Breast Cancer Detection

Objective:

The primary objective of this analysis is to apply unsupervised clustering techniques to the Wisconsin Breast Cancer Dataset to identify natural groupings of malignant (‘M’) and benign (‘B’) tumors without using labeled data during training.

Using algorithms such as K-Means, Gaussian Mixture Model (GMM), and Agglomerative Clustering, the study aims to separate the two diagnostic classes effectively.

This analysis benefits healthcare stakeholders by providing an automated, data-driven approach to tumor classification, potentially aiding early diagnosis and treatment planning.

By evaluating clustering performance against true labels, the study demonstrates the viability of unsupervised methods in medical diagnostics, offering cost-effective insights for clinicians and researchers.

Data Description:

The Wisconsin Breast Cancer Dataset, obtained from the UCI Machine Learning Repository through Kaggle, contains 569 samples of breast mass derived from fine needle aspirates (FNA). Each sample includes 30 real-valued features computed from digitized cell nuclei images, which capture various characteristics such as radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. These features are presented as mean, standard error, and worst (largest) values, resulting in 30 attributes.

The dataset also includes a diagnosis label (‘M’ for malignant and ‘B’ for benign), with 357 benign and 212 malignant cases provided for evaluation purposes only. Additional columns comprise a unique ‘id’ and an ‘Unnamed: 32’ column (which contains all NaN values).

Data Exploration and Preprocessing:

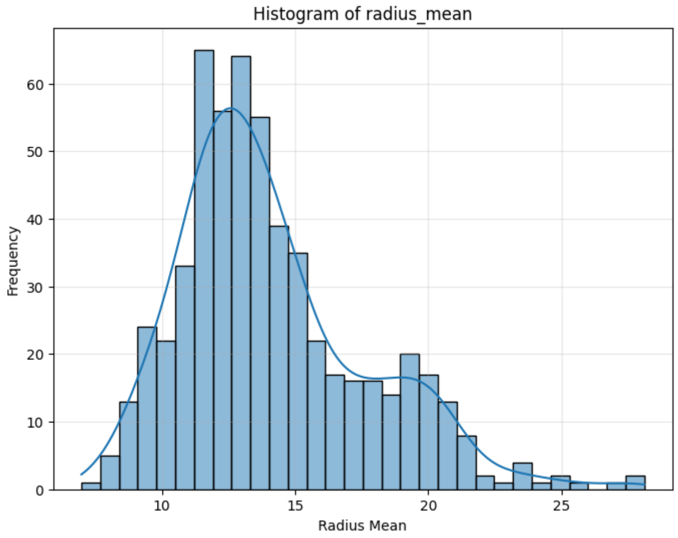
Exploration revealed a dataset with 569 samples and 33 columns: ‘id’, ‘diagnosis’, 30 features, and ‘Unnamed: 32’. The ‘id’ column had 569 unique values, and ‘Unnamed: 32’ was entirely NaN, leading to their removal. No other missing values were present, indicating a clean dataset. Feature scales varied significantly (e.g., ‘radius_mean’ ranged from 6.98 to 28.11, ‘area_mean’ from 143.5 to 2501)

Preprocessing steps included:

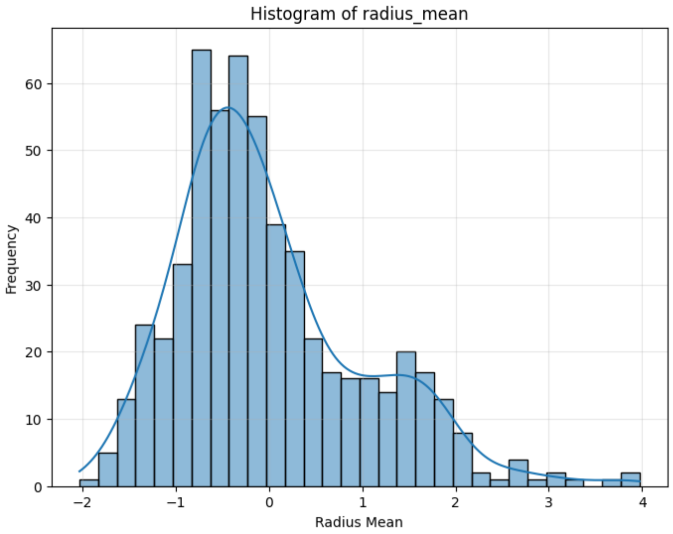
- **Column Removal**: Dropped ‘id’ and ‘Unnamed: 32’.
- **Label Encoding**: Converted ‘diagnosis’ (‘M’ → 0, ‘B’ → 1) for evaluation.
- **Standardization**: Applied ‘StandardScaler’ to normalize features.
- **PCA**: Retained 95% variance (10 components) for clustering and 2 components (63% variance) for visualization.

