1. Overview

The problem our code is trying to solve is classifying lumps in a patient’s breast as benign or cancerous. Breast cancer is the most common cancer worldwide and one of the most important health issues of the modern age. If our model can accurately identify benign and malignant cancer cells, we can allow patients to receive the treatment they need in order to fight breast cancer without wasting precious time. The success of this model can also cause cost reduction for Hospitals by optimizing treatment quickly, saving both costs and lives. This code specifically walks through data preparation, model training, and evaluation steps using machine learning techniques in Python.

1. Methodology

* Data Import: Training and test datasets are loaded from CSV files. These datasets contain measurements and features that are possibly indicative of cancerous conditions.
* Data Cleaning: The 'ID' column, which is not predictive, is removed to focus the analysis on relevant features. We then inspect and handle any missing or improperly formatted data, specifically in the 'BARE' column by imputing missing values (originally blanks) using KNN and 3 neighbor values.
* Data Preprocessing: We then standardized the data, ensuring that each feature contributes proportionately to the predictive models. We also checked for multicollinearity and identified that cell shape and size variables showed high multicollinearity.
* Model Training:
  + Simple Logistic Regression: A simple Logistic Regression model was trained for comparison purposes.
  + Ridge Classifier with Cross-Validation: A Ridge Classifier, which applies L2 regularization to a linear classification model, is trained with Leave-One-Out cross-validation to determine the best alpha value.
  + Logistic Regression with Lasso Regularization: Another 10-fold cross-validated Logistic Regression model with Lasso (L1 penalty) regularization is trained.
  + Best SVM model: 10-fold cross-validated with best parameters found via grid search.
* Model Evaluation & Selection: Each model's performance is assessed using recall, accuracy and AUC metrics. Recall is particularly important to help identify which model maximizes correct identification of malignant lumps. Also, we identified a subset model with 5 features based on bias-variance tradeoff and variable significance. Comparison between Simple Logistic model and SVM revealed that SVM had the highest recall on the test set while they had similar accuracy. Simple Logistic is selected for better interpretation, but in scenarios where interpretability is not important, SVM can be selected.
* Conclusion: We selected the Simple Logistic model as the best-performing model due to the relatively high recall score and best accuracy and AUC among four models. When the final choice came down to Simple Logistic or SVM, interpretability was taken into account where SVM lacked considerably.