YOUR FREEDOM IN LEARNING

Activity 1

COMP 466 Business Intelligence

Tarık Buğra Ay

28.03.2025

1. Introduction

This project was conducted to apply both unsupervised and supervised learning techniques to a medical dataset, as outlined in the course assignment (Activity 1). The dataset used is breast_cancer.arff, which contains clinical data related to breast cancer recurrence events.

The assignment objectives included:

1. Applying clustering algorithms to discover structure within the data and generate an output column (cluster label).

2. Using classification algorithms to build a supervised model trained on this cluster-based output.

3. Using the model to forecast outcomes by marking unknown targets in the dataset with '?'.

4. Evaluating and comparing classification model performance using accuracy, ROC AUC, and RMSE metrics.

5. Preparing and submitting both a project report (document) and a PowerPoint presentation.

In this project, we selected KMeans for clustering, and Decision Tree models for classification. We conducted a thorough feature engineering process to prepare the data and used several clustering evaluation metrics to determine the optimal number of clusters. The final report includes a comparison between model performance when predicting cluster labels versus predicting the actual recurrence outcomes.

2. Data Preparation

The dataset used in this project, breast_cancer.arff, contains 286 instances and 10 attributes describing patient data and treatment outcomes related to breast cancer recurrence. The data preparation phase consisted of two major parts: cleaning and feature engineering.

2.1 Data Cleaning

Since the original dataset was in ARFF format, it was manually parsed to extract attributes and values. Missing values were found in the following columns:

- node-caps: 8 missing values
- breast-quad: 1 missing value

These missing values were imputed using the mode (most frequent category) of their respective columns. Additionally, all string values were stripped of extra quotation marks for consistency. Column names were also cleaned and standardized (e.g., renaming 'Class' to Class, and 'irradiat' to irradiat).

After cleaning, the dataset had no missing values and was saved as a separate file named breast_cancer_cleaned.csv to preserve the original format before further transformation.

2.2 Exploratory Data Analysis

To better understand the distribution of each categorical feature, count plots were generated for attributes such as age, menopause, tumor-size, inv-nodes, node-caps, breast, breast-quad, and irradiat, grouped by the target class (recurrence-events vs no-recurrence-events). This visual analysis highlighted potential patterns and relationships in the data, especially around recurrence frequency in different subgroups.

