



UNIVERSITÀ  
DEGLI STUDI  
DI MILANO

# BREAST CANCER DIAGNOSIS

2025-2026

**Lorenzo Guidi 41301A**  
**Jaime Carnevale 32797A**  
**Umberto Mitolo 41384A**



# CANCRO AL SENO

## COS'È?

**Il tumore al seno è una malattia in cui le cellule della ghiandola mammaria crescono in modo anomalo e incontrollato, formando un tumore che può invadere i tessuti vicini e, se maligno, diffondersi ad altre parti del corpo.**



# OBIETTIVI

- **COMPRENDERE LA STRUTTURA DEL DATASET TRAMITE VISUALIZZAZIONE**
- **VALUTARE SE ALGORITMI DI CLUSTERING RIESCONO A SEPARARE NATURALMENTE LE DUE CLASSI**
- **ADDESTRARE UN CLASSIFICATORE PER DISTINGUERE TUMORI BENIGNI E MALIGNI**
- **COLLEGARE QUESTI PROCESSI AI MECCANISMI COGNITIVI DELLA PERCEZIONE.**



# FASI

1.VISUALIZZAZIONE  
DATI

2.CLUSTERING  
NON SUPERVISIONATO

3.CLASSIFICAZIONE  
SUPERVISIONATA

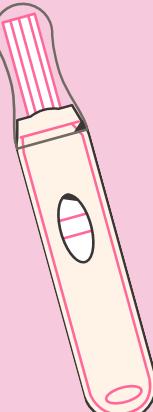
# VISUALIZZAZIONE

## IL NOSTRO DATASET

Per questo progetto abbiamo utilizzato un dataset messo a disposizione dal Winsconsin University Hospital, che contiene **569 istanze** che rappresentano le masse tumorali prelevate dalle pazienti. Per quanto riguarda il numero di **features** invece ne abbiamo **30** tutte di tipo **continuo**.

Come classi abbiamo:

- Maligno (M)**
- Benigno (B)**

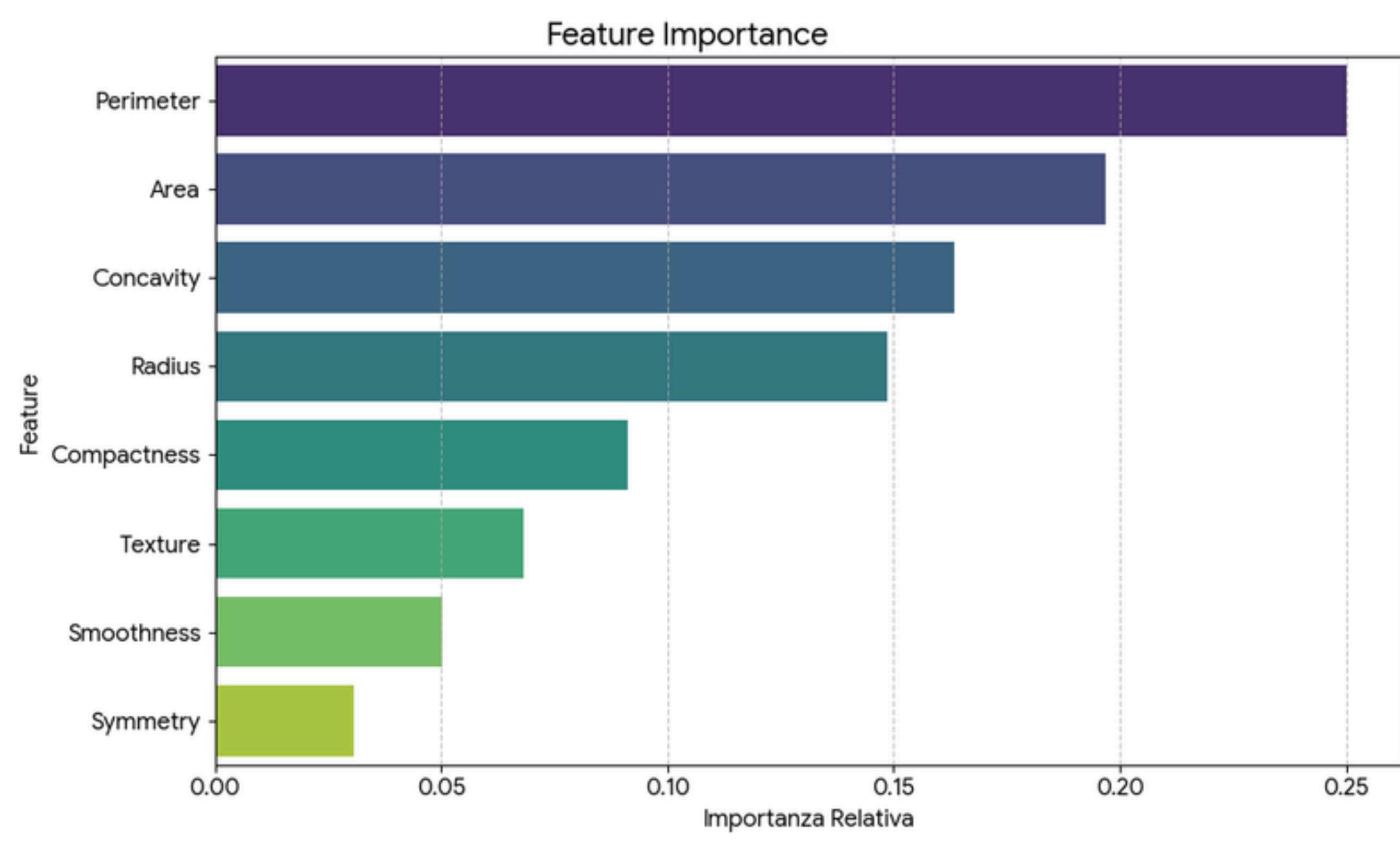


# FEATURES TOTALI

Elenco Completo delle 30 Features (Breast Cancer Wisconsin Diagnostic)

Mean (Media)	Standard Error (Errore Std)	Worst (Peggiore)
mean radius	radius error	worst radius
mean texture	texture error	worst texture
mean perimeter	perimeter error	worst perimeter
mean area	area error	worst area
mean smoothness	smoothness error	worst smoothness
mean compactness	compactness error	worst compactness
mean concavity	concavity error	worst concavity
mean concave points	concave points error	worst concave points
mean symmetry	symmetry error	worst symmetry
mean fractal dimension	fractal dimension error	worst fractal dimension

