

Breast Cancer Clustering & Classification

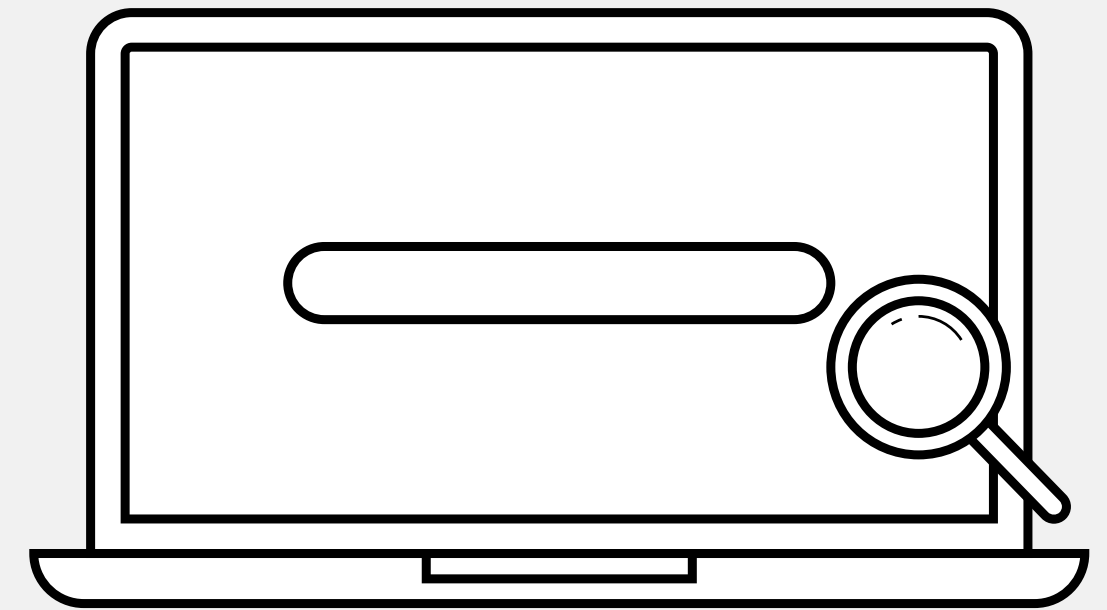
COMP 466 – Business Intelligence

Mustafa Eren Soyhan

042101008

Table Of Contents

- Introduction
- Project Objectives
- Dataset Overview
- Data Preparation
- Feature Engineering
- Clustering Analysis
- Cluster vs Actual Class Evaluation
- Classification Models & Results
- Forecasting Unknown Outcomes
- Model Comparison
- Conclusion



Introduction

This project explores how machine learning can be applied to predict breast cancer recurrence by combining both unsupervised and supervised methods. We used the breast-cancer.arff dataset, which includes a mix of clinical and demographic attributes. The goal was to discover patterns in the data through clustering, train classification models to learn those patterns, and forecast unknown outcomes effectively. The entire process—from data cleaning to model evaluation—was conducted in Python, with visualizations and metrics used to interpret model performance.



Objective

The project's objectives were to:

- Clean and preprocess the original dataset
- Apply clustering to generate a derived class column
- Use classification models to predict both clusters and real outcomes
- Simulate forecasting by hiding a portion of real labels
- Evaluate model performance with Accuracy, ROC AUC, and RMSE
- Present findings in a report and this presentation







Dataset Overview

The dataset consists of 286 patient records and 10 categorical attributes, such as age range, tumor size, number of involved lymph nodes, node-caps, and irradiation status. The target column, class, identifies whether a patient had recurrence or not. The dataset required transformation before machine learning models could be applied.

age	menopaus	tumor-size	inv-nodes	node-caps	breast	breast-qu	irradiat	deg-malig	Class	age_mid	tu
40-49	premeno	15-19	0-2	yes	right	left_up	no	3	recurrence	44.5	
50-59	ge40	15-19	0-2	no	right	central	no	1	no-recurrence	54.5	
50-59	ge40	35-39	0-2	no	left	left_low	no	2	recurrence	54.5	
40-49	premeno	35-39	0-2	yes	right	left_low	yes	3	no-recurrence	44.5	
40-49	premeno	30-34	5-Mar	yes	left	right_up	no	2	recurrence	44.5	
50-59	premeno	25-29	5-Mar	no	right	left_up	yes	2	no-recurrence	54.5	
50-59	ge40	40-44	0-2	no	left	left_up	no	3	no-recurrence	54.5	
40-49	premeno	14-Oct	0-2	no	left	left_up	no	2	no-recurrence	44.5	
40-49	premeno	0-4	0-2	no	right	right_low	no	2	no-recurrence	44.5	
40-49	ge40	40-44	15-17	yes	right	left_up	yes	2	no-recurrence	44.5	
50-59	premeno	25-29	0-2	no	left	left_low	no	2	no-recurrence	54.5	
60-69	ge40	15-19	0-2	no	right	left_up	no	2	no-recurrence	64.5	

irradiat_n	irradiat_y	age_range	tumor_size	inv_nodes	age_label	tumor_size	inv_nodes	malignancy_score
True	False	9	4	2	2	2	0	51
True	False	9	4	2	3	2	0	17
True	False	9	4	2	3	6	0	74
False	True	9	4	2	2	6	0	111
True	False	9	4	2	2	5	4	64
False	True	9	4	2	3	4	4	54
True	False	9	4	2	3	7	0	126
True	False	9	4	2	2	1	0	24
True	False	9	4	2	2	0	0	4
False	True	9	4	2	2	7	2	84
True	False	9	4	2	3	4	0	54



Data Preparation

Data Preparation

The original .arff file was parsed and converted to a pandas DataFrame. Missing values were identified in node-caps (8 rows) and breast-quad (1 row). These were imputed using the mode of the respective column. Additional cleaning included removing inconsistent quotes and fixing string formatting issues to ensure compatibility with encoding and model training.