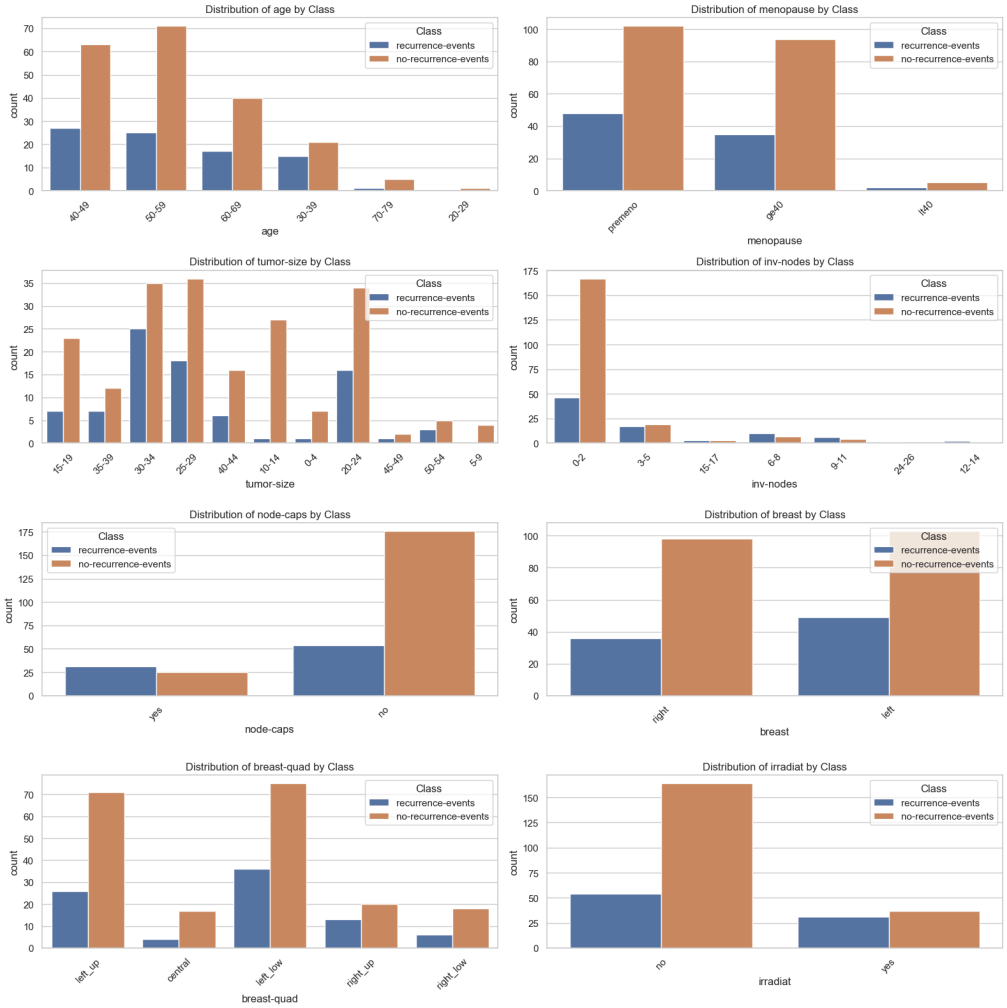

Figure 1. Shows the distribution of key categorical features in the dataset, comparing their frequency across different classes to highlight potential patterns or class-based differences.

2.3 Feature Engineering

To prepare the data for clustering and classification algorithms, the following transformations were applied:

- Range to numeric midpoints:
    - age, tumor-size, and inv-nodes were originally expressed as ranges (e.g., "40-49"). These were converted to numeric midpoints to provide ordinal information (e.g., "40-49" → 44.5).
- One-hot encoding:
    - Categorical variables were transformed using one-hot encoding to convert them into numerical binary features. This was done for the following columns:
        - menopause → menopause_ge40, menopause_lt40, menopause_premeno
        - node-caps → node_caps_yes, node_caps_no
        - irradiat → irradiat_yes, irradiat_no
        - breast → breast_left, breast_right
        - breast-quad → quad_central, quad_left_up, quad_left_low, quad_right_up, quad_right_low
- Preserved numeric features:
    - deg-malig was already a numeric feature and was kept as-is.
- Target column:
    - The Class column (recurrence-events or no-recurrence-events) was preserved for later analysis and comparison.

After feature engineering, the dataset contained 27 columns in total: 18 new numerical/binary features, the original class label, and the original columns for traceability. This transformed dataset was saved as breast_cancer_feature_engineered.csv.

3. Clustering

To explore the internal structure of the data without supervision, we applied the KMeans clustering algorithm to the feature-engineered dataset.

3.1 Feature Preparation

Before clustering, the following columns were removed from the dataset as they were either original categorical features or target labels:
age, menopause, tumor-size, inv-nodes, node-caps, breast, breast-quad, irradiat, and Class.

The remaining engineered numerical and one-hot encoded features were standardized using StandardScaler to ensure all variables contributed equally to distance-based clustering.

3.2 Selecting the Optimal Number of Clusters

To determine the ideal number of clusters k, we evaluated several values ranging from 2 to 10 using the following metrics:

- Elbow Method (Inertia): Assesses how compact the clusters are.
- Silhouette Score: Measures how well each sample fits within its cluster.
- Calinski-Harabasz Index: Evaluates between-cluster dispersion.
- Davies-Bouldin Index: Lower values indicate better clustering structure.