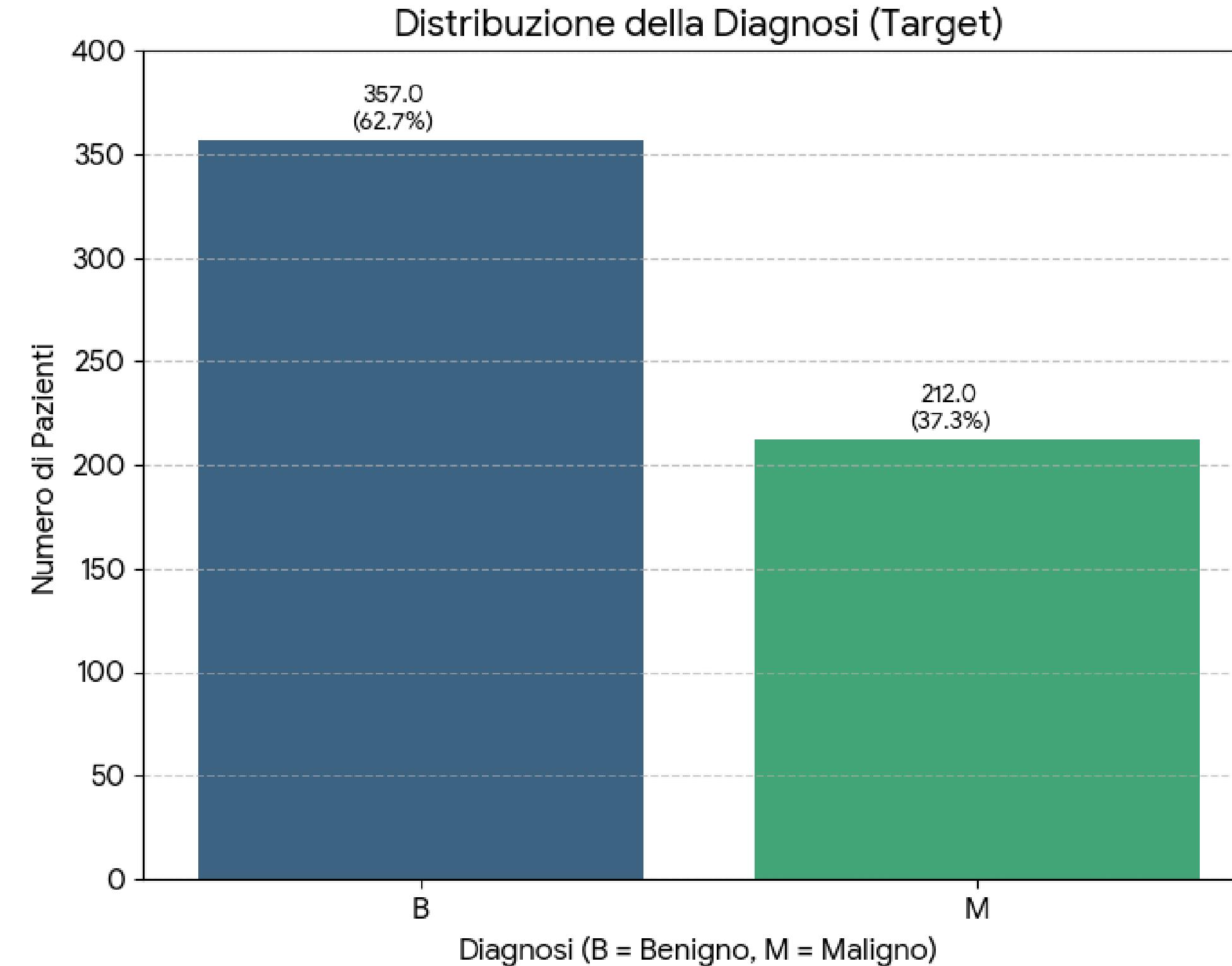
# FEATURES SELEZIONATE



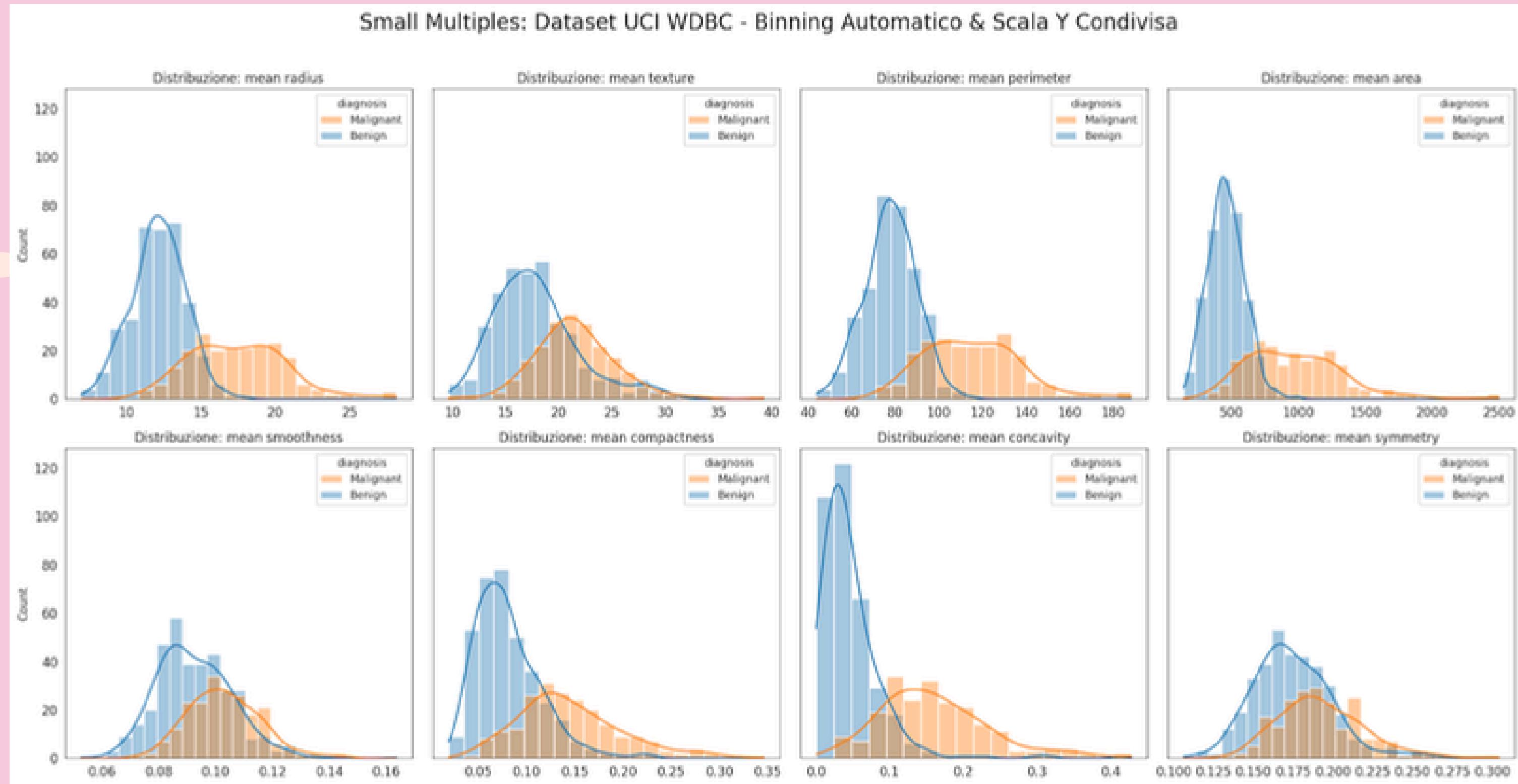
- PERIMETER
- AREA
- CONCAVITY
- RADIUS
- COMPACTNESS
- TEXTURE
- SMOOTHNESS
- SYMMETRY



