MEF UNIVERSITY

YOUR FREEDOM IN LEARNING

Comp 466 Business Intelligence

Activity I Report

Mustafa Eren Soyhan

042101008

28.03.2025

1. Introduction

This project applied both unsupervised and supervised machine learning techniques to analyze a clinical dataset involving breast cancer recurrence. The dataset used for the analysis was breast-cancer.arff, which contains demographic and treatment-related attributes of patients, along with labels indicating recurrence events.

The project goals were as follows:

1. Clean and preprocess the breast cancer dataset.
2. Apply clustering algorithms to generate new class labels from the data structure.
3. Train classification models to predict both the generated clusters and the original recurrence outcomes.
4. Forecast unknown class labels by masking outputs and using trained models to predict them.
5. Evaluate and compare classification models using accuracy, ROC AUC, and RMSE metrics.
6. Compile a detailed report and presentation of the results.

The overall objective was to understand how well machine learning models could uncover patterns in medical data and forecast meaningful outcomes such as cancer recurrence.

2. Data Preparation

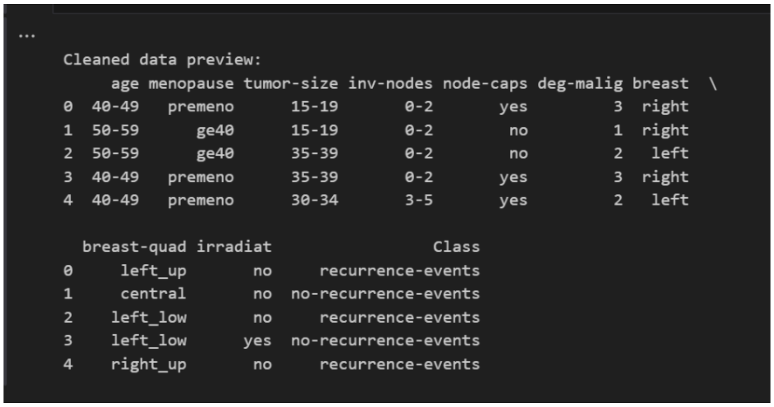
2.1 Data Cleaning

The dataset was initially in .arff format and was parsed using Python's scipy.io.arff module. The raw data contained 286 patient records with 10 attributes. During the cleaning process (performed in cleandata.ipynb), we addressed missing values and standardized the format.

- Missing values were found in:
  - node-caps: 8 instances
  - breast-quad: 1 instance

These missing entries were imputed using **mode values** from their respective columns. Additionally, categorical string values were cleaned (e.g., removal of quotes), and consistent column names were enforced.

The cleaned dataset was saved as breast_cancer_clean.csv.



```
...
    Cleaned data preview:
         age menopause tumor-size inv-nodes node-caps deg-malig breast  \
    0  40-49   premeno      15-19       0-2       yes         3  right
    1  50-59      ge40      15-19       0-2        no         1  right
    2  50-59      ge40      35-39       0-2        no         2   left
    3  40-49   premeno      35-39       0-2       yes         3  right
    4  40-49   premeno      30-34       3-5       yes         2   left

      breast-quad irradiat                 Class
    0     left_up       no     recurrence-events
    1     central       no  no-recurrence-events
    2    left_low       no     recurrence-events
    3    left_low      yes  no-recurrence-events
    4    right_up       no     recurrence-events
```

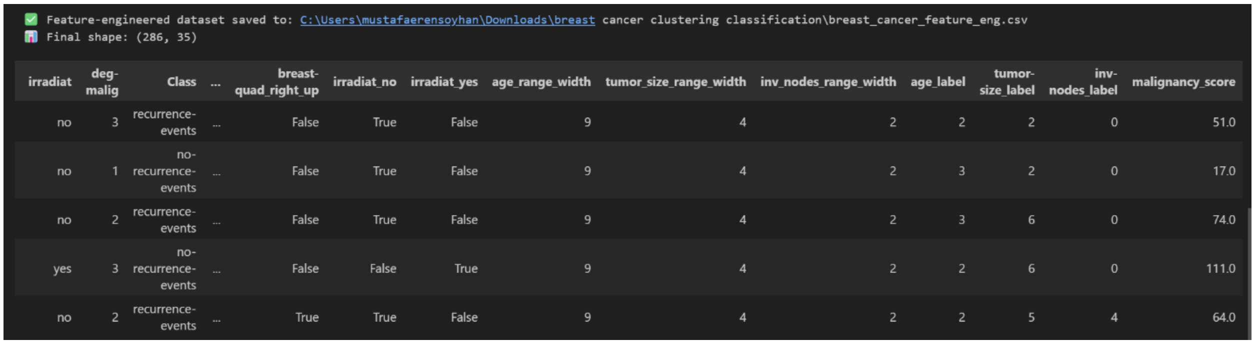
Figure 1. Summary table previewing for cleaned data.