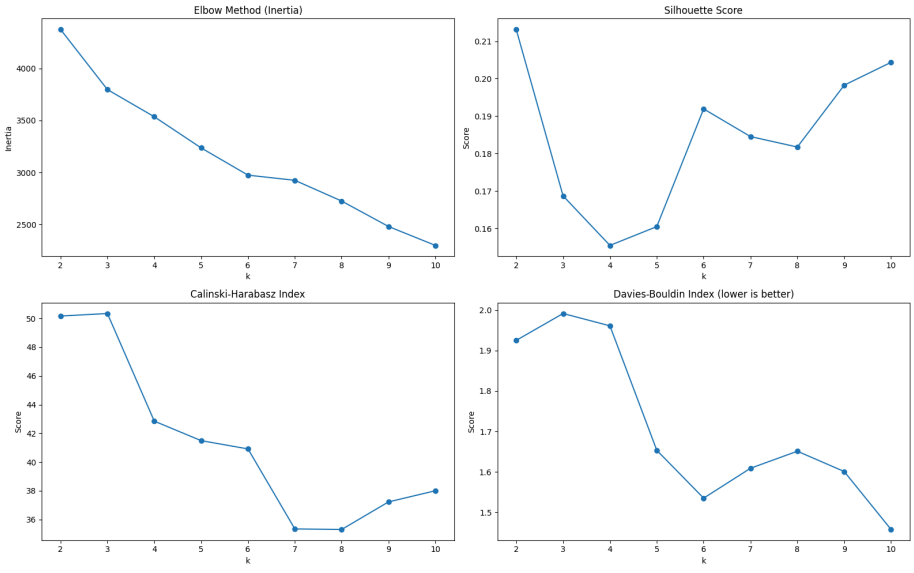

Figure 2. Results of metrics to find the number of clusters.

These metrics suggested multiple contenders, but k = 2 stood out with:

- Highest Silhouette Score
- Strong Calinski-Harabasz Index
- Reasonable inertia drop

Additionally, selecting 2 clusters aligns naturally with the binary structure of the original medical outcome (recurrence-events vs no-recurrence-events). Thus, k = 2 was chosen for interpretability and consistency with the domain context.

3.3 Applying KMeans Clustering

KMeans clustering was applied with k = 2 using the standardized feature set. The predicted cluster labels were saved as a new column: Cluster_Label.

The distribution of clusters was:

- Cluster 0: 60 instances
- Cluster 1: 226 instances

These clusters were then compared to the original class labels to evaluate how closely the unsupervised clustering matched the real-world labels.

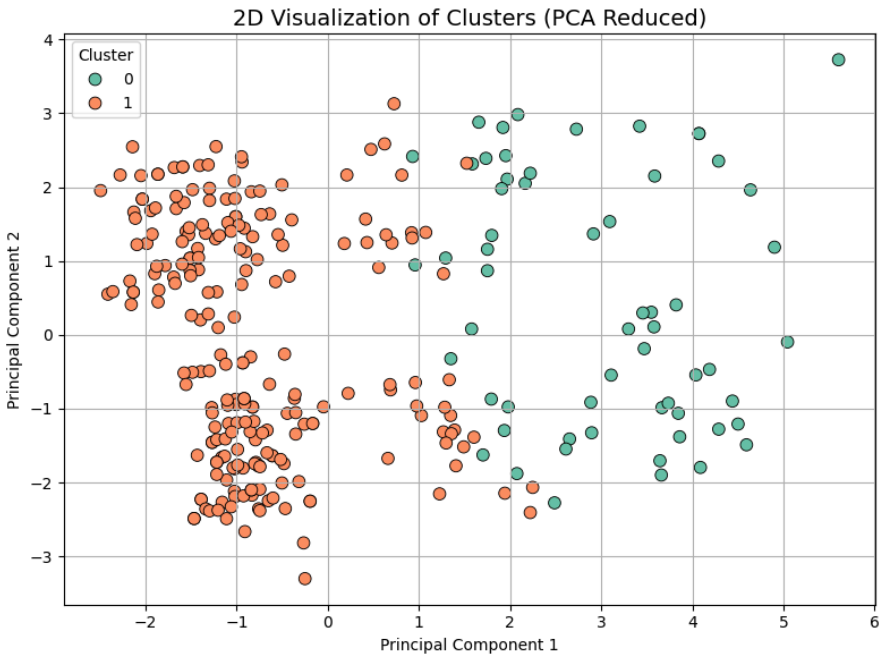

Figure 3. 2D Cluster Visualization of Clusters

3.4 Evaluating Clustering Against Actual Labels

To assess the alignment between clusters and the true medical outcomes, we:

- Encoded the Class column as a binary variable:
    - 0 = no-recurrence-events
    - 1 = recurrence-events
- Compared it to the Cluster_Label, accounting for possible label reversal (since KMeans assigns cluster IDs arbitrarily)
- Selected the label orientation that gave the highest accuracy

Evaluation Results:

- Adjusted Accuracy: 0.71
- ROC AUC Score: 0.69

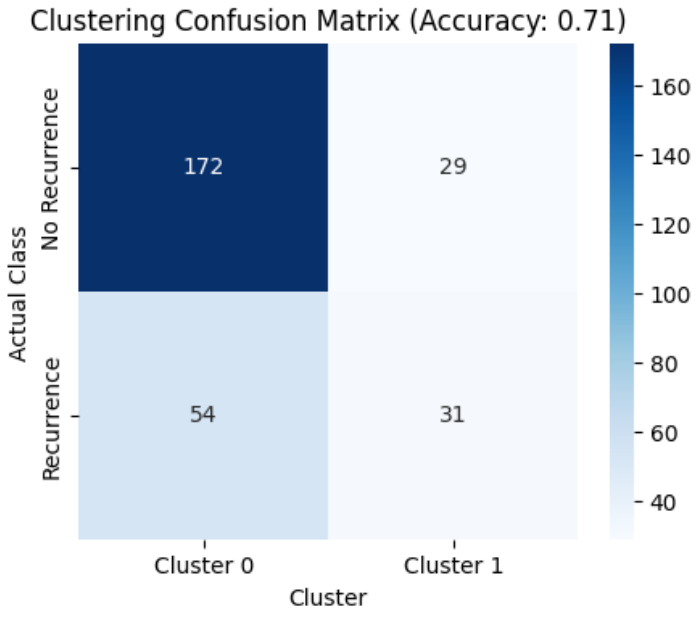

Figure 4. Cluster labels comparison to original labels.

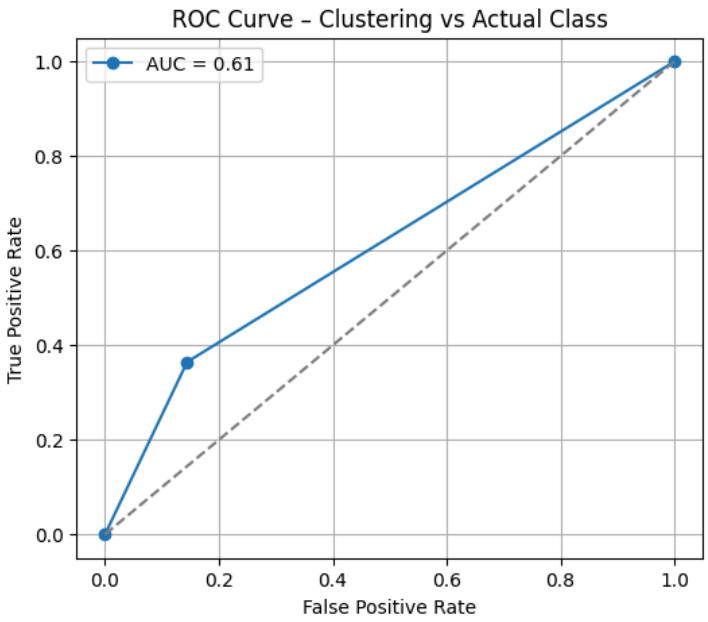

Figure 5. ROC Curve

A confusion matrix (fig 4.) and ROC curve (fig 5.) were also generated to visualize the performance of clustering relative to true class labels. The results suggest that the clusters partially capture the true class structure but do not perfectly replicate it as expected in an unsupervised setting.

4. Classification (Predicting Cluster Labels)

After clustering, the next step was to evaluate how well a supervised learning model could replicate the structure discovered by KMeans. To do this, we used the cluster assignments (Cluster_Label) as the target variable for training a Decision Tree classifier.

4.1 Data Preparation

To prepare the dataset for classification:

- We removed non-numeric columns and original class-related columns, including: Class, Class_Encoded, Cluster_Label, age, menopause, tumor-size, inv-nodes, node-caps, breast, breast-quad, and irradiat.

- The remaining features included engineered numeric midpoints and one-hot encoded binary variables.

The dataset was then split into training and test sets using an 80/20 split with stratification to maintain cluster distribution.

4.2 Decision Tree Classification

A Decision Tree classifier was trained to predict Cluster_Label from the remaining features.

Evaluation Metrics:

- Accuracy: 1.00
- ROC AUC: 1.00
- RMSE: 0.00

The classifier achieved perfect performance on the test set. This result can be explained by:

- The same features used to generate clusters were also used in training.
- Decision Trees are highly flexible and can exactly learn the logic used by KMeans to define clusters.

A confusion matrix (fig 7.) and ROC curve (fig 6.) further confirmed that the model could perfectly distinguish between the two clusters.