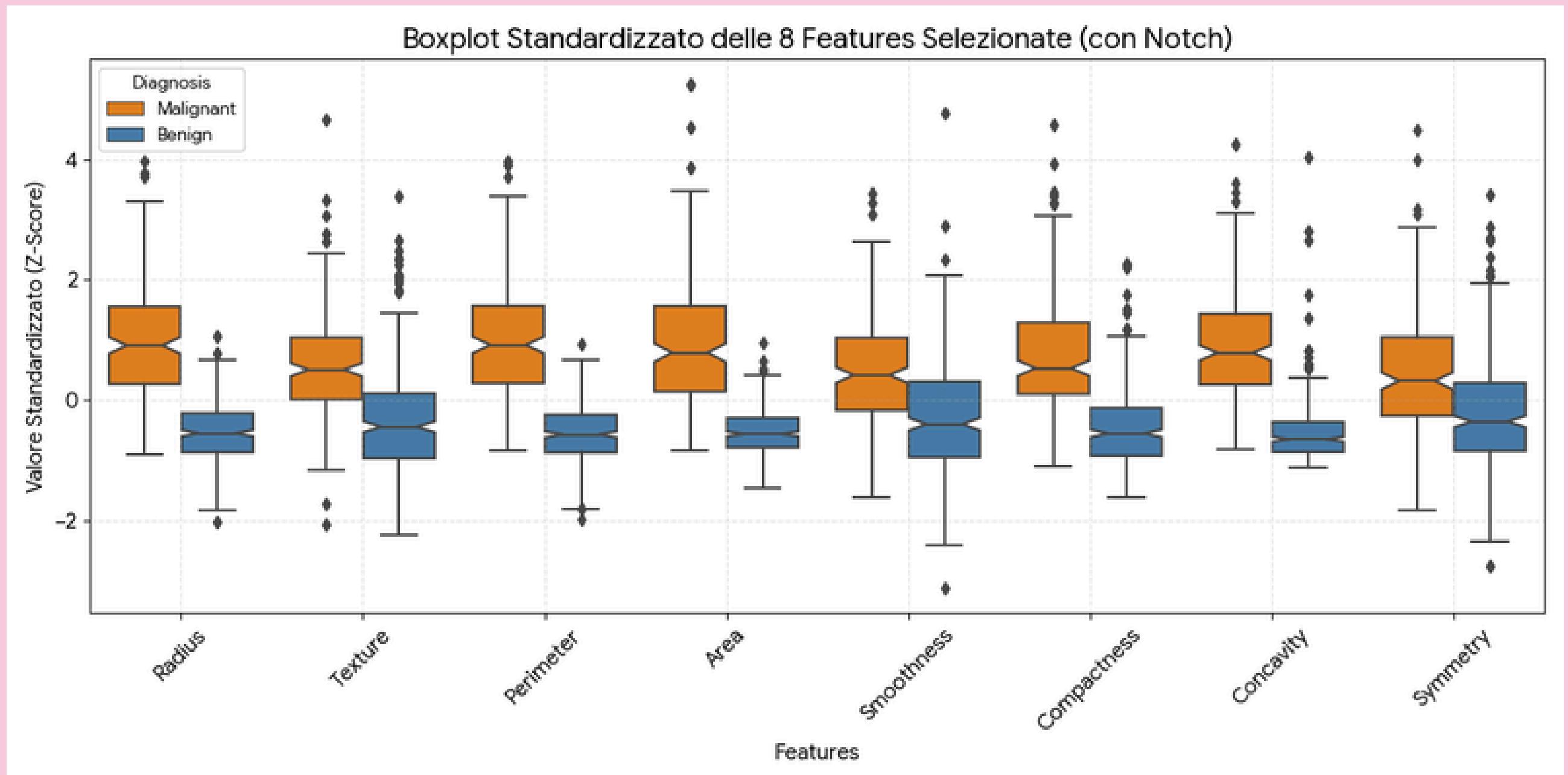
# DISTRIBUZIONE DEL TARGET



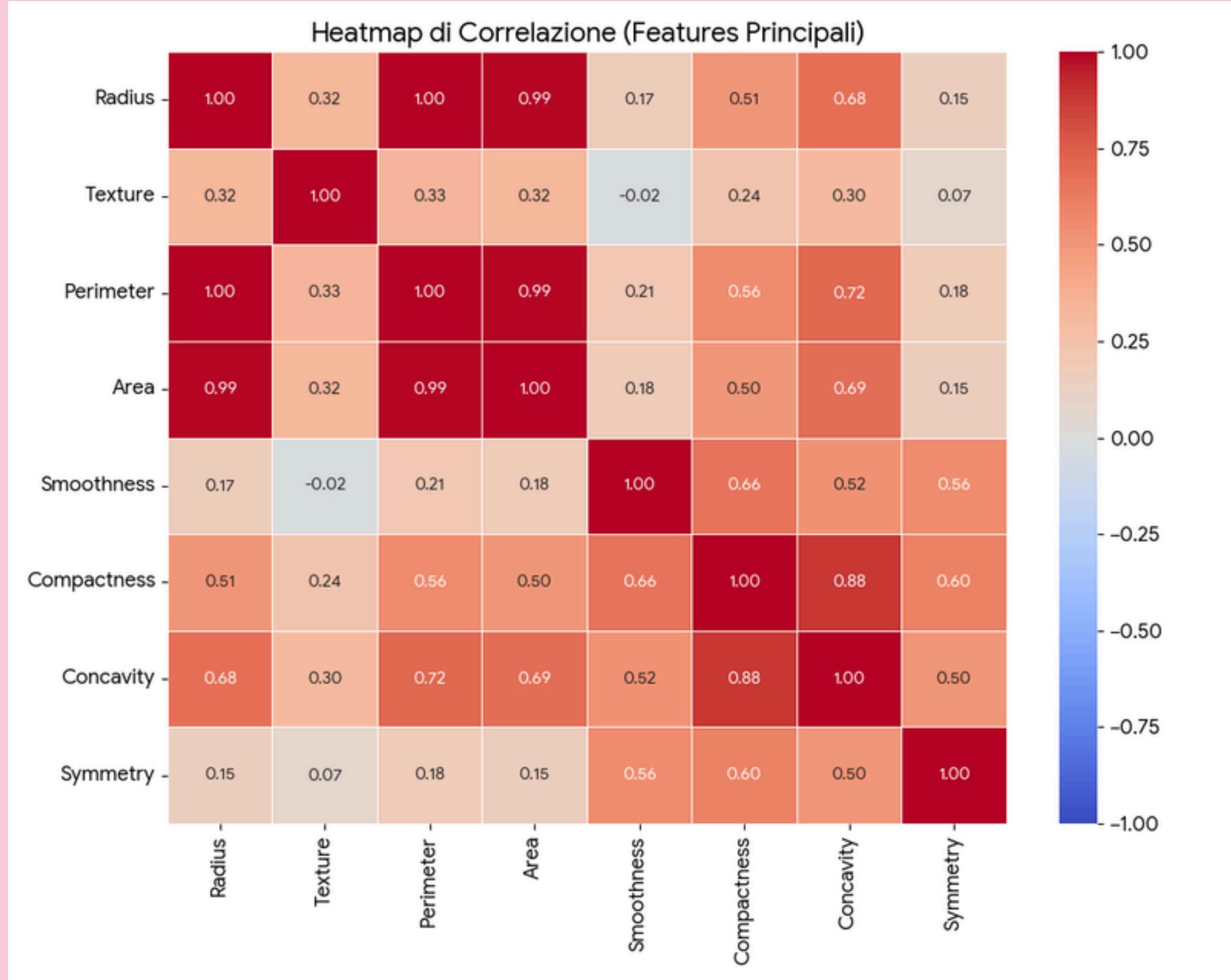
# ISTOGRAMMA DELLE FEATURES SELEZIONATE



# BOXPLOT



# HEATMAP DELLE FEATURES SELEZIONATE