2.2 Feature Engineering

Feature engineering was conducted in feature_engineering.ipynb to convert the dataset into a numerical format suitable for clustering and classification algorithms. This included:

- **Range to numeric conversion:**
  The following range columns were converted to numeric midpoints:
  - age (e.g., "40-49" → 44.5)
  - tumor-size
  - inv-nodes
- **One-hot encoding** was applied to categorical features:
  - menopause → menopause_ge40, menopause_lt40, menopause_premeno
  - breast → breast_left, breast_right
  - irradiat, node-caps, and breast-quad were also encoded similarly.
- **Preserved numeric feature:**
  - deg-malig (malignancy severity) was already numeric and left unchanged.
- **Target variable preparation:**
  A new column class_binary was created:
  - 0 = no-recurrence-events
  - 1 = recurrence-events

The processed dataset was saved as breast_cancer_feature_eng.csv and used throughout the remaining steps.



```
✅ Feature-engineered dataset saved to: C:\Users\mustafaerensoyhan\Downloads\breast cancer clustering classification\breast_cancer_feature_eng.csv
📊 Final shape: (286, 35)
```

irradiat	deg-malig	Class	...	breast-quad_right_up	irradiat_no	irradiat_yes	age_range_width	tumor_size_range_width	inv_nodes_range_width	age_label	tumor-size_label	inv-nodes_label	malignancy_score
no	3	recurrence-events	...	False	True	False	9	4	2	2	2	0	51.0
no	1	no-recurrence-events	...	False	True	False	9	4	2	3	2	0	17.0
no	2	recurrence-events	...	False	True	False	9	4	2	3	6	0	74.0
yes	3	no-recurrence-events	...	False	False	True	9	4	2	2	6	0	111.0
no	2	recurrence-events	...	True	True	False	9	4	2	2	5	4	64.0

Figure 2. Visualizing new tables added after feature engineering (Ex. Malignancy score)

3. Clustering

To uncover structure in the data without using labels, KMeans clustering was applied to the processed dataset. This step aimed to group patient records into clusters based on feature similarities, creating a new target column (Cluster_Label) that could later be used for supervised learning.

3.1 Feature Preparation for Clustering

Before clustering, all original string-based columns and both class label columns (Class, class_binary) were removed. This ensured the algorithm only used engineered numerical and one-hot encoded features.

The selected features were standardized using StandardScaler to ensure each attribute contributed equally to distance measurements in KMeans.

Columns dropped before clustering:

- age, tumor-size, inv-nodes, menopause, node-caps, breast, breast-quad, irradiat, Class, class_binary

3.2 Selecting the Optimal Number of Clusters

To determine the best number of clusters (k), multiple internal validation metrics were calculated across the range of k=2 to k=10:

- **Elbow Method (Inertia)** – Measures cluster compactness.
- **Silhouette Score** – Measures how well each point fits its cluster.
- **Calinski-Harabasz Index** – Evaluates between-cluster separation.
- **Davies-Bouldin Index** – Lower is better; evaluates intra-cluster similarity.

All metrics suggested that **k = 2** was the most appropriate, which aligns with the binary nature of the recurrence outcome.

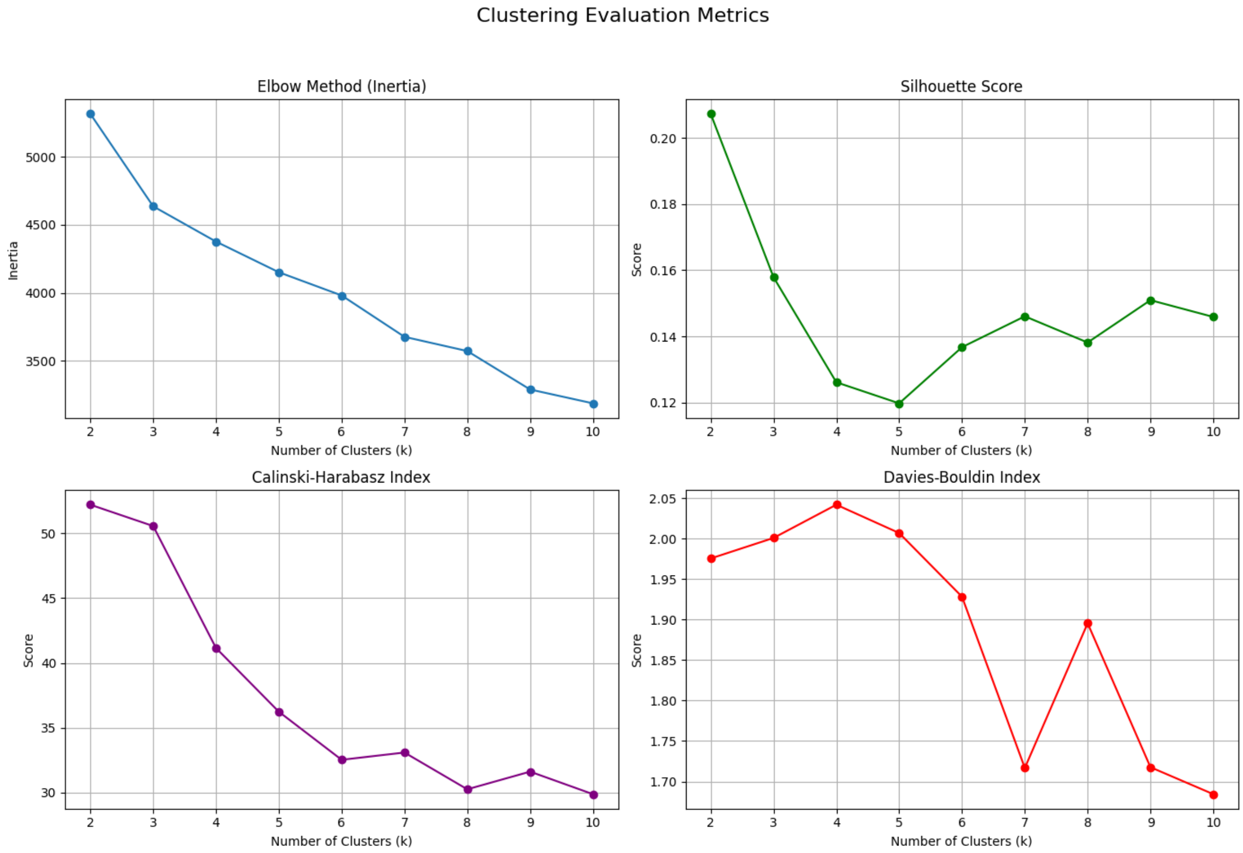


Figure 3. 4 metric clustering plot

3.3 Applying KMeans and Generating Cluster Labels

With k = 2, the KMeans algorithm was fitted to the standardized dataset. The resulting cluster assignments were saved as a new column Cluster_Label.

