■ APR Assignment 1 : Breast Cancer Classification using Support Vector Machine (SVM)

1. Introduction

Breast cancer diagnosis is a critical task in medical data analysis. Machine learning techniques, such as Support Vector Machines (SVM), are widely used for binary classification problems, including cancer detection. In this assignment, we trained and evaluated an SVM classifier on the Breast Cancer Wisconsin dataset using preprocessing, cross-validation, and performance evaluation metrics.

2. Dataset

- Source: Breast Cancer Wisconsin dataset (breastCancer.csv).
- Samples: 699 patient records.
- Features: 20 processed features (after encoding categorical variables and scaling numeric features).
- Target variable:
  Original labels: {2, 4}
  Mapped to {0, 1} $\rightarrow$ 0 = Benign, 1 = Malignant

3. Methodology

Preprocessing:
- Handled missing values (if any).
- Categorical variables encoded.
- Features standardized for SVM training.
- Data split: 80% training (559 samples) and 20% testing (140 samples).

Model:
- Algorithm: Support Vector Machine (SVM).
- Kernel: RBF (Radial Basis Function).
- Hyperparameter tuning: GridSearchCV with 5-fold cross-validation.
- Parameter space: C $\in$ {0.1,1,10}, gamma $\in$ {scale, auto}, kernel $\in$ {linear, rbf}.

4. Results

Best Parameters (via CV):
- Kernel: RBF
- C: 1
- Gamma: auto
- Best CV accuracy: 96.78%

Test Set Evaluation:
- Accuracy: 96.43%

Classification Report:
Class 0 (Benign): Precision=0.99, Recall=0.96, F1=0.97 (92 samples)
Class 1 (Malignant): Precision=0.92, Recall=0.98, F1=0.95 (48 samples)
Macro Avg F1=0.96, Weighted Avg F1=0.96

Cross-Validation Metrics:
- Mean CV Accuracy: 0.9642 ± 0.0136
- Mean CV ROC-AUC: 0.9902 ± 0.0053

5. Visualization

(Figures generated in Colab include Confusion Matrix, ROC Curve, Precision-Recall Curve)
- Confusion Matrix: Few misclassifications, high overall accuracy.
- ROC Curve: AUC ≈ 0.99, excellent separation.
- Precision-Recall Curve: High precision and recall trade-off, important in medical diagnosis.

6. Discussion

- The SVM model performed exceptionally well, achieving ~96% accuracy on the test set.
- High precision (0.99 for benign) ensures low false positives.
- High recall (0.98 for malignant) ensures most cancer cases are detected.
- ROC-AUC near 1.0 confirms strong discriminative ability.
- Limitations: Dataset size is moderate (699 samples). Real-world deployment requires larger datasets.

7. Conclusion

This study demonstrates that an SVM with RBF kernel is highly effective for breast cancer classification. With careful preprocessing and cross-validation, the model achieved high accuracy and robustness. Such techniques can support doctors in early and reliable breast cancer diagnosis.