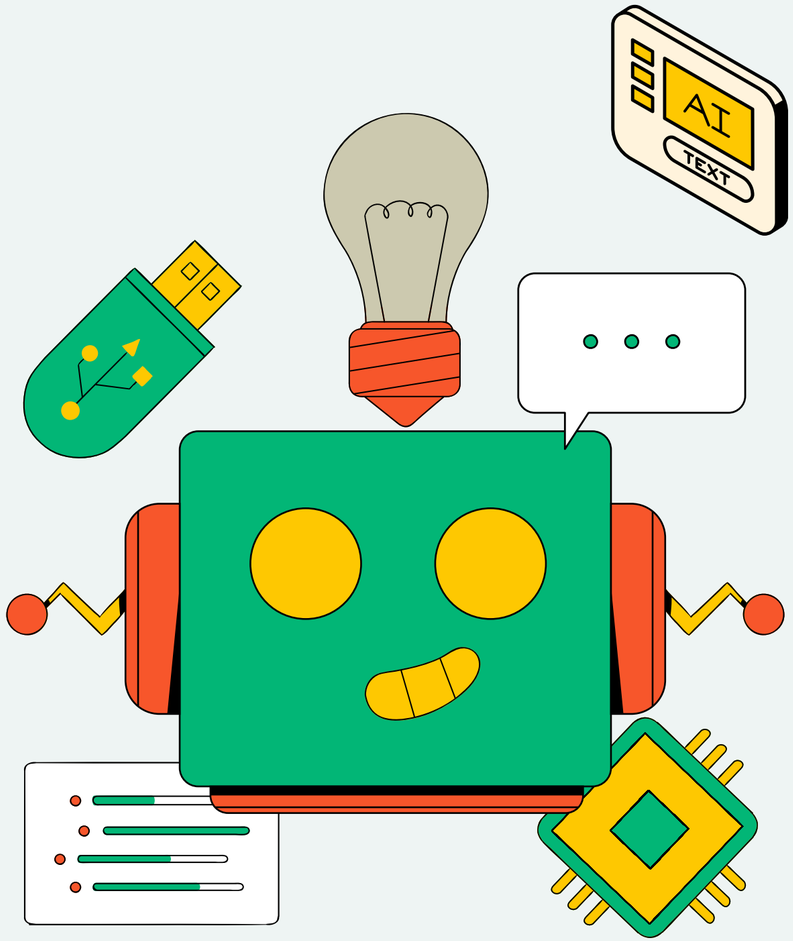
MEF UNIVERSITY

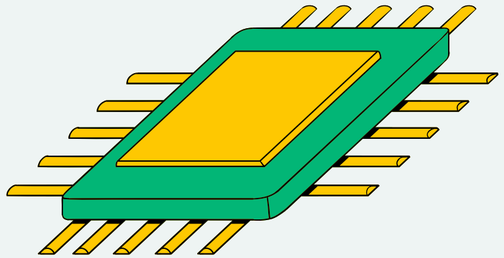
YOUR FREEDOM IN LEARNING



COMP 466 BUSINESS INTELLIGENCE ACTIVITY 1 PRESENTATION

PRESENTED BY:

TARIK BUĞRA AY



PRESENTATION OUTLINE

- Introduction
- Project Objective
- Dataset Overview
- Data Preparation
- Feature Engineering
- Clustering
- Clustering vs Actual Class
- Classification Results
- Model Comparison
- Conclusion

INTRODUCTION

This project explores how machine learning can be applied to analyze breast cancer data using both clustering and classification techniques.

The goal is to:

- **Discover natural groupings in the data**
- **Build predictive models to classify outcomes**
- **Evaluate and compare performance using common ML metrics**

OBJECTIVE OF THE PROJECT

The aim of this project is to apply machine learning techniques to a breast cancer dataset to:

- Identify natural groupings using clustering algorithms
- Use those clusters as pseudo-labels for classification
- Compare results with real class labels
- Evaluate models using:
  - Accuracy
  - ROC AUC Score
  - RMSE

DATASET OVERVIEW

- Dataset Name: ` breast_cancer.arff`
- Source: University Medical Centre, Institute of Oncology, Ljubljana
- Instances: 286 patients
- Features: 9 input attributes + 1 class label
- Class Labels:
  - no-recurrence-events
  - recurrence-events