# CLUSTERING CON K-MEANS

DISTANZA EUCLIDEA

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

ASSIGNMENT STEP

UPDATE STEP

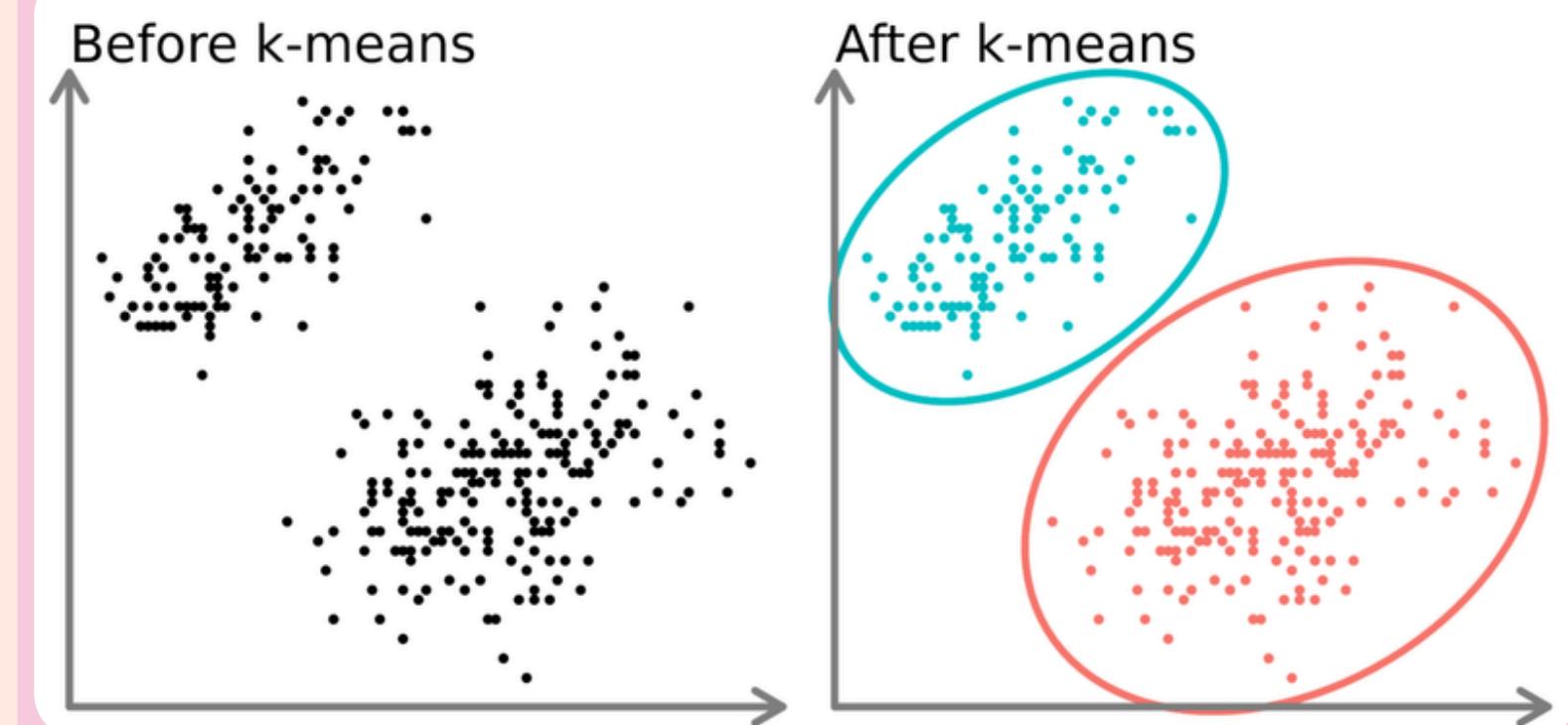
# PERCHE' USARE K-MEANS?