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



Feature Engineering

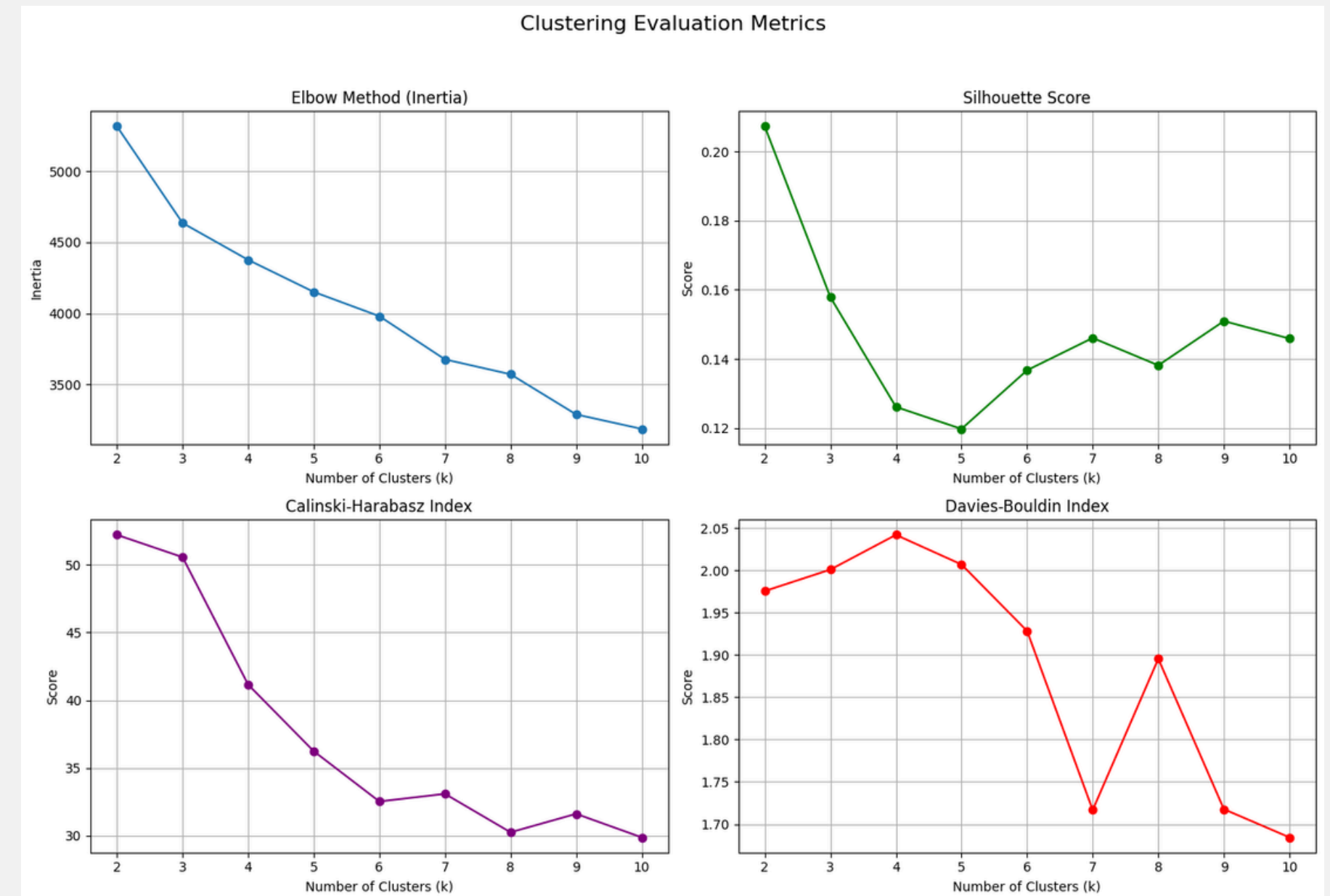
To make the dataset suitable for machine learning models:

- Ranged values like age, tumor-size, and inv-nodes were converted to numeric midpoints.
- Categorical features were one-hot encoded.
- A new binary target column `class_binary` was created:
 - 0 = No Recurrence
 - 1 = Recurrence