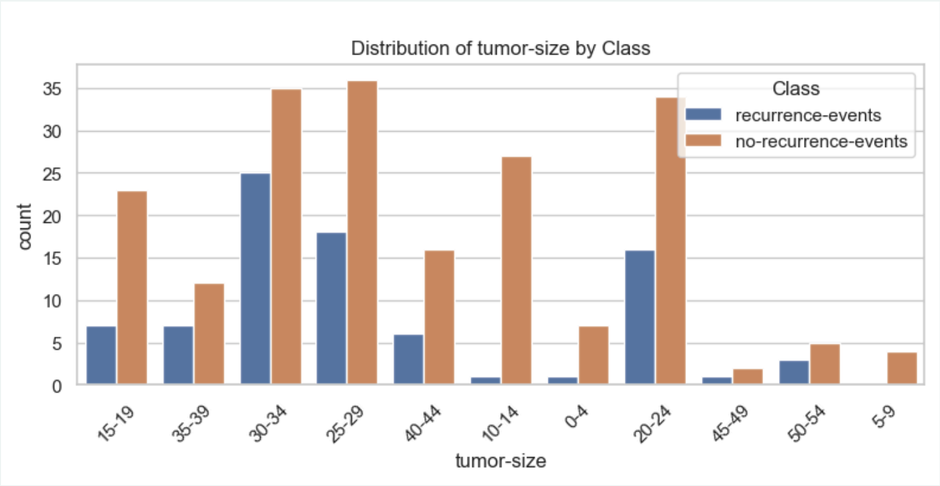


Figure 1. Distribution of tumor size by class.

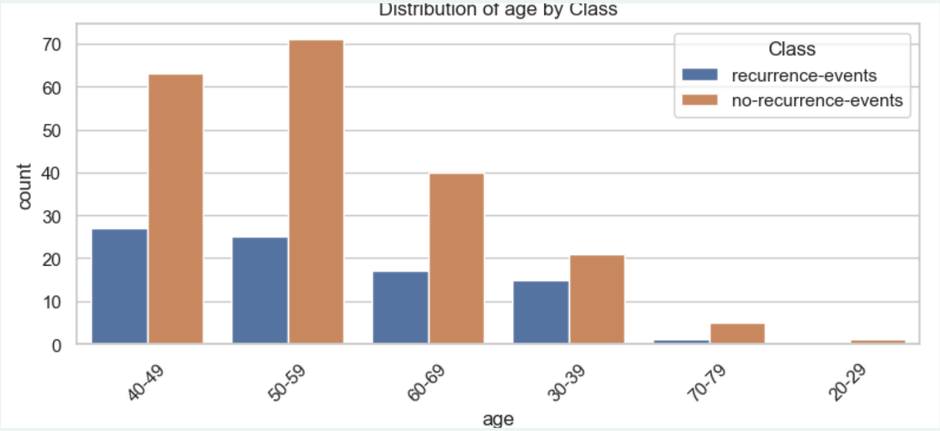


Figure 2. Distribution of age by class.

DATA PREPARATION

Steps Taken to Prepare the Data

- Parsed `.arff` file manually to extract data and attributes
- Handled missing values:
  - `node-caps`: filled with mode
  - `breast-quad`: filled with mode
- Removed extra quotes and cleaned column names
- Saved cleaned version as CSV for further processing