Standardization was essential due to feature scale disparities, enabling algorithms to treat features equally and improving clustering accuracy.

Figure below shows the distribution of ‘radius_mean’ before and after standardization



Before Standardization



After Standardization

Model fitting and results:

Three clustering algorithms—K-Means, GMM, and Agglomerative Clustering—were trained with hyperparameter variations to optimize performance.

Model 1: K-Means:

K-Means was configured with `n_clusters = 2`. Parameters tested included:

- `init`: 'k-means++', 'random'
- `n_init`: 10, 20, 50

The elbow method confirmed `n_clusters = 2`, and silhouette scores guided parameter selection, favouring `k-means++` and `n_init = 10`.

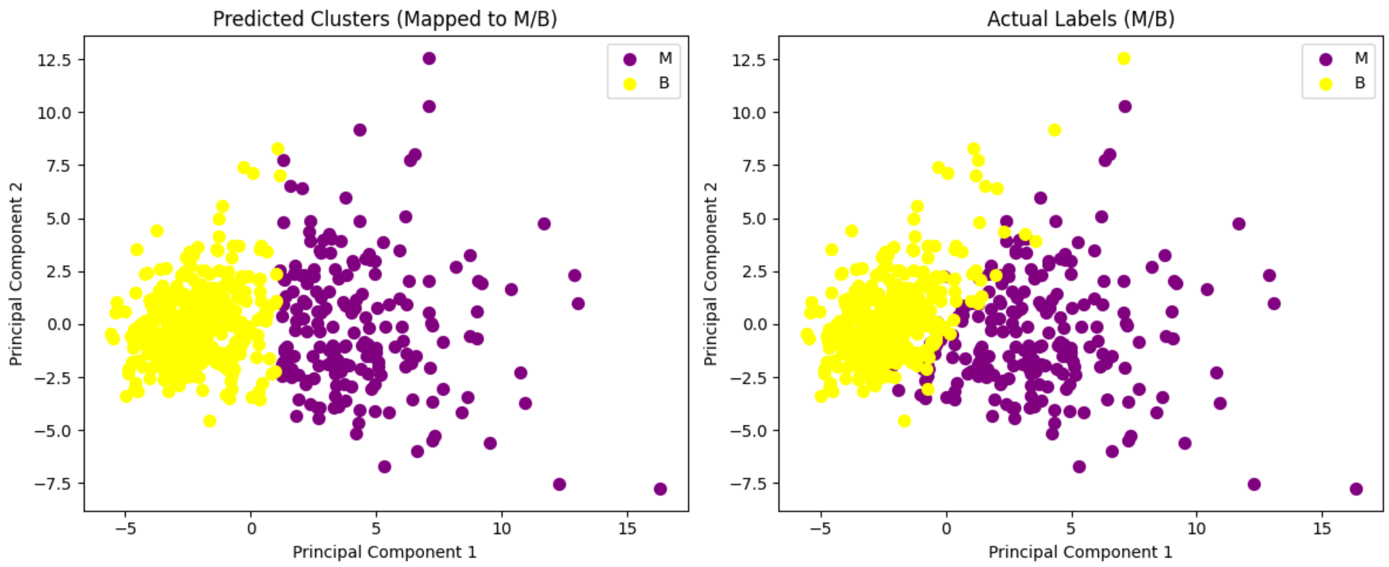
Results:

The final K-Means model achieved:

- Silhouette Score: 0.358
- Davies-Bouldin Score: 1.260
- ARI: 0.67
- Accuracy: 0.91

Final Plots:

Figure below compares predicted clusters (mapped to 'M'/'B') with actual labels in 2D PCA space, showing clear separation.