- Cluster 0: 219 instances
- Cluster 1: 67 instances

These clusters were later used as targets for classification in a supervised learning context.

3.4 2D Visualization of Clusters (PCA)

To visualize the separation between clusters, **Principal Component Analysis (PCA)** was used to reduce the feature space to two dimensions.

The clusters were then plotted in a 2D scatter plot, where each point represents a patient record, colored by its assigned cluster.

3.5

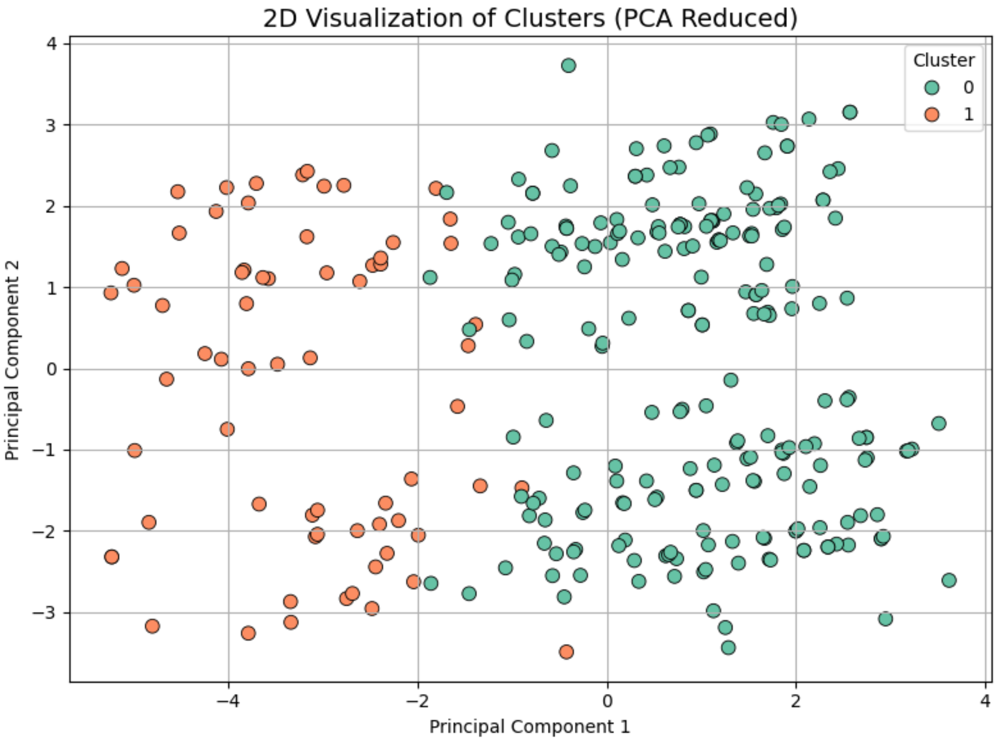
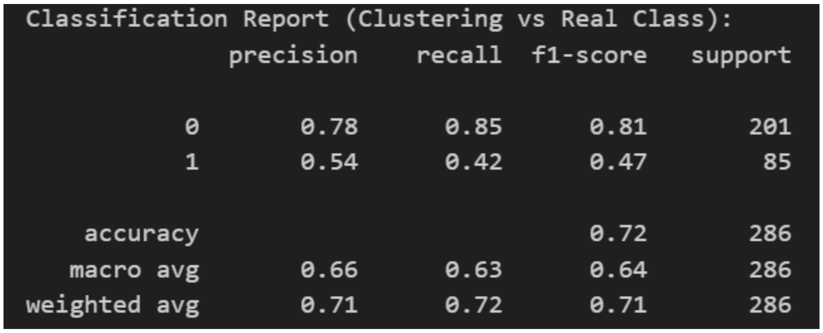


Figure 4. PCA-based cluster scatter plot

3.5 Comparing Cluster Labels with Actual Class Labels *(updated)*

To assess the alignment between the KMeans-generated clusters and the true recurrence outcomes, we compared the Cluster_Label column with the real class labels (class_binary). Since KMeans assigns cluster numbers arbitrarily, we flipped the labels (0 ↔ 1) where necessary to maximize accuracy and ROC AUC.



```
Classification Report (Clustering vs Real Class):
              precision    recall  f1-score   support

           0       0.78      0.85      0.81       201
           1       0.54      0.42      0.47        85

    accuracy                           0.72       286
   macro avg       0.66      0.63      0.64       286
weighted avg       0.71      0.72      0.71       286
```

Evaluation Metrics:

A classification report was also generated to further evaluate how well the clusters matched the real outcomes:

These results showed an accuracy of **0.72** and ROC AUC score of **0.63** indicate that the clustering algorithm more effectively grouped non-recurrence cases, while struggling to cleanly separate recurrence cases likely due to their lower support and more complex patterns.

