BREAST CANCER CLASSIFICATION USING SVM

1. OVERVIEW

This Jupyter Notebook presents a case study on breast cancer classification using machine learning techniques, specifically Support Vector Machines (SVM). The goal is to predict whether a tumor is malignant (cancerous) or benign (non-cancerous) based on 30 different features derived from cell nucleus measurements.

Key Objectives:

- Load and explore the breast cancer dataset
- Perform exploratory data analysis (EDA) to understand feature distributions
- Train an SVM model to classify tumors
- Evaluate and improve model performance using scaling and hyperparameter tuning

2. DATASET DESCRIPTION

Source: The dataset is sourced from the UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Dataset (https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic).

Dataset Characteristics:

- Number of Instances: 569
- Number of Features: 30 (numeric)
- Target Variable:
    - 0 → Malignant (212 cases)
    - 1 → Benign (357 cases)

Features: The dataset contains measurements of cell nuclei, including:

- Mean Values: mean radius, mean texture, mean perimeter, etc.
- Standard Errors: radius error, texture error, etc.
- "Worst" Values: Largest (worst) values observed (worst radius, worst texture, etc.)

3. METHODOLOGY

Step 1: Problem Statement

- Goal: Classify breast tumors as malignant or benign
- Features Used: 30 numerical features derived from digitized images of fine needle aspirates (FNA)
- Class Distribution:
    - Malignant: 212 cases
    - Benign: 357 cases

Step 2: Data Importing

- Libraries Used: pandas, numpy, matplotlib, seaborn, sklearn
- Dataset Loaded Using: sklearn.datasets.load_breast_cancer()
- Converted into a Pandas DataFrame for easier manipulation

Step 3: Data Visualization

- Pairplot: Visualized relationships between key features (mean radius, mean texture, etc.) with hue as the target variable

- Countplot: Displayed the distribution of malignant vs. benign cases
- Scatterplot: Examined relationships between mean area and mean smoothness
- Heatmap: Visualized feature correlations to identify multicollinearity

Step 4: Model Training (SVM)

- Train-Test Split: 80% training, 20% testing (test_size=0.2)
- Model Used: SVC() (Support Vector Classifier with default parameters)
- Training: svc_model.fit(X_train, Y_train)

Step 5: Model Evaluation

- Confusion Matrix: Evaluated model performance on test data
- Classification Report: Provided precision, recall, and F1-score

Step 6: Model Improvement

- A. Feature Scaling: Applied min-max scaling to normalize data
- B. Hyperparameter Tuning (GridSearchCV): Searched for optimal SVM parameters

4. RESULTS & CONCLUSION

Key Findings:

- The SVM model achieved 97% accuracy in classifying tumors
- Feature scaling and hyperparameter tuning significantly improved performance
- The model can be useful for early breast cancer detection, especially in resource-limited settings

Future Improvements:

- Computer Vision Integration: Use deep learning to classify tumors directly from tissue images
- Larger Dataset: Train on more diverse data to improve generalizability
- Explainability: Use SHAP/LIME to interpret model decisions

5. CODE STRUCTURE

Step 1 - Problem Statement Description: Initial problem definition Key Functions Used: None

Step 2 - Data Importing Description: Loading and preparing data Key Functions Used: load_breast_cancer(), pd.DataFrame()

Step 3 - Data Visualization Description: Exploratory data analysis Key Functions Used: sns.pairplot(), sns.heatmap()

Step 4 - Model Training Description: Training the SVM model Key Functions Used: train_test_split(), SVC()

Step 5 - Model Evaluation Description: Assessing model performance Key Functions Used: confusion_matrix(), classification_report()

Step 6 - Model Improvement Description: Enhancing model accuracy Key Functions Used: GridSearchCV(), Min-Max Scaling

6. HOW TO RUN

- Ensure dependencies are installed (pandas, numpy, matplotlib, seaborn, scikit-learn)
- Execute cells sequentially
- Modify hyperparameters (C, gamma) for further tuning

7. REFERENCES

- UCI Breast Cancer Dataset: https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic
- Scikit-learn Documentation: SVM: https://scikit-learn.org/stable/modules/svm.html