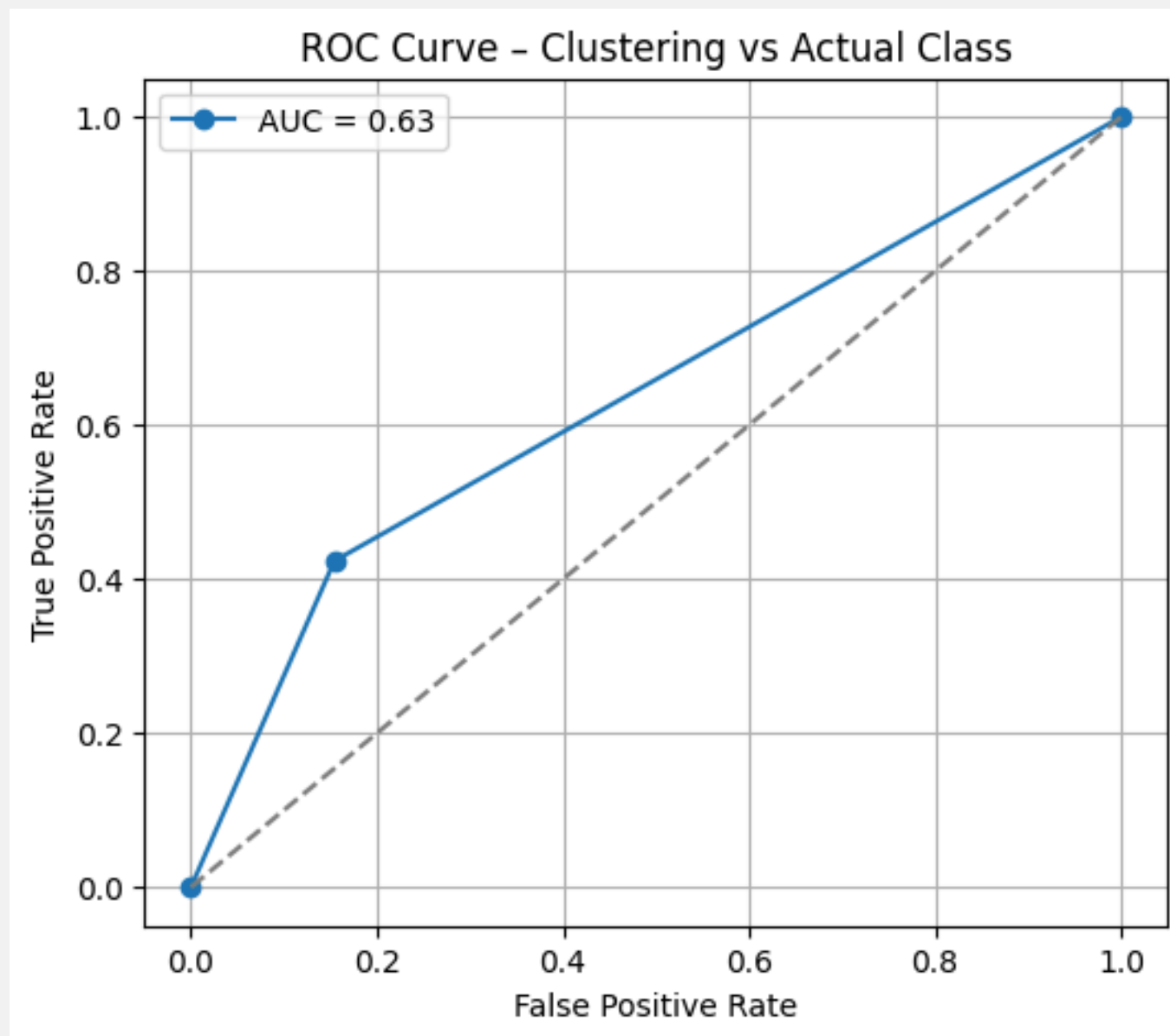
This transformation resulted in a clean, fully numeric dataset ready for clustering and classification.

Clustering

- KMeans clustering was applied to identify natural groupings in the dataset without using any target label. The number of clusters (k) was optimized by evaluating several internal metrics including the Elbow Method, Silhouette Score, Calinski-Harabasz Index, and Davies-Bouldin Index.
- These metrics consistently indicated that k = 2 was optimal, which aligns with the known binary outcome (recurrence vs no recurrence). The resulting cluster assignments were saved as Cluster_Label.