- Resulting dataset:
  - Shape: 286 rows × 10 columns

```
age,menopause,tumor-size,inv-nodes,node-caps,deg-malig,breast,breast-quad,irradiat,Class
40-49,premeno,15-19,0-2,yes,3,right,left_up,no,recurrence-events
50-59,ge40,15-19,0-2,no,1,right,central,no,no-recurrence-events
50-59,ge40,35-39,0-2,no,2,left,left_low,no,recurrence-events
40-49,premeno,35-39,0-2,yes,3,right,left_low,yes,no-recurrence-events
40-49,premeno,30-34,3-5,yes,2,left,right_up,no,recurrence-events
50-59,premeno,25-29,3-5,no,2,right,left_up,yes,no-recurrence-events
50-59,ge40,40-44,0-2,no,3,left,left_up,no,no-recurrence-events
40-49,premeno,10-14,0-2,no,2,left,left_up,no,no-recurrence-events
40-49,premeno,0-4,0-2,no,2,right,right_low,no,no-recurrence-events
40-49,ge40,40-44,15-17,yes,2,right,left_up,yes,no-recurrence-events
50-59,premeno,25-29,0-2,no,2,left,left_low,no,no-recurrence-events
60-69,ge40,15-19,0-2,no,2,right,left_up,no,no-recurrence-events
50-59,ge40,30-34,0-2,no,1,right,central,no,no-recurrence-events
50-59,ge40,25-29,0-2,no,2,right,left_up,no,no-recurrence-events
40-49,premeno,25-29,0-2,no,2,left,left_low,yes,recurrence-events
30-39,premeno,20-24,0-2,no,3,left,central,no,no-recurrence-events
50-59,premeno,10-14,3-5,no,1,right,left_up,no,no-recurrence-events
60-69,ge40,15-19,0-2,no,2,right,left_up,no,no-recurrence-events
50-59,premeno,40-44,0-2,no,2,left,left_up,no,no-recurrence-events
50-59,ge40,20-24,0-2,no,3,left,left_up,no,no-recurrence-events
50-59,lt40,20-24,0-2,no,1,left,left_low,no,recurrence-events
60-69,ge40,40-44,3-5,no,2,right,left_up,yes,no-recurrence-events
50-59,ge40,15-19,0-2,no,2,right,left_low,no,no-recurrence-events
40-49,premeno,10-14,0-2,no,1,right,left_up,no,no-recurrence-events
30-39,premeno,15-19,6-8,yes,3,left,left_low,yes,recurrence-events
```

Figure 3. A snippet from data.

FEATURE ENGINEERING

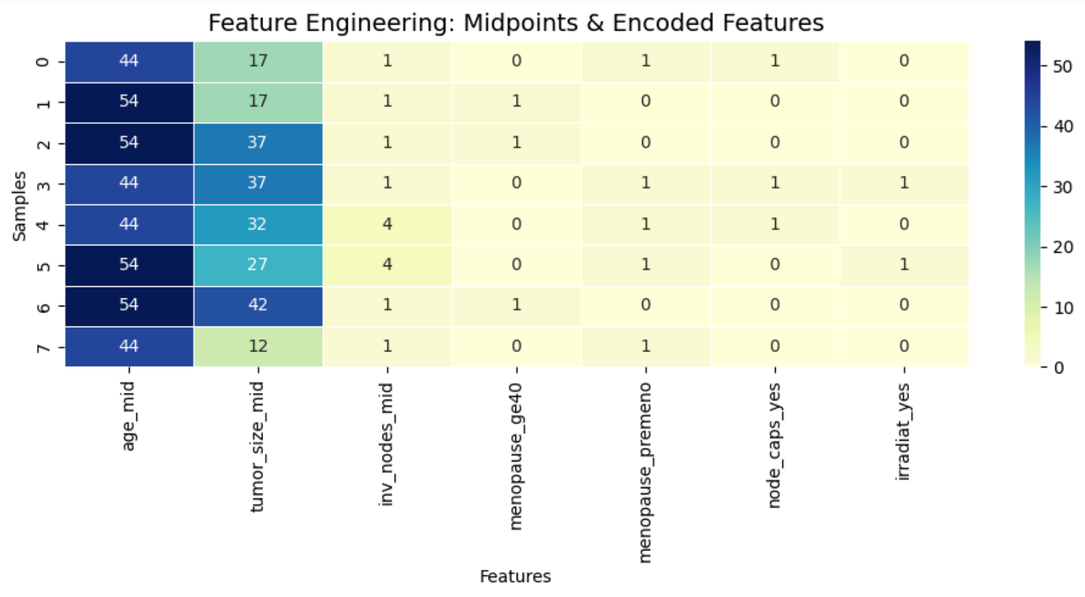


Figure 4. Feature Engineered Columns

Feature Engineering Steps

To prepare the dataset for clustering and classification, we:

- bConverted categorical ranges to numeric midpoints:
  - `age`, `tumor-size`, `inv-nodes`
- Applied one-hot encoding to:
  - `menopause`, `node-caps`, `irradiat`
  - `breast`, `breast-quad`
- Retained `deg-malig` as numeric