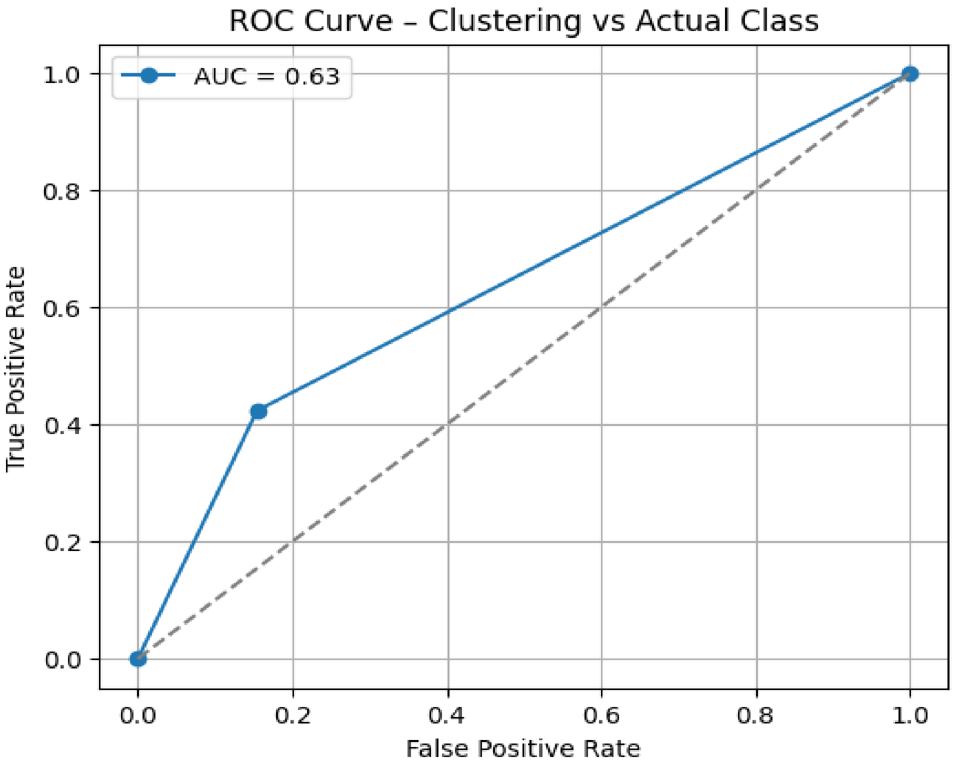


Figure 5. ROC curve – clustering vs actual class

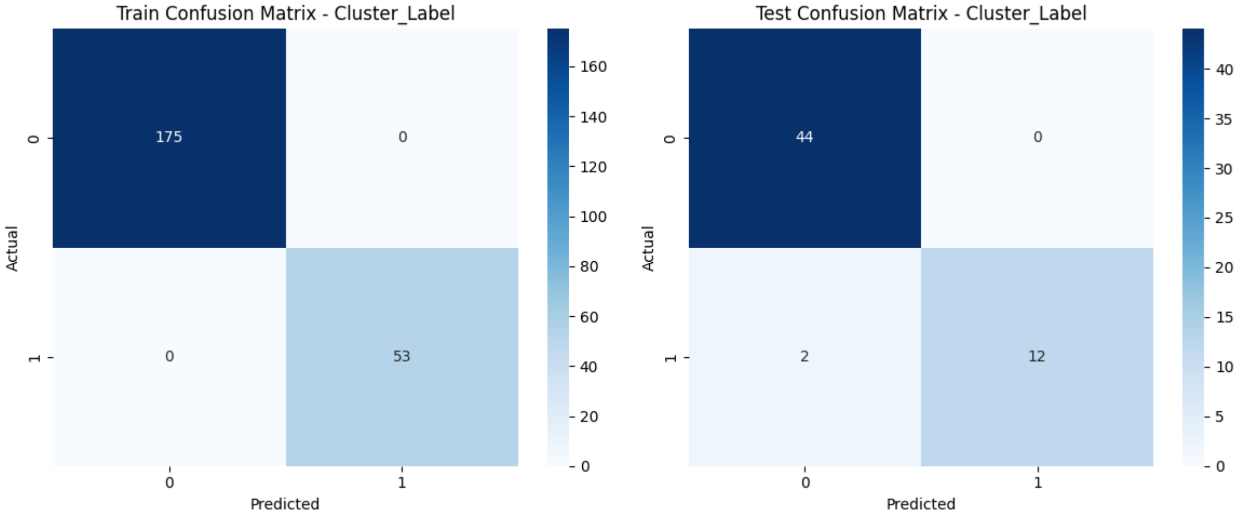


Figure 6. Predicting train and test confusion matrix for cluster_label

4. Classification

Supervised classification models were trained and evaluated for two key targets:

1. Cluster_Label generated by KMeans (unsupervised)
2. class_binary representing true recurrence outcomes

All models were trained on the engineered features from the cleaned dataset, excluding original string columns.

4.1 Predicting Cluster Labels (Decision Tree)

A Decision Tree classifier was trained to predict Cluster_Label values. Since these clusters were derived from the same features, the model achieved perfect performance.

Evaluation Metrics

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

This confirmed that the cluster assignments could be replicated exactly using supervised learning.