Clustering vs Actual Class



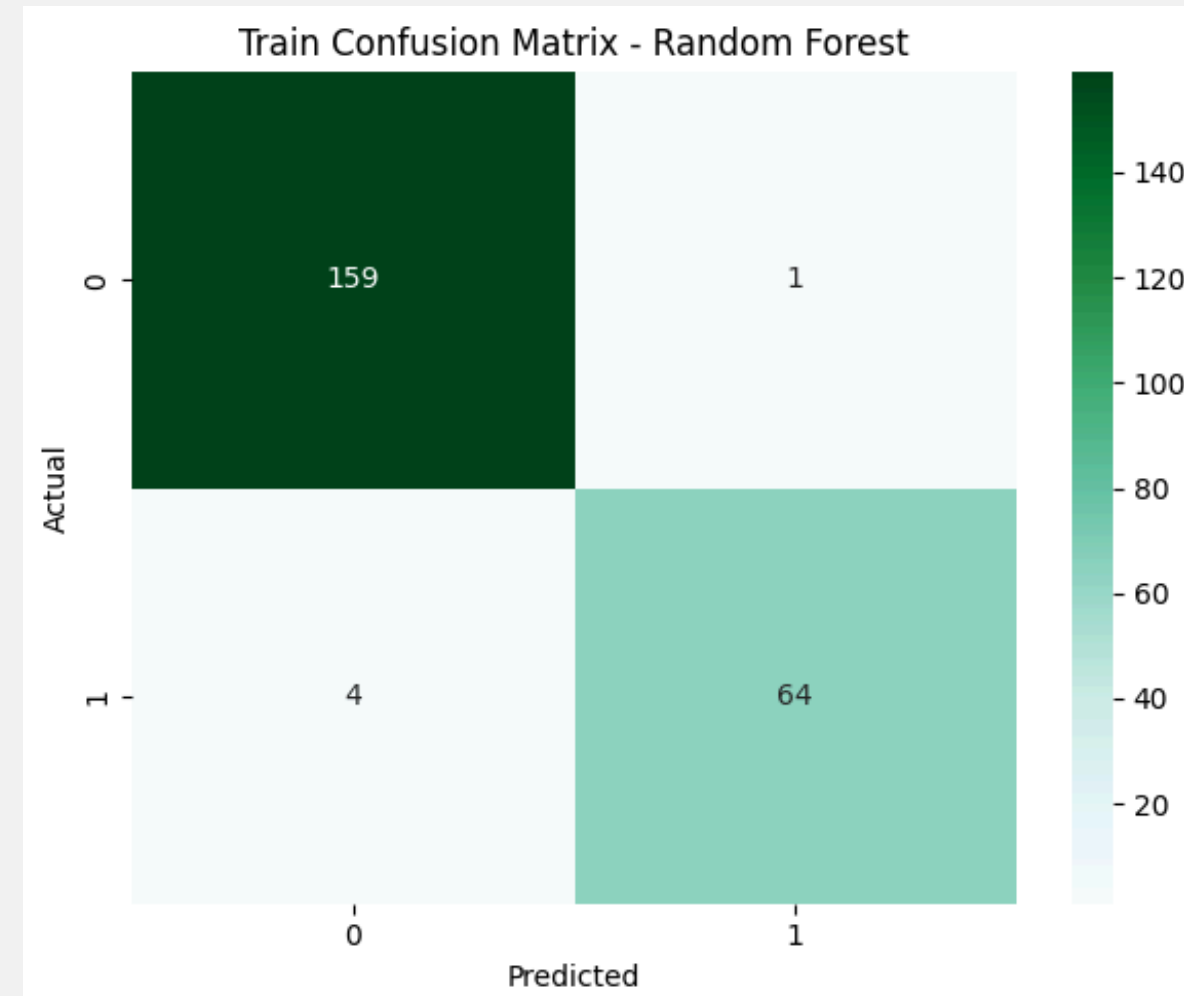
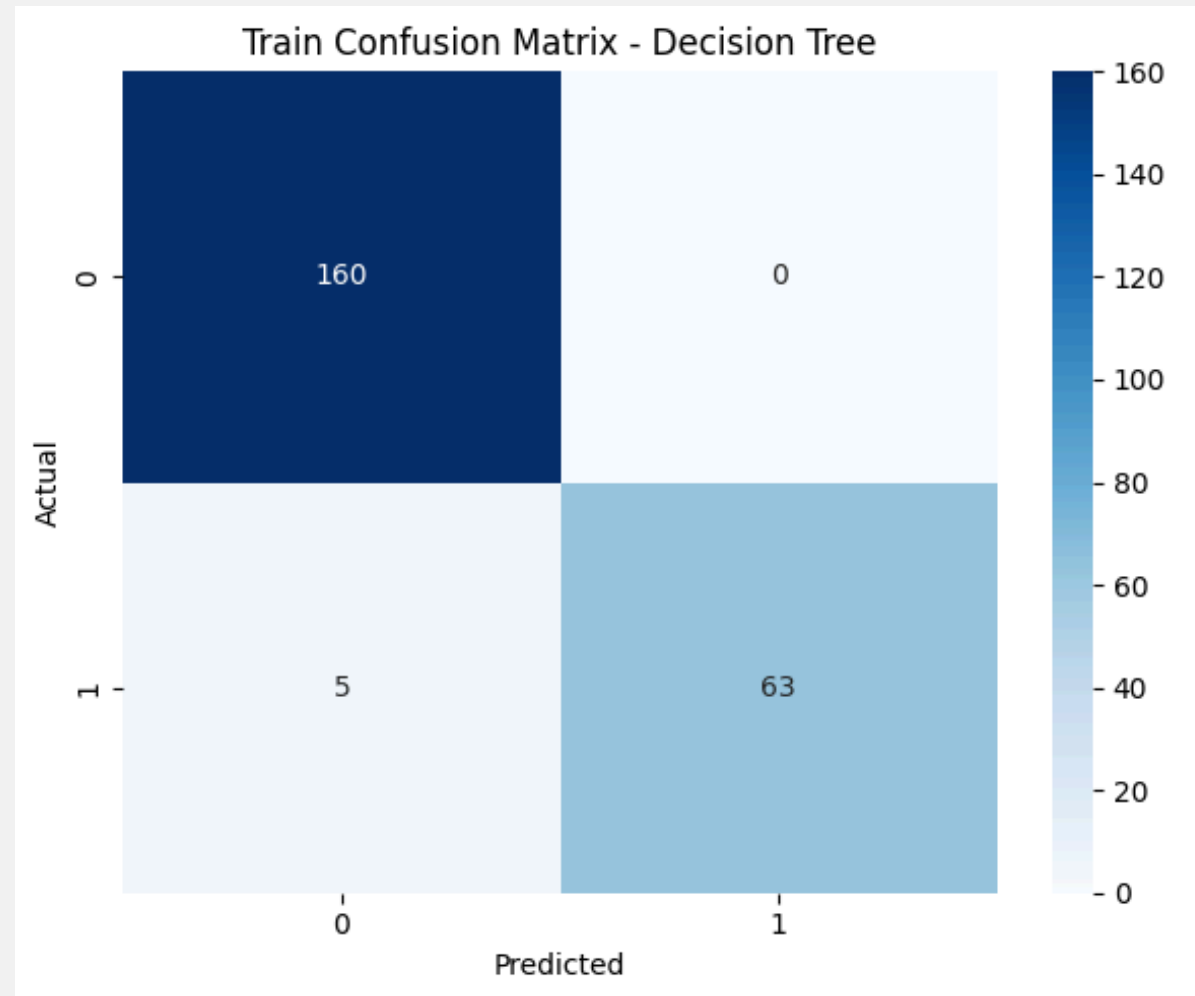
To assess the validity of the clusters, we compared the Cluster_Label column to the actual class_binary label. The labels were flipped where needed to maximize alignment. After adjustment:

- **Accuracy: 0.72**
- **ROC AUC: 0.63**
- The clusters aligned well with the non-recurrence class but were less precise for recurrence cases, as shown in the ROC and confusion matrix visuals.

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

Classification Results

Two models were used to predict the true recurrence class:



Decision Tree

Decision Tree Classifier

- Accuracy: 0.67
- ROC AUC: 0.65
- RMSE: 0.57

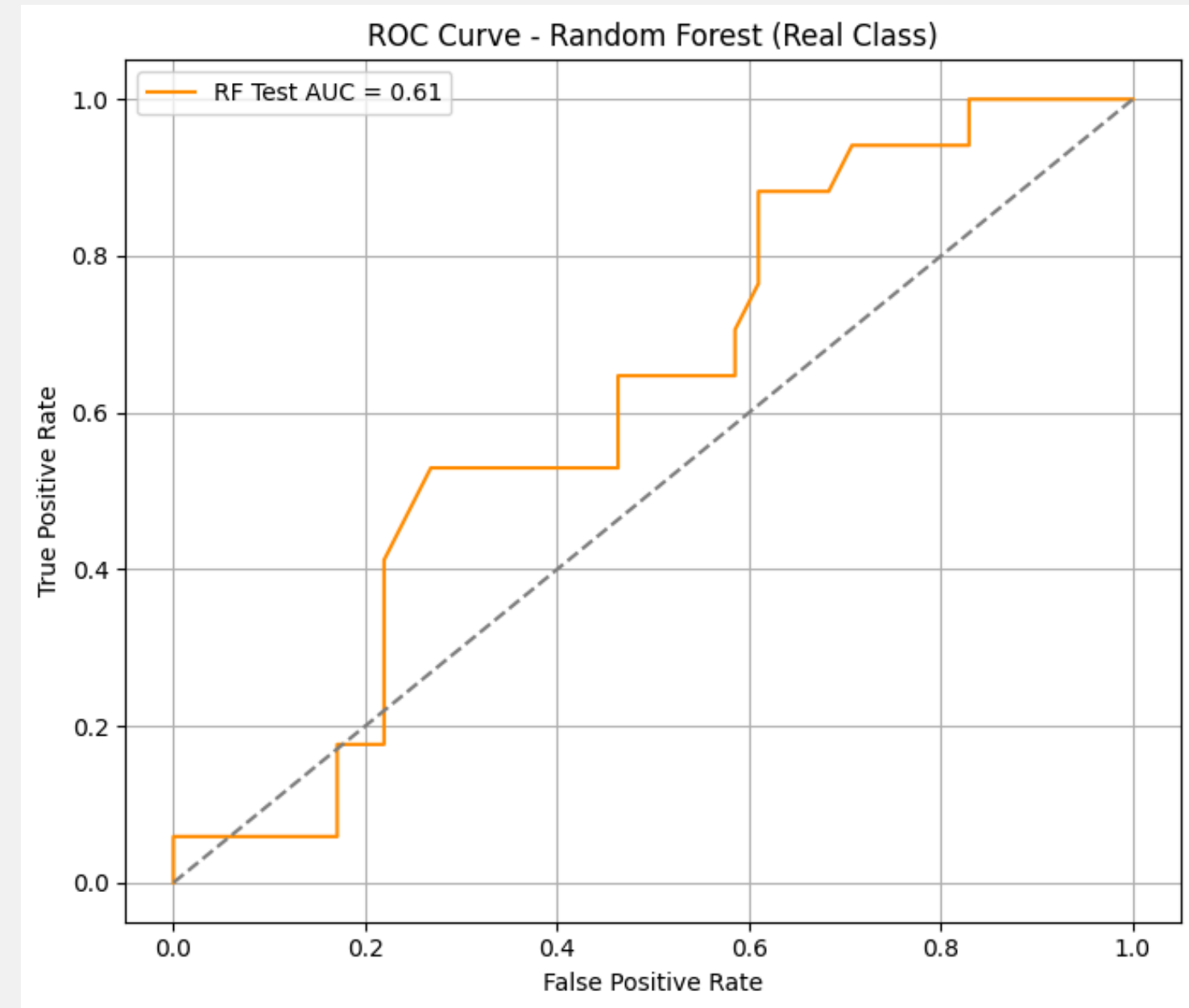
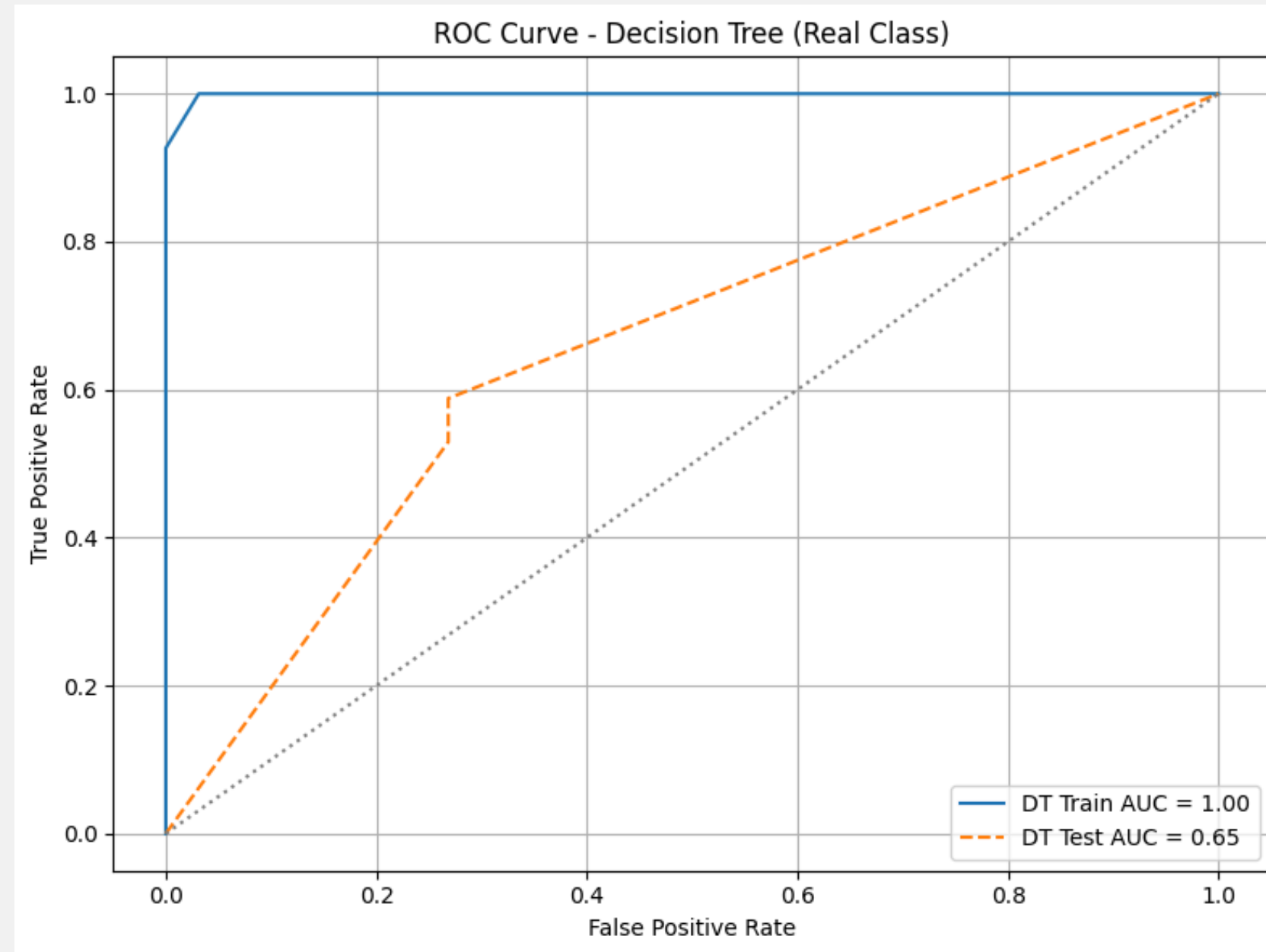
Random Forest