Final dataset:
→ 286 rows × 27 columns

CLUSTERING

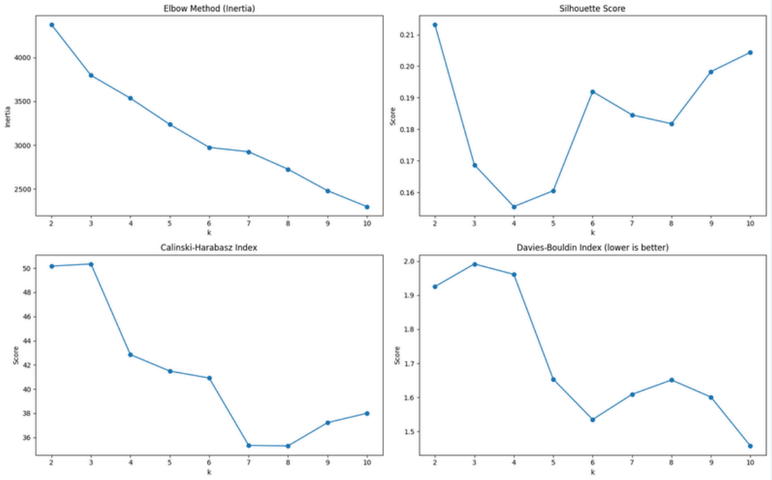


Figure 5. Evaluation metric chart (4 subplots) — Elbow, Silhouette, CH, DBI

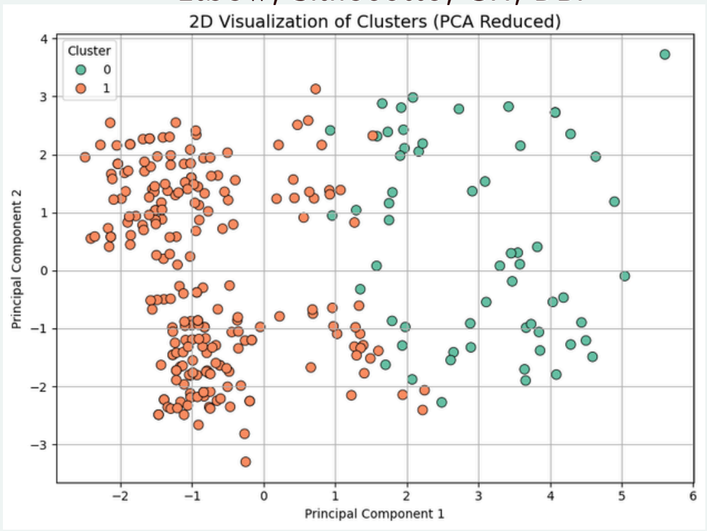


Figure 6. PCA cluster visualization

Clustering with KMeans

- Standardized numeric features before clustering
- Applied KMeans to discover natural groupings in the data
- Evaluated multiple values of ˜ k˜ using:
  - Elbow Method (Inertia)
  - Silhouette Score
  - Calinski–Harabasz Index
  - Davies–Bouldin Index

Chosen Number of Clusters: ˜ k = 2˜

- Best balance across metrics
- Matches the known recurrence/no-recurrence classes
- Simple for downstream analysis

CLUSTERING VS ACTUAL CLASS

Evaluating Clustering Against Actual Class

- After clustering, we compared the generated cluster labels with the real medical outcomes (`Class` column).

- Mapped clusters to match the closest class distribution
- Adjusted cluster labels where needed
- Evaluation metrics:

- Accuracy: 0.71
- ROC AUC Score: 0.69
- RMSE: 0.46

- This showed partial alignment between clusters and real-world outcomes.

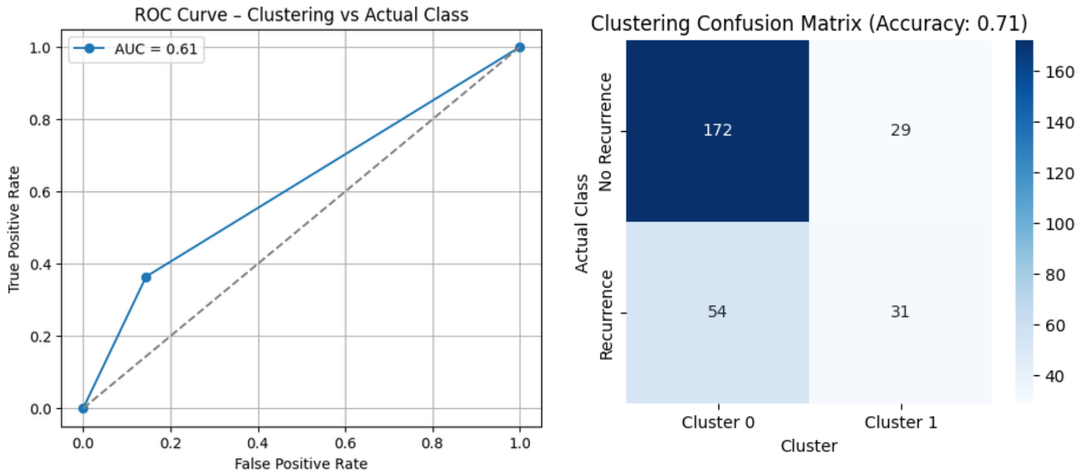


Figure 7. ROC Curve and Confusion Matrix

CLASSIFICATION RESULTS

Decision Tree Classifier Results

We trained two classification models using the same features:

1. Model 1 – Predicting Cluster Labels
   - Accuracy: 1.00
   - ROC AUC: 1.00
   - RMSE: 0.00

2. Model 2 – Predicting Actual Class Labels
   - Accuracy: 0.71
   - ROC AUC: 0.64
   - RMSE: 0.53

Model 1 replicated clustering perfectly.
Model 2 showed more realistic performance .

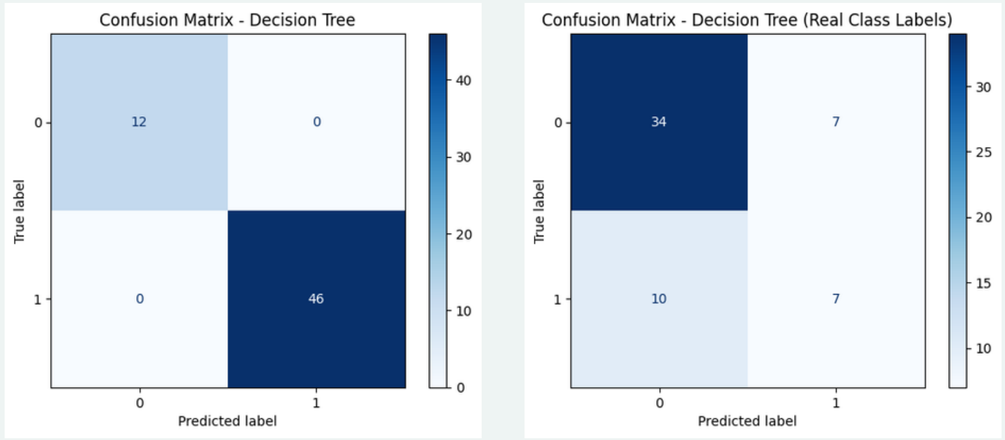


Figure 8. Confusion Matrixes for Classifications.

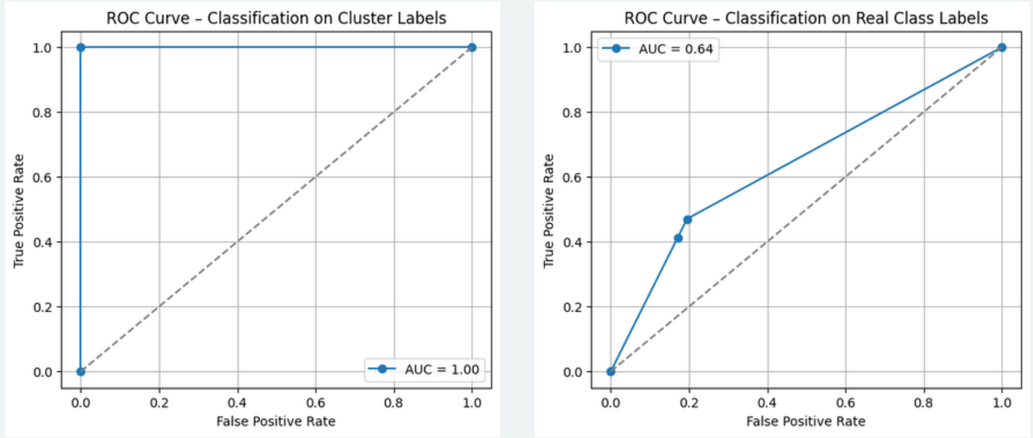


Figure 9. ROC Curves for Classifications.

CONCLUSION

- KMeans clustering revealed patterns in the breast cancer dataset that partially aligned with real recurrence labels.
- Classification models performed very differently:
  - Cluster labels were easy to learn and replicate.
  - Real class labels were harder to predict but gave more realistic results.
- Feature engineering played a key role in shaping both clustering and classification performance.

🔗 Project Repository GitHub:

- https://github.com/TarikBugraAy/breast-cancer-clustering-classification?tab=readme-ov-file