Predicted Clusters (Mapped to M/B)

Actual Labels (M/B)

Model 2: Gaussian Mixture Model (GMM)

Hyperparameter Tuning:

GMM used `n_components = 2`. Tested parameters were:

- `covariance_type`: `full`, `diag`, `spherical`
- `n_init`: 3, 5, 10

BIC plots and silhouette scores selected `covariance_type = spherical` and `n_init = 3`.

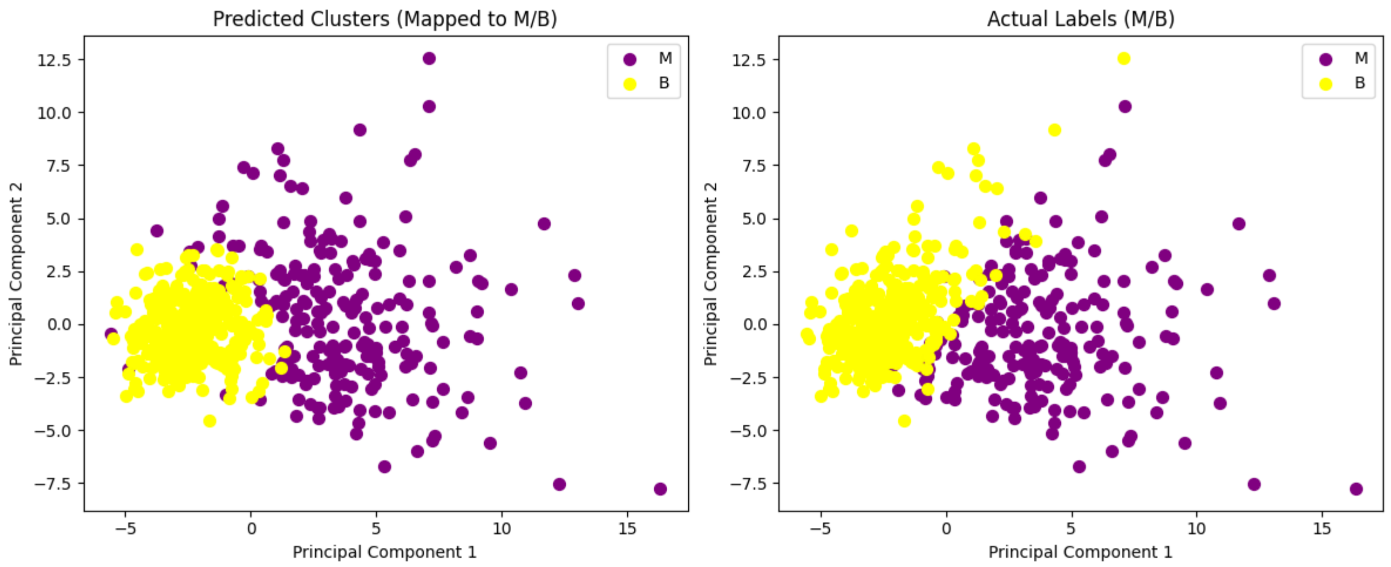
Results:

The final GMM model achieved:

- Silhouette Score: 0.326
- Davies-Bouldin Score: 1.403
- ARI: 0.551
- Accuracy: 0.872

Final Plots:

Figure below shows predicted vs. actual labels in 2D PCA space, indicating moderate separation.



Model 3 - : Agglomerative Clustering

Hyperparameter Tuning:

Agglomerative Clustering was applied with `n_clusters = 2`. Parameters included:

- `linkage`: `ward`, `complete`, `average`.

A dendrogram confirmed `n_clusters = 2` using `linkage = ward` based on silhouette scores.