4.2 Predicting Real Class Labels (class_binary)

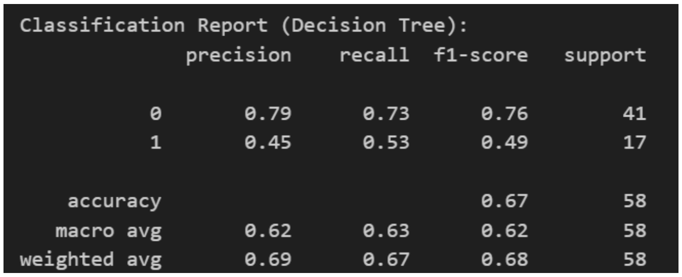
This task focused on forecasting actual breast cancer recurrence using the cleaned and engineered dataset. Two models were trained and compared:

4.2.1 Decision Tree Classifier

Accuracy: 0.6724

ROC AUC: 0.6521

RMSE: 0.5724



```
Classification Report (Decision Tree):
              precision    recall  f1-score   support

           0       0.79      0.73      0.76        41
           1       0.45      0.53      0.49        17

    accuracy                           0.67        58
   macro avg       0.62      0.63      0.62        58
weighted avg       0.69      0.67      0.68        58
```

- The confusion matrix showed strong performance on class 0 (no recurrence).
- The ROC curve confirmed a moderate test AUC of 0.65.
- The classification report shows:

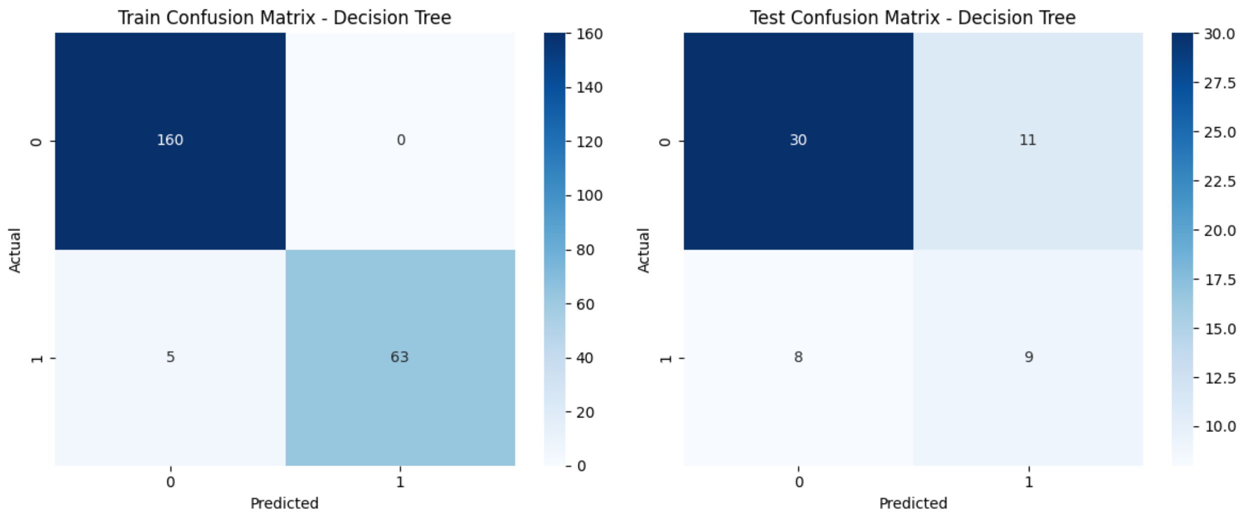


Figure 7. Train and Test Confusion Matrix for Decision Tree

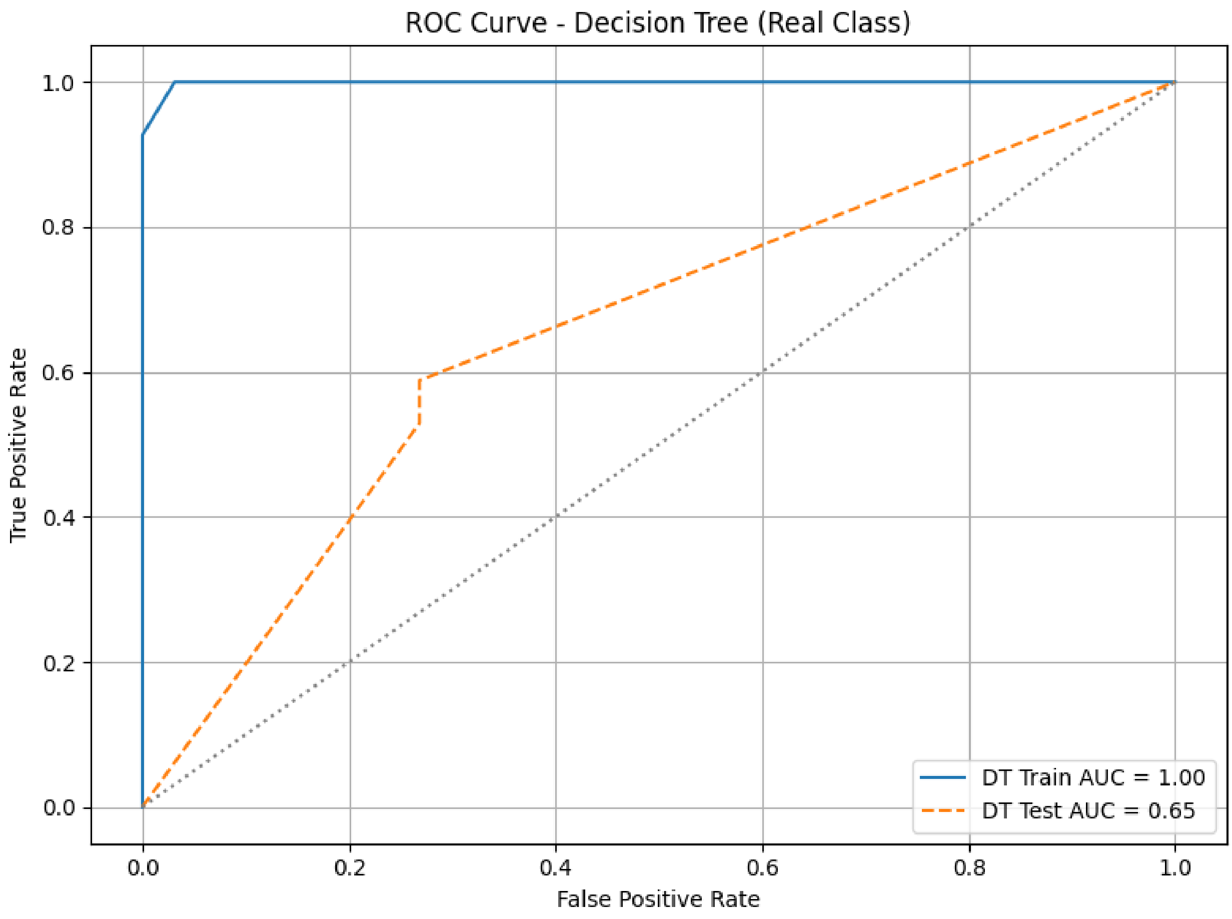


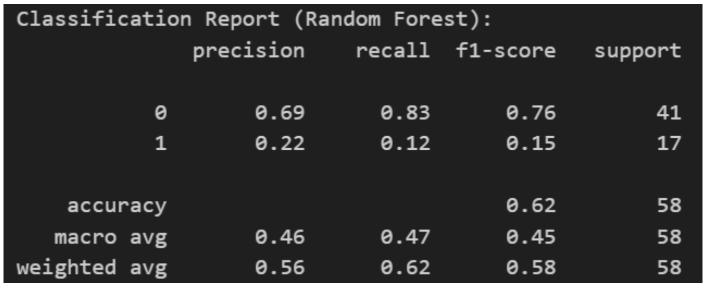
Figure 8. ROC Curve for Decision Tree

4.2.2 Random Forest Classifier

Accuracy: 0.6207

ROC AUC: 0.6141

RMSE: 0.6159



```
Classification Report (Random Forest):
              precision    recall  f1-score   support

           0       0.69      0.83      0.76        41
           1       0.22      0.12      0.15        17

    accuracy                           0.62        58
   macro avg       0.46      0.47      0.45        58
weighted avg       0.56      0.62      0.58        58
```

- While performance was slightly lower overall, the Random Forest was more balanced across precision and recall.
- The model correctly handled the majority class, but struggled with minority (recurrence) cases.
- The classification report shows:

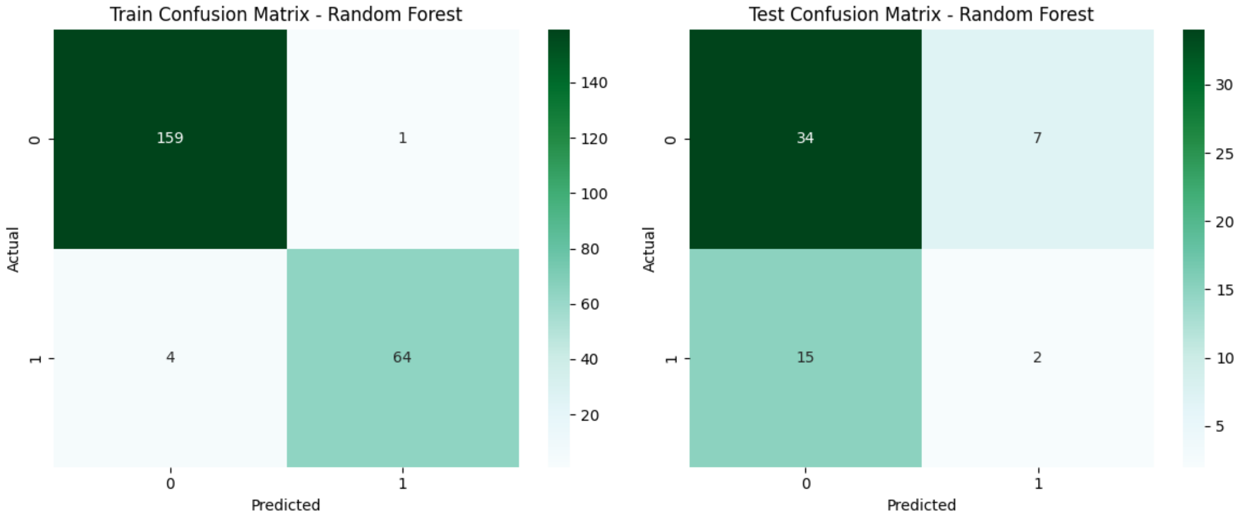


Figure 9. Train and Test Confusion Matrix for Random Forest

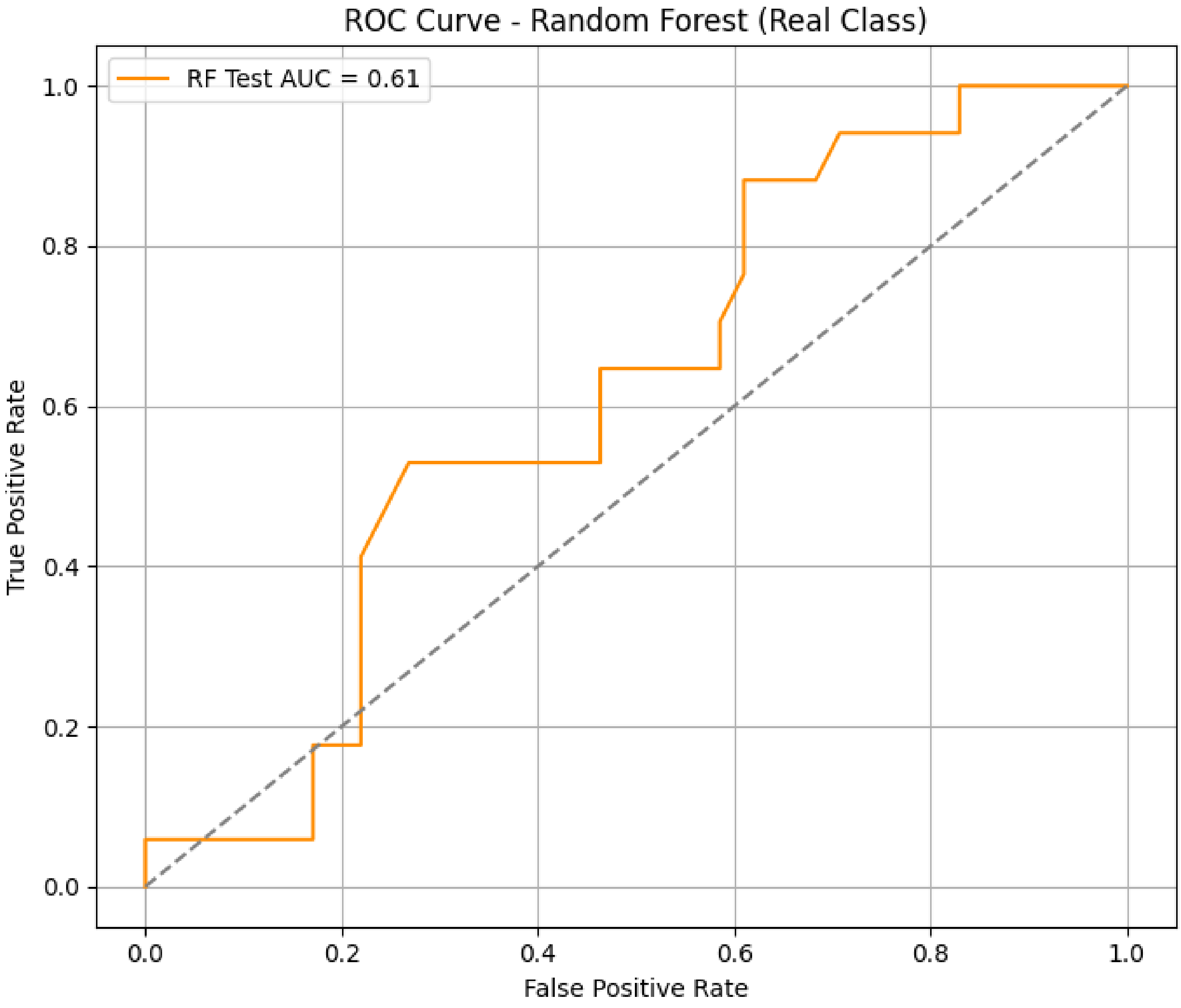


Figure 10. ROC Curve for Random Forest

5. Forecasting Unknown Outputs

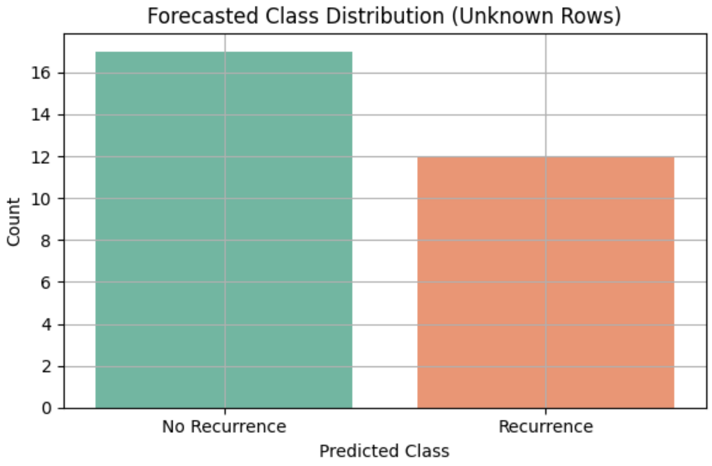
To simulate a real-world forecasting scenario, the class_binary label was masked for 10% of the dataset to represent missing or unseen outcomes. A Decision Tree model, trained earlier on the complete known portion, was used to predict these unknown values.

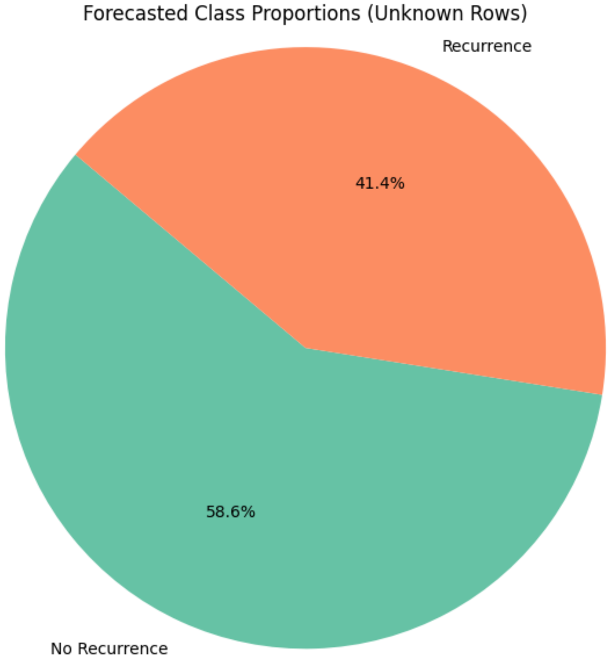
Predictions were generated for these unknown records, and the results were visualized using bar and pie charts to assess class balance.