**1) FACILMENTE INTERPRETABILE**

**2) FUNZIONA BENISSIMO CON FEATURE  
CONTINUE**

**3) RIFLETTE I PRINCIPI PERCETTIVI DELLA  
GESTALT**

**4) COMPUTAZIONALMENTE LEGGERO**



# K-MEANS

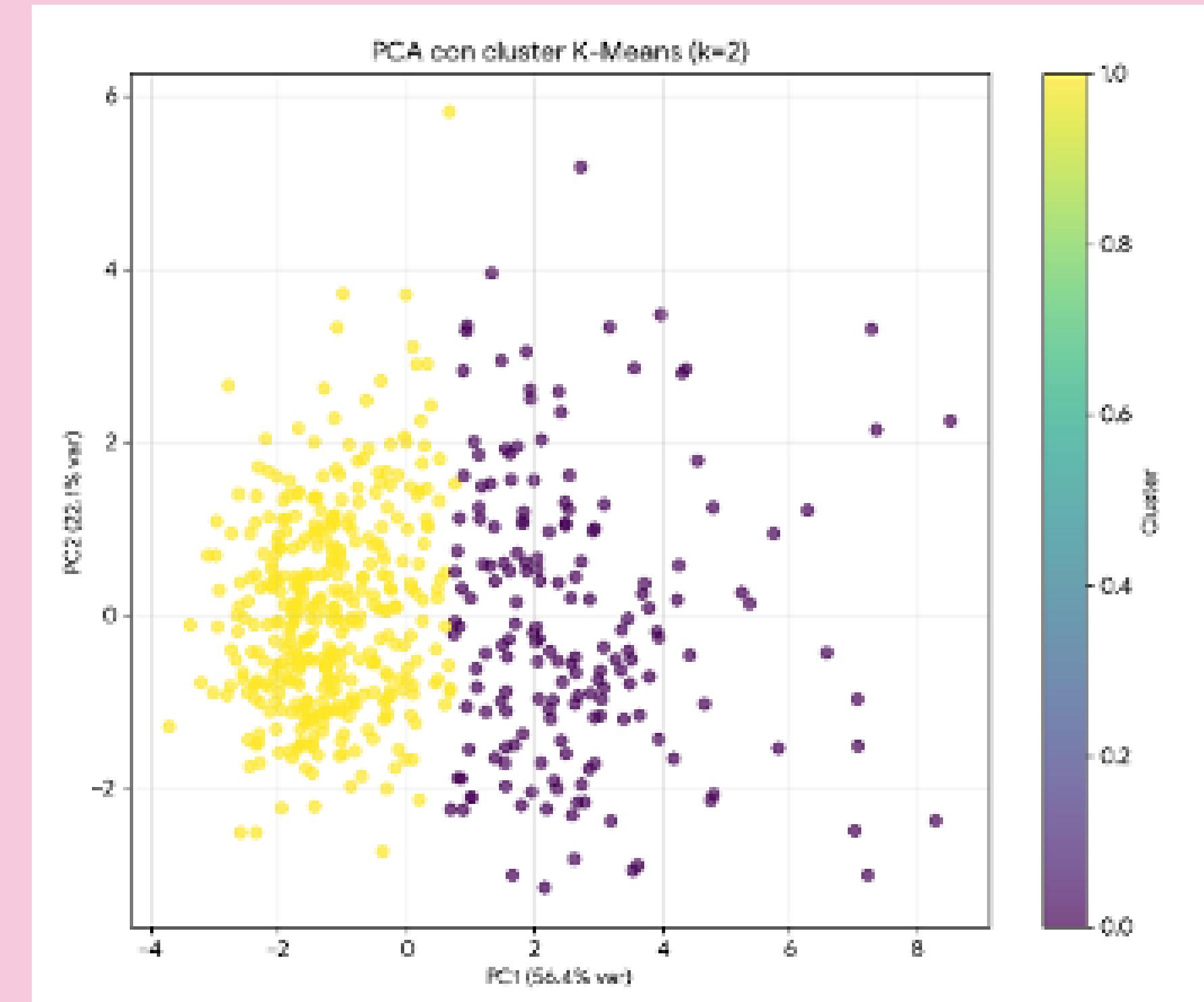
## PREPROCESSING: STANDARDIZZAZIONE DEI DATI

```
from sklearn.preprocessing import StandardScaler  
  
scaler = StandardScaler()  
X_scaled = scaler.fit_transform(X)
```

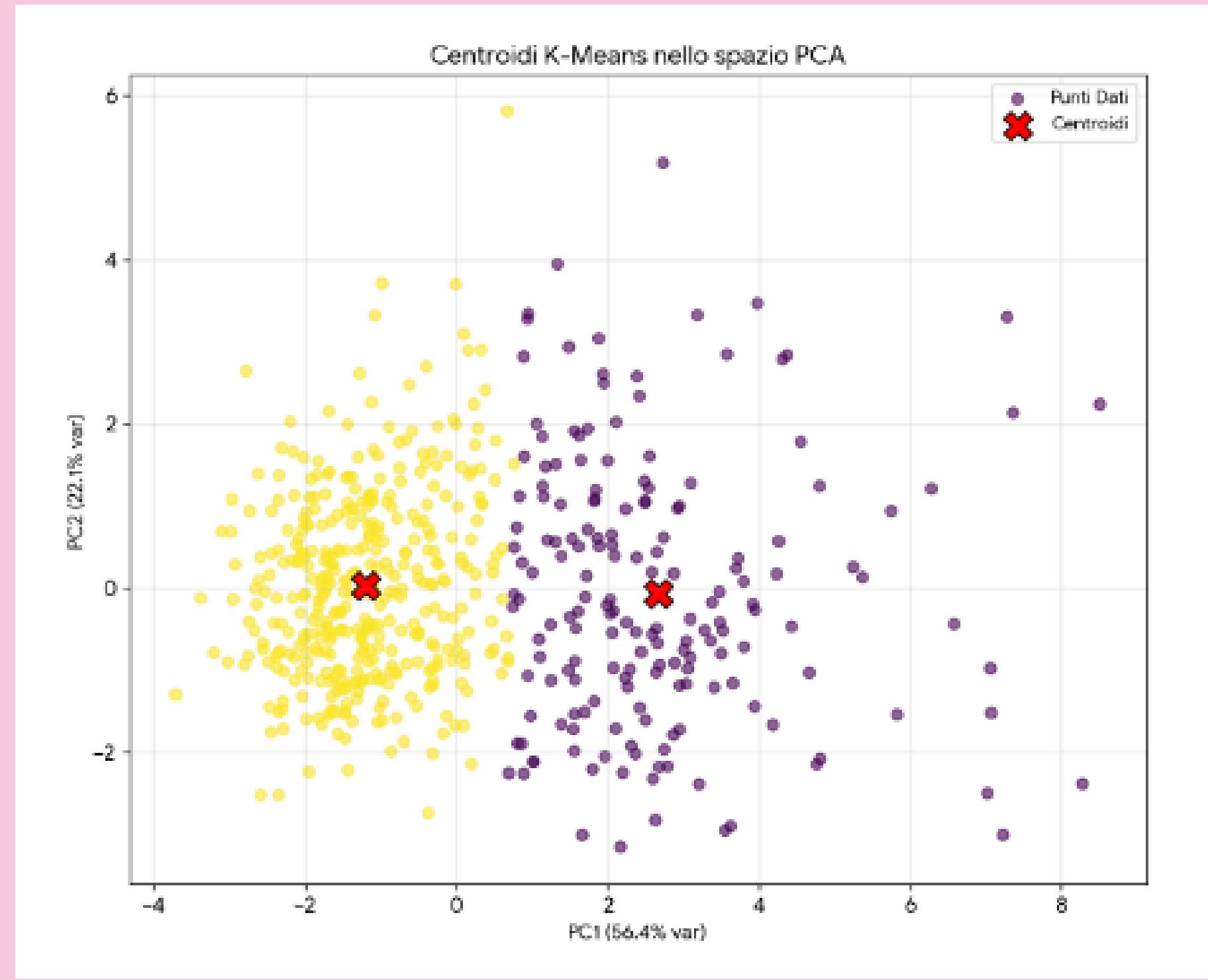
## ADDESTRAMENTO ED ASSEGNAZIONE DEI CLUSTER

```
from sklearn.cluster import KMeans  
  
kmeans = KMeans(n_clusters=2, random_state=42)  
labels = kmeans.fit_predict(X_scaled)
```

