Breast Cancer Classification Report

AAYUSH BHARUKA

---

1. Dataset Description and Preprocessing Steps

The dataset used in this study consists of 4,024 entries with 16 features, including demographic details, tumor characteristics, and survival status. The target variable is "Status," which indicates the classification outcome.

Preprocessing Steps:

- **Handling Categorical Data**: Label encoding was applied to transform categorical features into numerical values.
- **Missing Values Handling**: Numerical columns with missing values were imputed using the mean.
- **Feature Scaling**: Standardization was performed using StandardScaler to normalize numerical features.
- **Train-Test Split**: The dataset was split into 70% training and 30% testing subsets.

---

2. Exploratory Data Analysis (EDA)

EDA was performed to understand the dataset's distribution and key characteristics.

Key Findings:

- **Count of Patients by Age**: The number of patients peaks between **45-50 years**.
- **Tumor Size Distribution**: Tumor sizes range between **10 to 50**, with some outliers, and the highest count occurs at size **20**.
- **Survival Months Range**: Survival months mostly range from **50 to 100 months**.
- **Feature Correlation**: A heatmap revealed strong correlations between certain tumor characteristics and survival status.
- **Feature Distributions**: Histograms and density plots showed a wide range of values for age, tumor size, and other features.
- **Pairplot Analysis**: Some features demonstrated clear separability between classes.

Visualizations:

- **Count plot of "Status"** to check the class balance.
- **Heatmap of feature correlations** to identify significant relationships.
- **Pairplots and histograms** for visualizing feature distributions.

---

3. Model Performance & Results

Four classification models were implemented: Naive Bayes, K-Nearest Neighbors (KNN), Decision Tree, and Logistic Regression. Their performance was evaluated using a confusion matrix and key metrics.

Naive Bayes:

- **Confusion Matrix:**
    - [[888 135] [100 85]]
- **Accuracy:** 0.805464
- **Precision:** 0.820310
- **Recall:** 0.805464
- **F1-Score:** 0.812177

KNN (k=3):

- **Confusion Matrix:**
    - [[982 41] [114 71]]
- **Accuracy:** 0.871689
- **Precision:** 0.855853
- **Recall:** 0.871689
- **F1-Score:** 0.858130

KNN (k=5):

- **Confusion Matrix:**
    - [[993 30] [123 62]]
- **Accuracy:** 0.873344
- **Precision:** 0.856725
- **Recall:** 0.873344
- **F1-Score:** 0.854836

KNN (k=7):

- **Confusion Matrix:**
    - [[1002 21] [134 51]]
- **Accuracy:** 0.871689
- **Precision:** 0.855439
- **Recall:** 0.871689
- **F1-Score:** 0.846838

Decision Tree:

- **Confusion Matrix:**
    - [[992 31] [98 87]]
- **Accuracy:** 0.893212
- **Precision:** 0.883628
- **Recall:** 0.893212

- **F1-Score:** 0.883098

Logistic Regression:

- **Confusion Matrix:**
    - [[993 30] [99 86]]
- **Accuracy:** 0.893212
- **Precision:** 0.883618
- **Recall:** 0.893212
- **F1-Score:** 0.882714

---

4. Model Comparison & Conclusions

Model	Accuracy	Precision	Recall	F1-Score
Naive Bayes	0.805464	0.820310	0.805464	0.812177
KNN (k=3)	0.871689	0.855853	0.871689	0.858130
KNN (k=5)	0.873344	0.856725	0.873344	0.854836
KNN (k=7)	0.871689	0.855439	0.871689	0.846838
Decision Tree	0.893212	0.883628	0.893212	0.883098
Logistic Regression	0.893212	0.883618	0.893212	0.882714

- **Naive Bayes**: Performs well with smaller datasets and assumptions of feature independence, but may not capture complex relationships.
- **KNN**: Performance varies with different values of k, with k=5 showing the best balance between bias and variance.
- **Decision Tree**: Using post-pruning (max_depth=5) helps reduce overfitting while maintaining good performance.
- **Logistic Regression**: Provides interpretability and decent performance, though it may struggle with non-linearly separable data.

Final Conclusion:

- The best-performing models based on accuracy and F1-score were **Decision Tree and Logistic Regression**, both achieving an accuracy of **0.893212**. However, the **F1-score is slightly better in the case of the Decision Tree**.
- Further tuning of hyperparameters and feature engineering could enhance performance.
- A combination of models (e.g., ensemble methods) might improve classification results.

This study provides a foundational approach to breast cancer classification using machine learning. Future work could involve deep learning techniques and larger datasets for improved accuracy.