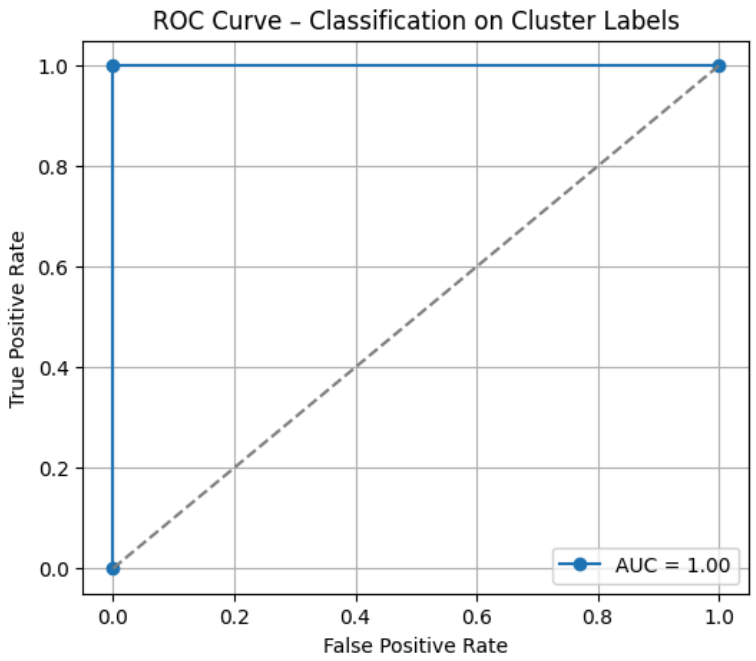

Figure 6. ROC Curve

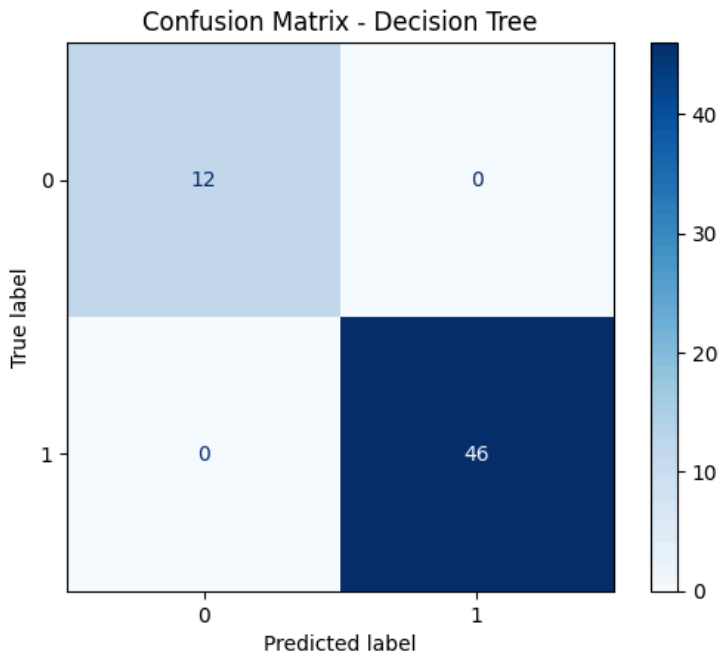

Figure 7. Confusion Matrix.

5. Classification (Predicting Real Class Labels)

To evaluate how well our features could predict actual medical outcomes, a second classification task was performed. This time, the target variable was the original class label, which indicates whether a patient experienced recurrence-events or no-recurrence-events.

5.1 Data Preparation

As in the previous step, non-feature columns were removed from the dataset, including:

- Class, Cluster_Label, and Class_Encoded (which was used as the target)
- Original string-based columns such as age, menopause, tumor-size, etc.

The remaining input features consisted of the numeric midpoints and one-hot encoded variables created during feature engineering.

The dataset was again split into an 80/20 training and test set with stratification to preserve the balance of the class labels.

5.2 Decision Tree Classification

A Decision Tree classifier was trained to predict the real Class column. This provided a more realistic evaluation of how well our features could predict actual medical recurrence outcomes.

Evaluation Metrics:

- Accuracy: 0.71
- ROC AUC: 0.64
- RMSE: 0.53

The classifier showed modest performance. Unlike the cluster-label prediction, the model could not perfectly classify the true medical labels, which is expected given the more complex and potentially noisy nature of real-world outcomes.

The ROC curve (fig 8.) and confusion matrix (fig 9.) highlighted that:

- The model performed better in identifying non-recurrence cases.
- It had more difficulty correctly identifying recurrence-events, which were fewer in number.