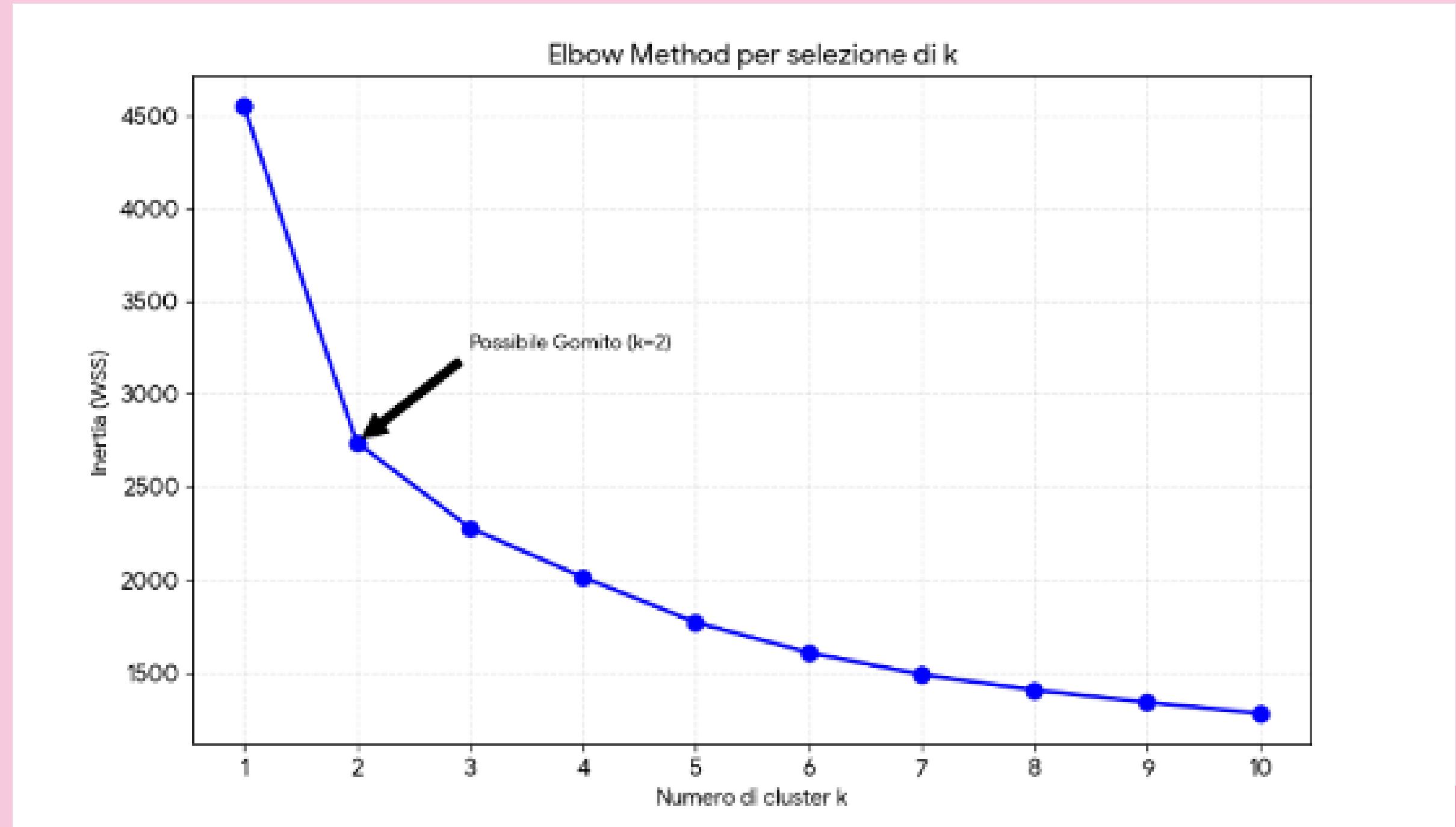
# PCA SCATTER PLOT



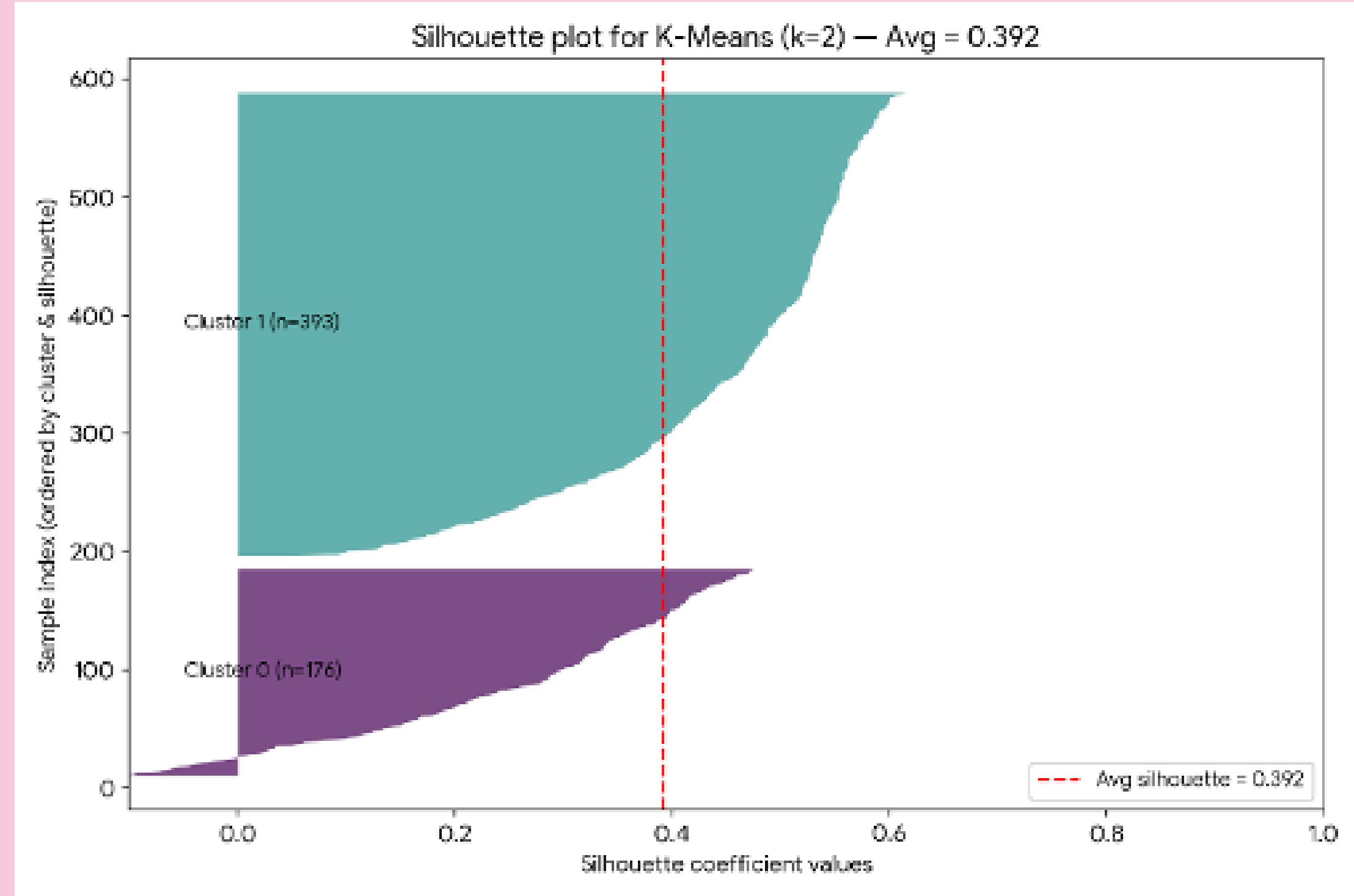
# PCA SCATTER PLOT CON CENTROIDI



# ELBOW PLOT (K VS INERTIA)



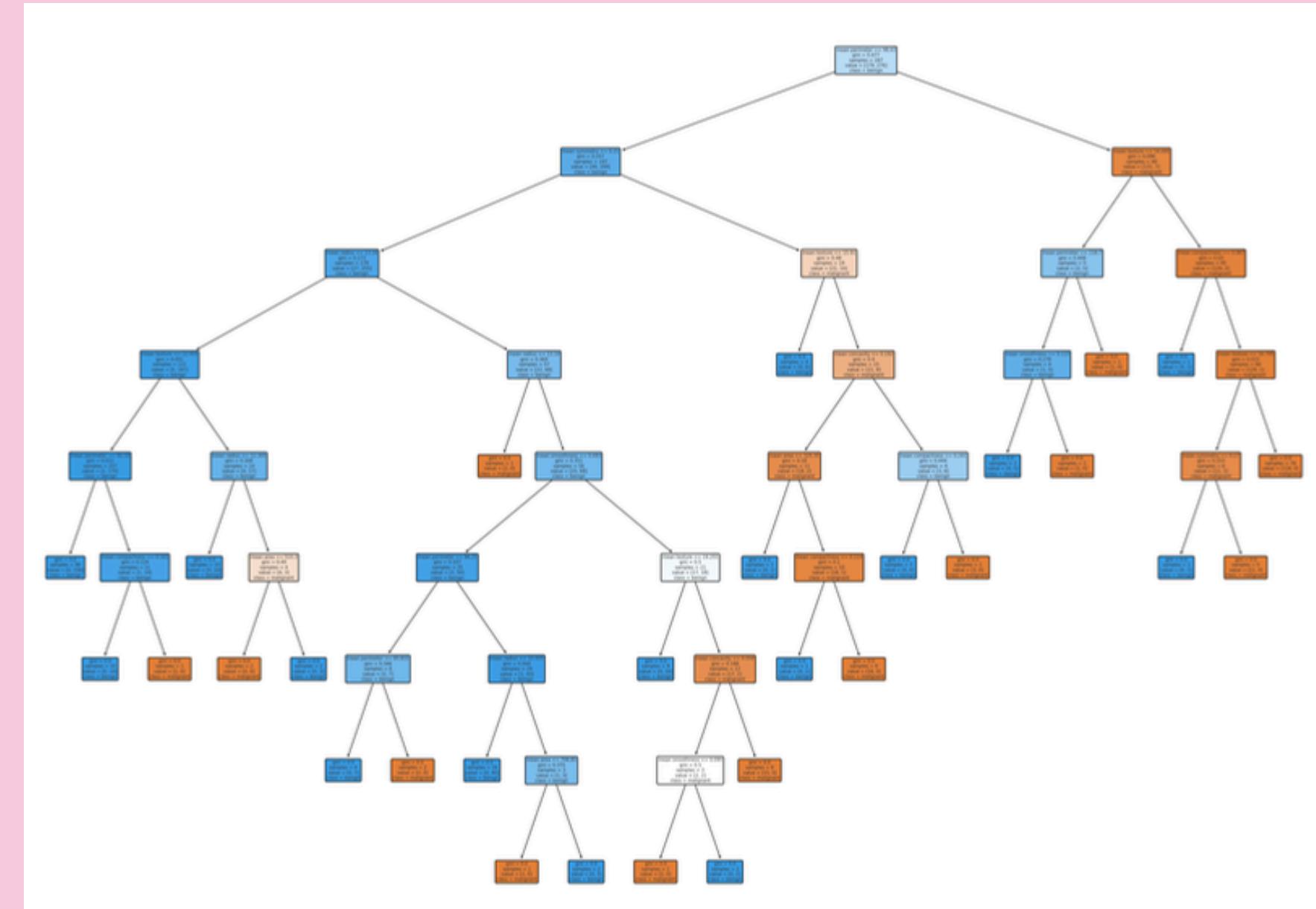
# SILHOUETTE PLOT



# CLASSIFICAZIONE

# PERCHE' USARE RANDOM FOREST?

- 1) OTTIMO PER DATASET PICCOLI  
COME IL NOSTRO E CON CON  
POCHE FEATURES**
  - 2) EVITA L'OVERFITTING**
  - 3) FACILE DA INTERPRETARE**
  - 4) OTTIMO ANCHE CON DATASET  
SBILANCIATI**