Random Forest Classifier

- Accuracy: 0.62
- ROC AUC: 0.61
- RMSE: 0.62

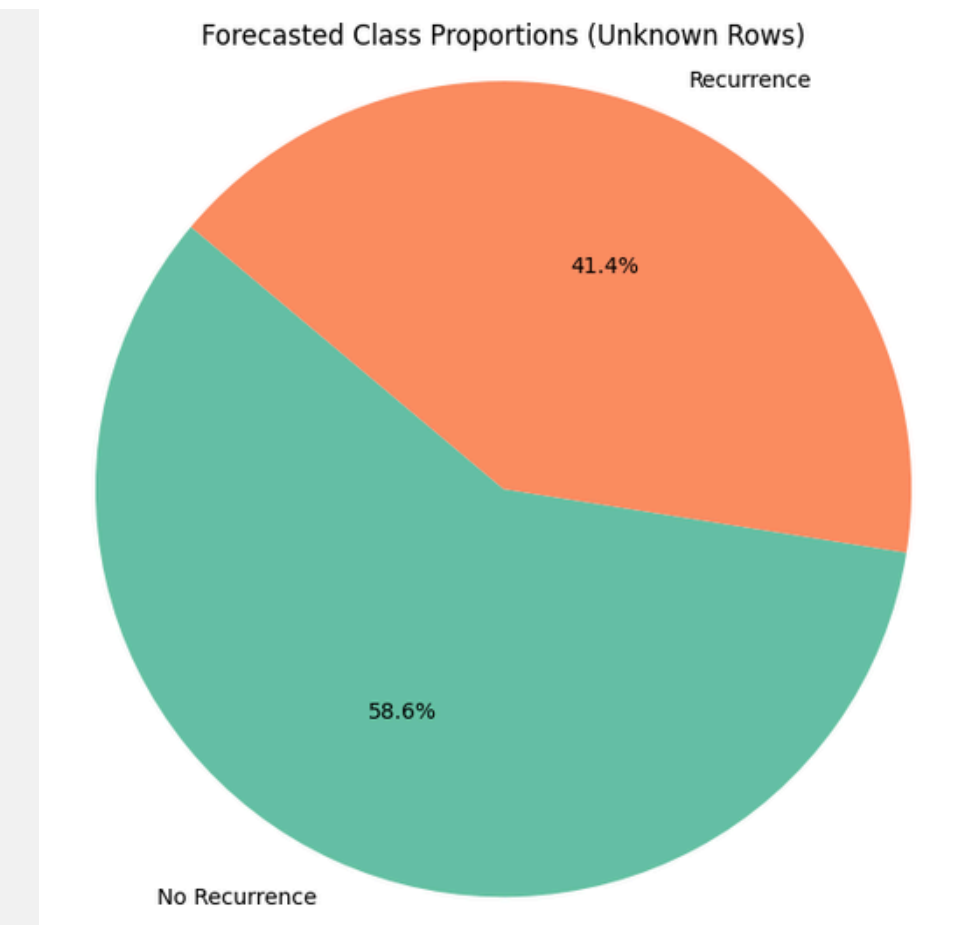
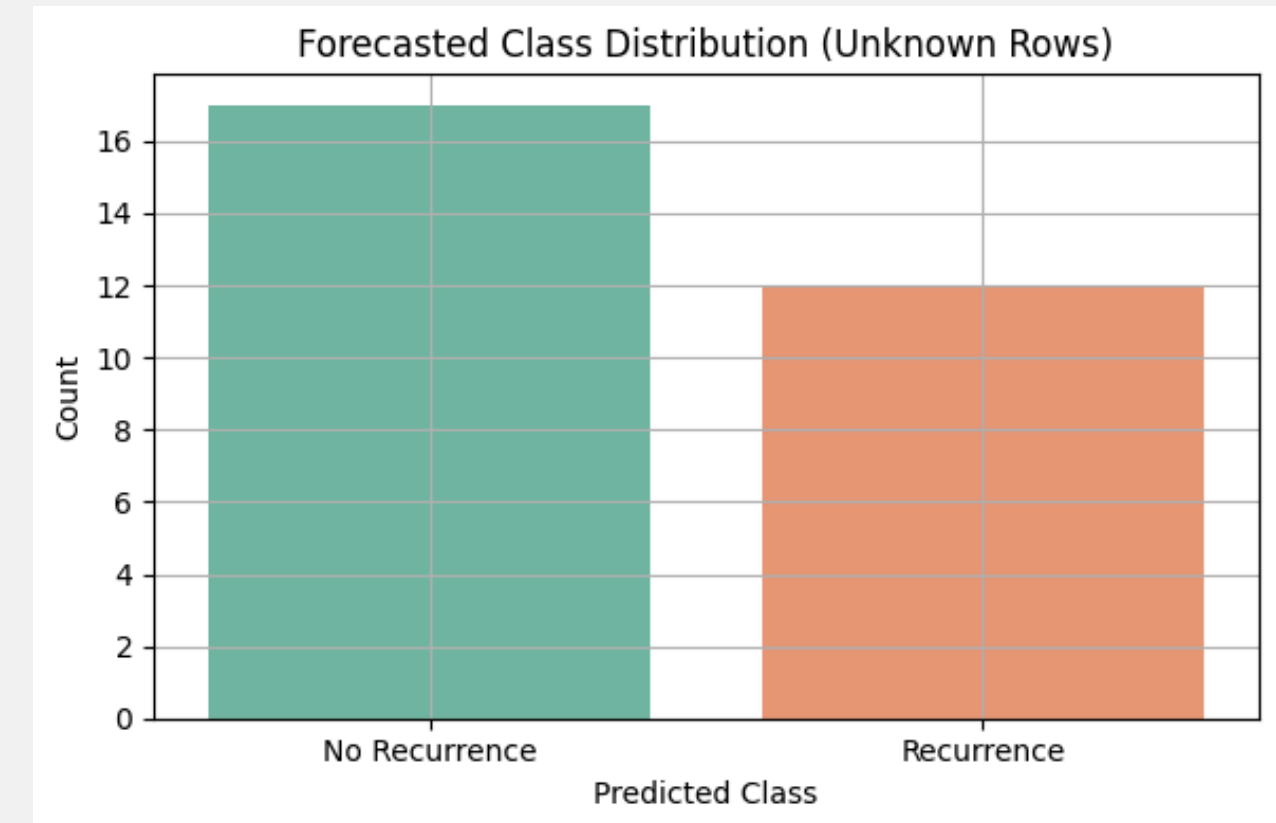
Decision Tree performed better overall, especially in recognizing the recurrence class. Both models struggled with minority class precision due to class imbalance.

