Name : Ahmad Bello Rilwan

Student Number : 23080952

Report on Tumor Classification Using the Breast Cancer Wisconsin (Original) Dataset

Breast cancer is one of the most common deadly diseases in women globally. Since early detection significantly improves survival chances, this study makes use of valuable tools offered by machine learning for diagnosis. It therefore presents an empirical assessment of three classification models: **Logistic Regression, Decision Tree, and Random Forest** in predicting breast cancer based on clinical features as either benign or malignant. This is based on the Breast Cancer Wisconsin (Original) dataset obtained from the UCI repository that contains diagnostic attributes helpful in determining malignancy such as clump thickness and uniformity of cell shape among others. To find out the best classifier, data preprocessing followed by application of the models and evaluation was carried out.

Data Preparation

The dataset contained 699 records and 11 attributes or fields of data; with one being the target class. Missing values in **Bare_Nuclei** were addressed by adding a code that converted every column to numeric and coerced errors to NaN. The rows containing missing values were eliminated. From the boxplots, outliers were seen in features like **Mitoses and Bare_Nuclei**; these features will be kept because their biological relevance makes them plausible. Logistic Regression requires feature standardization, so standardization was applied as per the univariate analysis which showed that most features were right-skewed. The original target labels (2 = benign, 4 = malignant) were recoded to 0 and 1 respectively for model compatibility and it showed a slight imbalance (65% benign, 35% malignant). Not all metrics were emphasized during evaluation, though recall and F1-score are the metrics emphasized.

Data Mining Analysis

A heatmap showed strong correlations between features like **Uniformity_of_Cell_Size, Uniformity_of_Cell_Shape, and Bare_Nuclei.** Weaker correlated features were kept to see their contribution in the model. In Logistic Regression Residual analysis, mostly false positives were seen and only one false negative was there.

Logistic Regression is binary classification and it is interpretable. Decision Tree works well when there are non-linear patterns and the data is not scaled; Random Forest is an ensemble model which gives healthy results and less overfitting. All models were subjected to accuracy, recall, precision, F1-score, confusion matrix, and ROC-AUC evaluations.

Model Development

Data was divided in the ratio 80/20 for training and testing purposes with stratified sampling. Logistic Regression utilized standardized features and was tuned using GridSearchCV (best: C=0.01, penalty=L2). The model attained an overall accuracy of 96.35%, with a recall of 0.98 and AUC of 0.99. The main predictors were **Bare_Nuclei and Uniformity_of_Cell_Size.** Decision Tree model showed an accuracy of 97.08% and executed perfect recall in its base form. Tuning diminished performance marginally, implying that default possibly captured patterns more effectively. Random Forest matched the Decision Tree's performance having 97.08% accuracy and an equivalent AUC of 0.98 when assessed against it. Able to confirm its dependability through stable metrics across folds, optimization applied.

Model Comparison

All models performed well. Logistic Regression showed strong results and was the only one to produce a false negative. Default Decision Tree had perfect recall; that recall decreased slightly when optimized. Random Forest matched the accuracy of the Decision Tree but with added stability. Below is the summary table: Random Forest emerged as the most reliable, balancing high accuracy with perfect recall and robustness of the model. Logistic Regression can be considered a strong baseline, though it is less robust.

Model	Accuracy (%)	Recall (Malignant)	AUC
Logistic Regression	96.35	0.98	0.99
Decision Tree	97.08	1.00	0.98
Random Forest	97.08	1.00	0.98

ReferenceBreast Cancer Wisconsin (Original) Dataset, UCI Machine Learning Repository:

https://archive.ics.uci.edu/dataset/15/breast+cancer+wisconsin+original