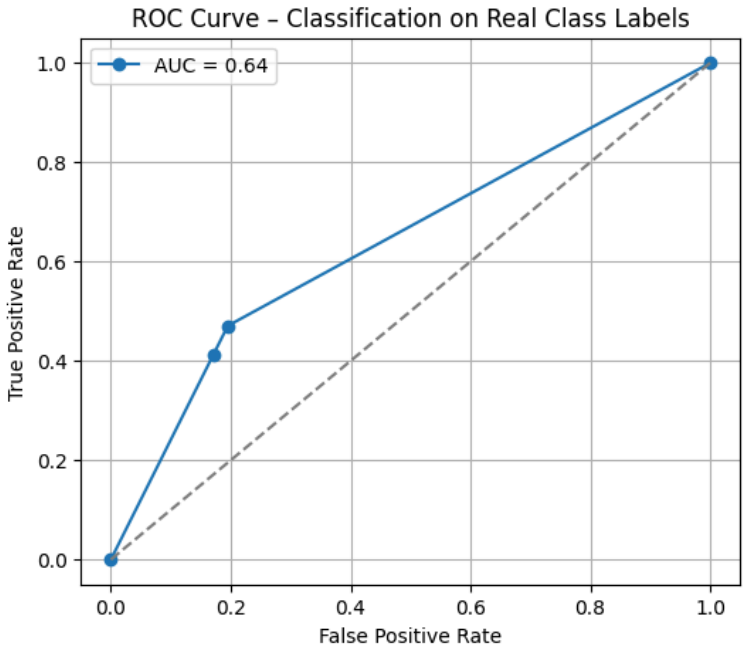

Figure 8. ROC curve.

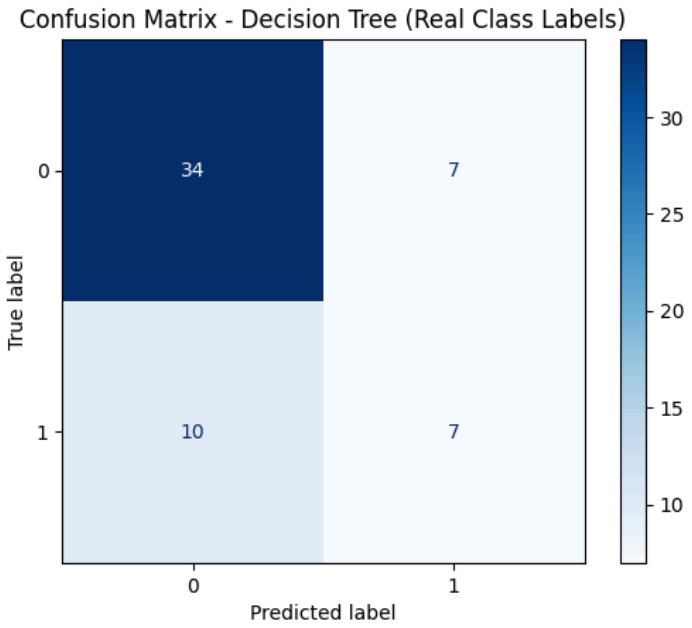

Figure 9. Confusion matrix.

6. Comparison

The two classification models showed clear differences in performance. The model trained to predict cluster labels achieved perfect scores across all evaluation metrics (accuracy, ROC AUC, RMSE), which can be explained by the fact that the model was learning to replicate a structure generated from the same features. In contrast, the model trained to predict the actual class labels (recurrence-events vs no-recurrence-events) achieved more modest results, with an accuracy of 0.71 and a ROC AUC of 0.64. This reflects the increased difficulty of modeling real-world outcomes, which are often influenced by factors not captured in the available features. The comparison highlights the difference between learning algorithm-generated patterns versus attempting to predict complex medical realities.

7. Conclusion

This project demonstrated the application of both unsupervised and supervised learning techniques on a breast cancer dataset. KMeans clustering was used to explore hidden structure in the data, and classification models were trained to predict both the generated cluster labels and the actual medical outcomes. While the cluster-based classification achieved perfect performance due to its alignment with the input features, the real class prediction yielded more modest results, reflecting the inherent complexity of real-world health data. The project provides insight into how feature engineering and learning objectives impact model performance and interpretation.

The full source code, notebooks, and data files used in this project are available in the GitHub repository below:

GitHub Repository: https://github.com/TarikBugraAy/breast-cancer-clustering-classification