Classification ROC Results

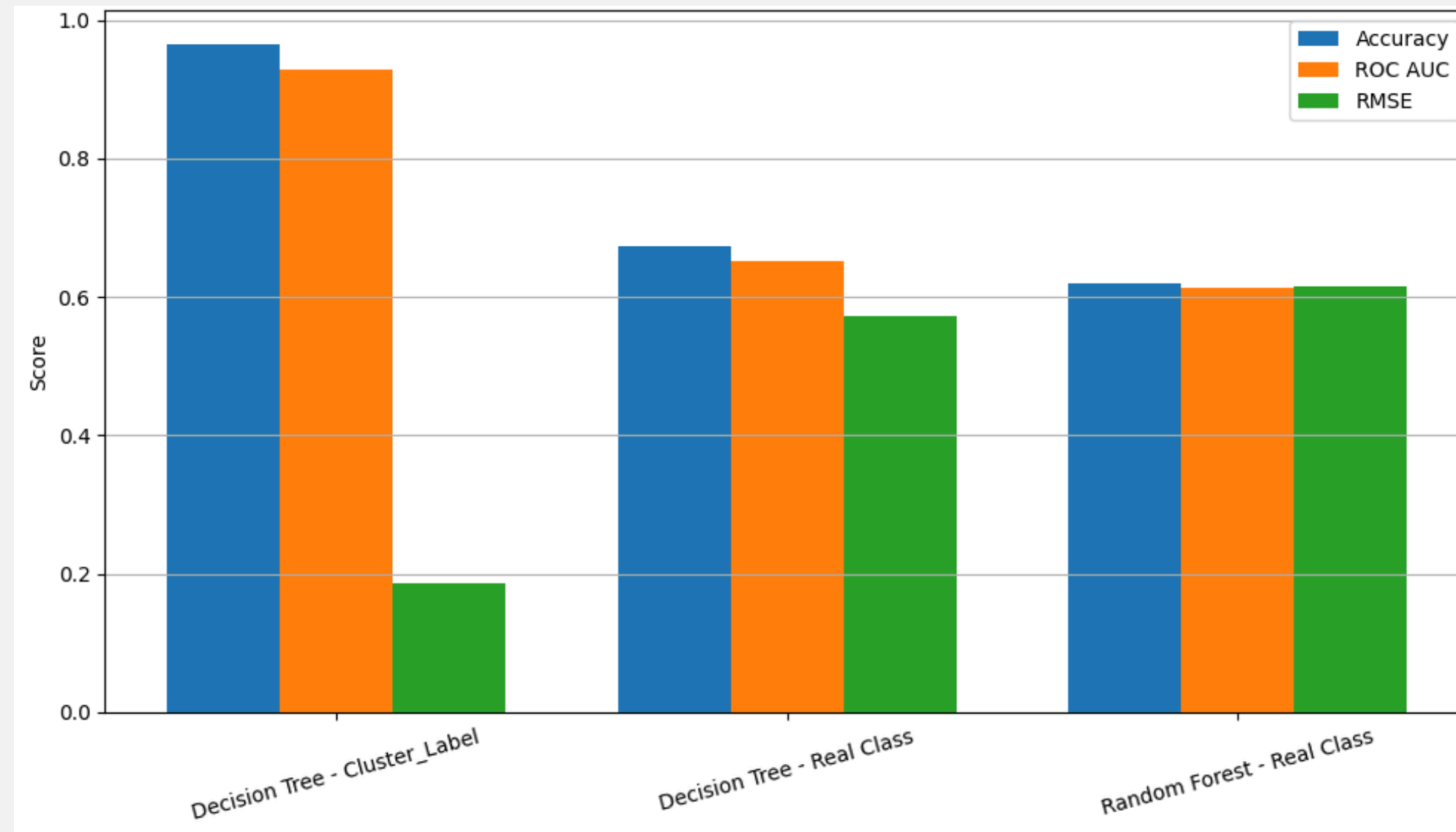


Forecasting Unknowns

To simulate a real-world prediction task, 10% of the `class_binary` values were masked to represent unknown or missing outcomes. A trained Decision Tree model was then used to forecast these values. The predictions were distributed across both classes, with approximately 59% predicted as "No Recurrence" and 41% as "Recurrence". This step tested the model's generalization ability and illustrated how it might behave in a production setting with live data.



Comparison of Models



All models were compared using Accuracy, ROC AUC, and RMSE. Decision Tree on Cluster_Label unsurprisingly performed best. When predicting real recurrence, Decision Tree also slightly outperformed Random Forest on all metrics.

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

Conclusion

This project applied machine learning techniques to analyze and forecast breast cancer recurrence. Clustering revealed meaningful patterns in the data, while classification models were able to predict both derived clusters and actual recurrence labels with reasonable accuracy. Among the models tested, Decision Tree performed best for real class prediction. Forecasting on unknown data showed the model's potential in real-world applications. Overall, the results demonstrate the effectiveness of combining unsupervised and supervised learning for medical outcome prediction.



The background is a light gray color decorated with various hand-drawn blue doodles. These include several loops and swirls at the top, a star-like shape on the right, a wavy line at the bottom center, and several checkmarks at the bottom right. There are also some scribbled shapes on the left and bottom left.

**Thank you
very much!**