

Cluster Interpretation

Cluster Characteristics

- **Cluster 1: Low-Risk Benign Tumors**

This cluster includes patients with smaller tumor sizes and lower values across most tumor characteristics. It indicates a group primarily consisting of benign tumors.

Key features of this cluster include lower mean radius, perimeter, and area, suggesting that these tumors are less aggressive.

- **Cluster 2: High-Risk Malignant Tumors**

Patients in this cluster exhibit larger tumor sizes and higher values in critical features like radius, perimeter, and area, indicating a higher likelihood of malignancy.

This cluster is characterized by significant values in compactness and concavity metrics, which are often associated with more aggressive tumor behavior.

- **Cluster 3: Moderate-Risk Tumors**

This cluster contains patients with intermediate tumor characteristics, showing a blend of benign and malignant features.

The features in this cluster reflect moderate values for radius and area, suggesting that these tumors may require closer monitoring due to their transitional risk profile.

Report Summary

Data Exploration Process

Data Loading: I started by loading the breast cancer dataset from a CSV file, which contains various tumor characteristics.

Data Cleaning: I checked for missing values and outliers. Fortunately, there were no significant outliers, and all data points were complete.

Feature Selection: I focused on relevant features for clustering, such as radius, perimeter, area, and texture metrics, which are crucial in differentiating tumor types.

Normalization: To ensure that all features contributed equally to the clustering process, I standardized the features using normalization techniques.

K-Means Clustering Results

Optimal Value of k: I used the Elbow Method to determine the optimal number of clusters. The plot indicated that $k=3$ was the best choice, where the inertia (sum of squared distances) started to plateau.

Clustering Execution: I applied K-Means clustering with $k=3$, resulting in three distinct clusters that represent different tumor risk levels.

Characteristics of Each Cluster

Cluster 1: Patients with benign tumors characterized by smaller tumor sizes and lower metrics across various features.

Cluster 2: Patients with malignant tumors characterized by larger tumor sizes and higher values in radius, perimeter, and area.

Cluster 3: Patients with moderate-risk tumors that exhibit mixed characteristics of both benign and malignant tumors.

Insights and Observations

The clustering analysis successfully highlighted the clear distinctions between benign and malignant tumors based on tumor size and shape characteristics.

Identifying moderate-risk tumors can provide valuable insights for targeted monitoring and treatment strategies, potentially improving patient outcomes.

Visualizations

Elbow Method Plot: This plot illustrates the relationship between the number of clusters (k) and the inertia. The optimal k is identified at the "elbow" point, which is $k=3$.

Silhouette Score Plot: This plot displays the silhouette scores for different values of k , indicating how well-separated the clusters are. Higher scores suggest better-defined clusters.

Cluster Scatter Plots: These visualizations show the clusters in a 2D space, using features like radius and area to illustrate the separation between different tumor types. These visualizations help in understanding the clustering process and the distinct characteristics of each cluster in the dataset.