# ALGORITMO RANDOM FOREST

```
from sklearn.datasets import load_breast_cancer
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import classification_report, confusion_matrix
import pandas as pd

rf_standard = RandomForestClassifier(
    n_estimators=150,           # 150 alberi
    random_state=42             # per riproducibilità
)

# Addestramento
rf_standard.fit(X_train, y_train)

"""
X_train, X_test, y_train, y_test = train_test_split(
    X, y, test_size=0.2, stratify=y, random_state=42
)

print("\n==== CLASSIFICATION REPORT (STANDARD) ===")
print(classification_report(y_test, y_pred, target_names=["malignant", "benign"]))

print("\n==== CONFUSION MATRIX (STANDARD) ===")
print(confusion_matrix(y_test, y_pred))

print("\n==== METRICHE PRINCIPALI ===")
print("Accuracy : ", accuracy_score(y_test, y_pred))
```

**IMPORT INDISPENSABILI**

**COSTRUZIONE RANDOM FOREST**

**DIVISIONE DEL SET**

**VALUTAZIONE**



# METRICHE

## PRECISIONE

$$\text{Precision} = \frac{TP}{TP + FP}$$



## F1-SCORE

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$



$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$



## ACCURATEZZA

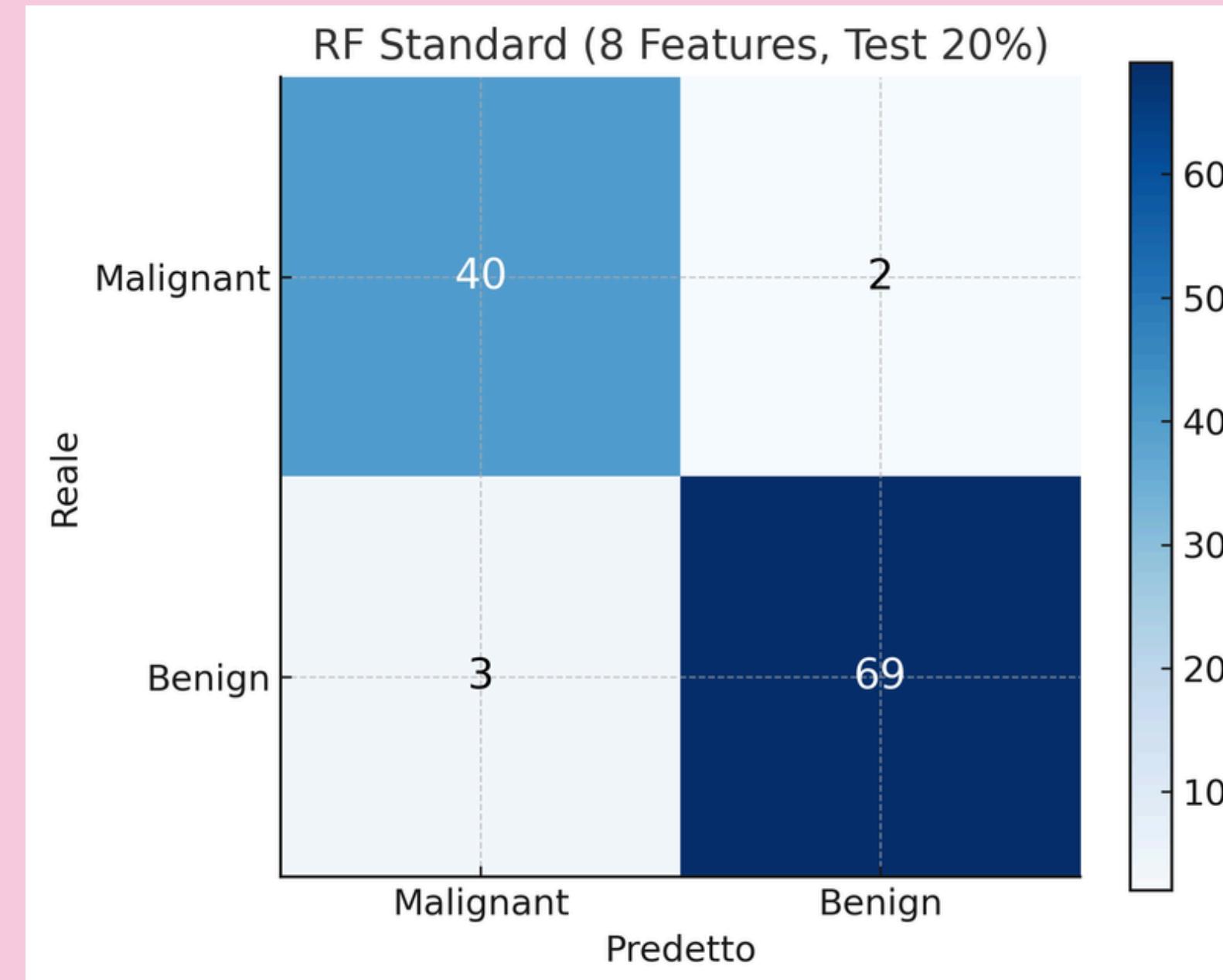


$$\text{Recall} = \frac{TP}{TP + FN}$$

## RECALL



# MATRICE DI CONFUSIONE MODELLO STANDARD



# CONCLUSIONI

- 1. ANALISI ESPLORATIVA CON HEATMAP E BOXPLOT HA RIVELATO CHE IL PROBLEMA DIAGNOSTICO È UN PROBLEMA GEOMETRICO**
- 2. CON ALGORITMO KMEANS È STATA CONFIRMATA L'ESISTENZA DI UNA STRUTTURA BINARIA NATURALE (TUTTAVIA C'È UNA "ZONA AMBIGUA")**
- 3. SIAMO RIUSCITI A DISTINGUERE TUMORI BENIGNI E MALIGNI TRAMITE RANDOM FOREST**
- 4. LA VISUALIZZAZIONE PCA CI MOSTRA COME IL KMEANS COSTRUISCA CATEGORIE MENTALI BASATE SULLA SIMILARITÀ, PROPRIO COME PROPRIO COME LA MENTE UMANA (= CATEGORIZZAZIONE)**



# FONTI

UCI MACHINE LEARNING REPOSITORY – BREAST CANCER  
WISCONSIN (DIAGNOSTIC)

MACQUEEN, J. (1967). SOME METHODS FOR CLASSIFICATION  
AND ANALYSIS OF MULTIVARIATE OBSERVATIONS.

BISHOP, C. (2006). PATTERN RECOGNITION AND MACHINE LEARNING.

ROUSSEEUW, P. J. (1987). SILHOUETTES: A GRAPHICAL AID...,  
JOURNAL OF COMPUTATIONAL AND APPLIED MATHEMATICS

SCIKIT-LEARN DOCUMENTAZIONE UFFICIALE