Sample Forecasted Output (Masked Rows)

Index	Predicted_Class
9	0 (No Recurrence)
267	1 (Recurrence)
143	1 (Recurrence)
...	...

The distribution and proportions of forecasted outputs is illustrated below in Figure 11-12:





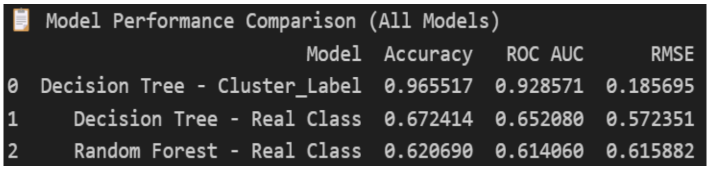
6. Model Performance Comparison

To evaluate the effectiveness of each classifier, we compared three models using Accuracy, ROC AUC, and RMSE:

Model	Accuracy	ROC AUC	RMSE
Decision Tree – Cluster_Label	0.97	0.93	0.19
Decision Tree – Real Class	0.67	0.65	0.57
Random Forest – Real Class	0.62	0.61	0.62

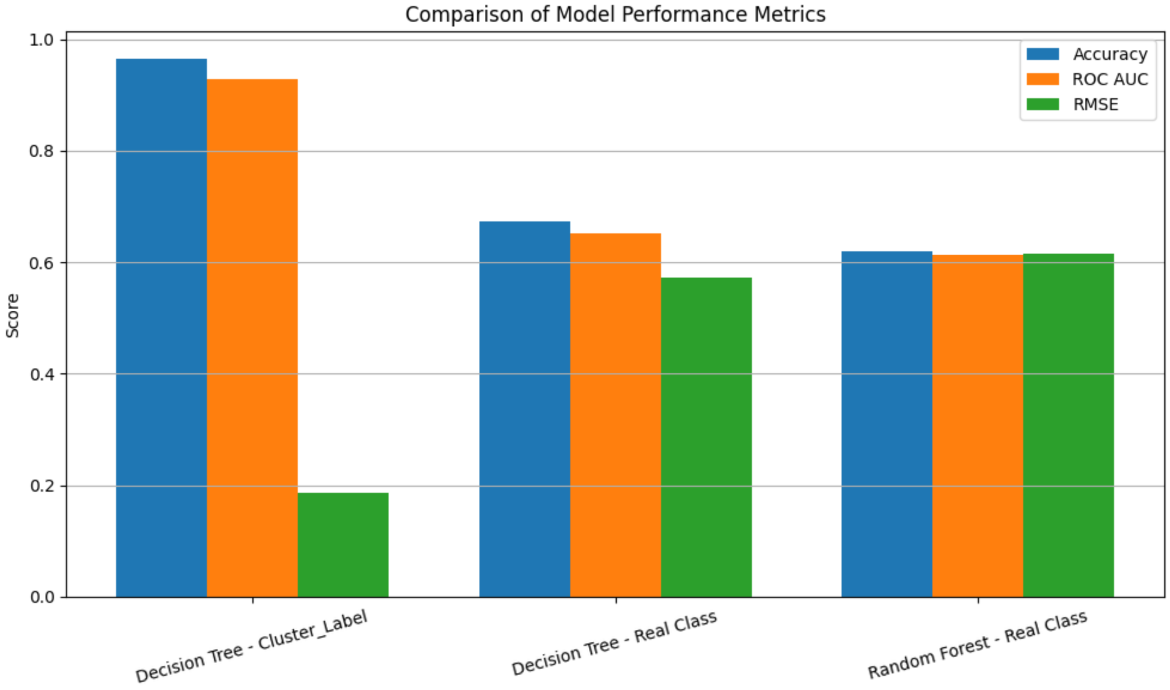
These results show that:

- The Decision Tree on Cluster_Label performed exceptionally (as expected).
- The Decision Tree outperformed Random Forest on real class prediction in both accuracy and AUC.
- Random Forest had slightly higher RMSE, suggesting more errors on minority cases.



```
📋 Model Performance Comparison (All Models)
                            Model  Accuracy   ROC AUC      RMSE
0  Decision Tree - Cluster_Label  0.965517  0.928571  0.185695
1     Decision Tree - Real Class  0.672414  0.652080  0.572351
2     Random Forest - Real Class  0.620690  0.614060  0.615882
```

Figure 13. Comparison of model performance metrics



7. Conclusion

This project demonstrated the practical application of clustering and classification techniques on breast cancer recurrence data. KMeans clustering revealed natural groupings in the data, while decision trees and random forests were trained to predict both generated clusters and real outcomes. The best predictive performance was observed in the Decision Tree model trained on true class labels, which achieved 67% accuracy and a 0.65 ROC AUC. Forecasting was successfully simulated by masking class labels, and predicted distributions were visualized. Overall, the combination of preprocessing, feature engineering, modeling, and evaluation provided a full pipeline aligned with the project objectives.