurement. The compactness measure for Cluster 3 has a mean of 0.8512, with a minimum of 0.8081 and a maximum of 0.8944. This shows the seeds in this cluster have a high degree of compactness, like the other two clusters. The seeds in Cluster 3 have an average kernel length of 5.301, with a minimum of 4.899 and a maximum of 5.717. This suggests the seeds tend to have relatively short kernels compared to the other clusters. The kernel width for Cluster 3 has a mean of 2.918, with a minimum of 2.630 and a maximum of 3.298. This indicates the seeds have narrower kernels compared to the other clusters. The average asymmetry coefficient for Cluster 3 is 4.731, with a minimum of 2.221 and a maximum of 8.456. This relatively high asymmetry coefficient suggests the seeds in this cluster have an asymmetric or irregular shape, like the other clusters. The mean kernel groove length for Cluster 3 is 5.176, with a minimum of 4.794 and a maximum of 5.491. This aligns with the shorter kernel length observed for this cluster.

A screenshot of a computer code

Description automatically generated

Figure 6: Comparing Clustering Results with the True Class Labels

Figure 6 shows a comparison of the clustering results with their true class label. The elements of the table suggest that the clustering results match the true class labels reasonably well. For example, most of the seeds in Cluster 1 belong to Class 1, and so on. Overall, the comparison suggests that the K-means algorithm has been reasonably effective in partitioning the seeds dataset.



Figure 7: Dendrogram for Hierarchical Clustering

Figure 7 shows a visual representation of the hierarchical clustering of the seeds data. This type of plot can offer additional insights beyond the K-means clustering analysis. The dendrogram depicts the relationships and merging of data points as the clustering process iterates. The y-axis represents the height, which corresponds to the distance of dissimilarity between clusters. The x-axis lists the individual observations. The dendrogram indicates a clear separation between the three main clusters, as evident from the relatively large distances between the final cluster merges. This aligns with the previous findings that the K-means algorithm was able to identify three distinct groups of seeds. Within each of the three main clusters, there appear to be further subclusters or smaller groupings of data points. This suggests that there may be additional substructures within the broader clusters assignments that could be worth further exploring. The height at which the clusters merge provides information about the level of similarity or dissimilarity between them. The taller the vertical lines, the more distinct the clusters are from one another. The dendrogram also highlights potential outliers or anomalous data points that do not fit neatly into the main cluster structures. These can be identified as the single branches that join the main clusters at relatively high heights.

Based on the analysis of the Seeds dataset using K-means and the dendrogram, a few conclusions can be drawn. The most important physical attributes for differentiating the seed varieties appear to be the kernel length, kernel width, and asymmetry coefficient. These features showed the clearest separation between the three main clusters, which likely correspond to the three known wheat seed types. The compactness measure was also a useful attribute, as all three clusters exhibited a high degree of compactness, indicating this is a consistent characteristic across the seed varieties. While the area and perimeter measurements helped distinguish the clusters, they were not as pivotal as the kernel-related features in terms of separating the seed types. These size-based attributes may be better suited for monitoring overall quality control rather than varietal differentiation. The dendrogram analysis suggests there may be additional subclusters within the broader groupings, which could be worth further investigation.

Moving forward, a focused investigation on the kernel-related attributes combined with a deeper dive into the hierarchical structure revealed by the dendrogram, could be significant for enhancing quality control processes and gaining a more comprehensive understanding of the wheat seed varieties. Additionally, exploring alternative clustering techniques may help refine the partitioning of the data and provide even greater insights.