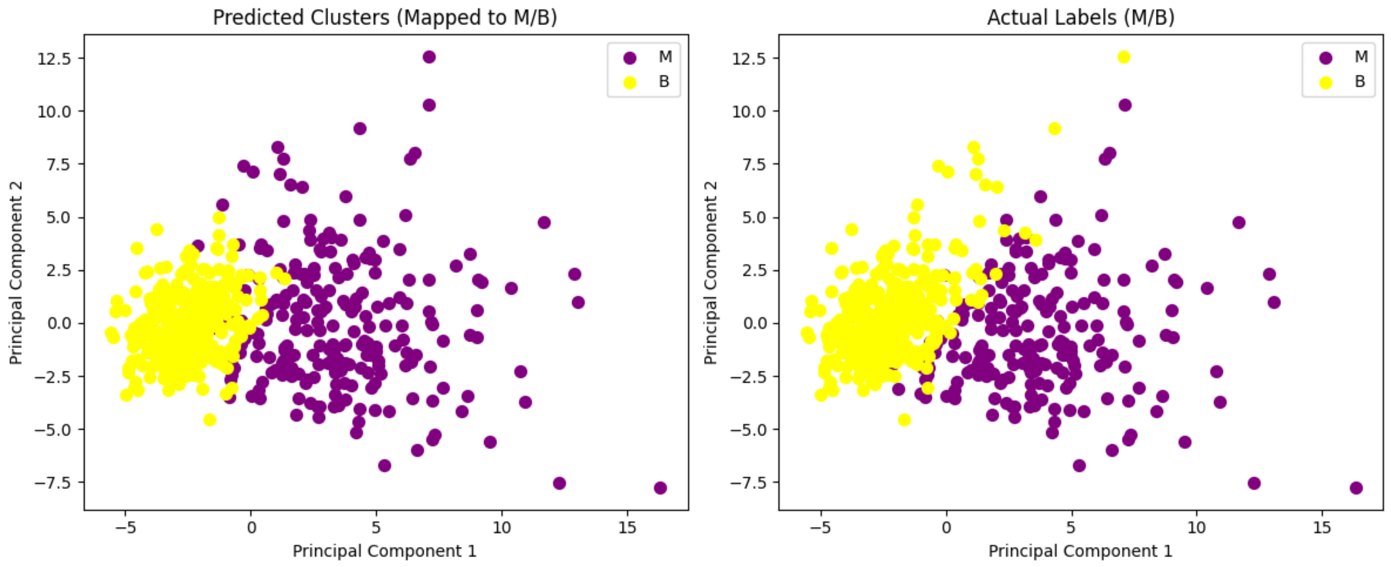
Results:

The final model achieved:

- Silhouette Score: 0.29
- Davies-Bouldin Score: 1.38
- ARI: 0.70
- Accuracy: 0.91

Final Plots:

Figure below shows a comparison of predicted vs. actual labels, demonstrating good alignment.



Model Recommendation:

K-Means is selected as the final model due to its performance metrics: Silhouette Score (0.358), ARI (0.67), and Accuracy (0.91), which surpass those of GMM and Agglomerative Clustering.

K-Means also demonstrates robustness with the dataset's well-separated structure, validated by the elbow method and scatter plots, which meets clinical requirements for clear tumor classification.

GMM's sensitivity to initialization and Agglomerative Clustering's computational complexity make K-Means the most practical choice.

Key Findings and Insights:

The analysis showed effective unsupervised clustering for the Wisconsin Breast Cancer Dataset:

- **K-Means**: Highest ARI (0.67), clear cluster separation.
- **GMM**: Moderate ARI (0.551), limited by covariance assumptions.
- **Agglomerative Clustering**: Strong ARI (0.70), less efficient than K-Means because of its computational complexity on larger datasets.
- **PCA**: 10 components retained 95% variance; 2 components (63%) aided visualization.
- **Standardization**: Improved algorithm performance with normalized feature scales.
- **Clinical Potential**: High ARI values indicate clustering can aid diagnostic tools.

DBSCAN was unsuitable, forming one cluster and labeling many points as outliers due to uniform density post-PCA, highlighting its sensitivity to density variations.

Next Steps:

- **Explore Algorithms:** Evaluate the effectiveness of Spectral Clustering or OPTICS for identifying nonlinear structures.
- **Feature Selection:** Utilize correlation analysis to prioritize key features such as 'radius_mean'.
- **Semi-Supervised Learning:** Integrate partial labels to enhance model performance.
- **External Validation:** Test the generalizability of models by applying them to new datasets.
- **Clinical Collaboration:** Work with medical experts to validate clusters for their diagnostic value.