Brian J. Davis

Dr. Tanvi Banerjee

CS-7830

18 February 2025

Assignment 1 Report

A single-feature model using only mean_texture served as an initial baseline for this assignment, though its negative R^2 and Adjusted R^2 values on both the training and testing subsets indicated a limited capacity to explain variation in tumor_size. This outcome suggested that mean_texture, by itself, offers only minimal predictive power for the task at hand.

Extending the model to include lymph_node_status proved more informative. The resulting two-feature approach demonstrated lower MSE for both the training and testing subsets, and the training subset's R^2 became positive, reflecting some explanatory value in the added variable. However, the model still yielded a negative R^2 on the testing subset, which implies weaker generalization and insufficient predictive strength when faced with new data.

A more systematic feature selection approach, using forward stepwise regression, prioritized mean_perimeter and lymph_node_status based on Bayesian Information Criterion (BIC). This yielded further improvements over both the single and two-feature models, as evidenced by a reduced test MSE and a less negative R^2. By comparison, backward stepwise regression (initially incorporating all ten candidate predictors) encountered numerical instability, generating NaN values for both MSE and R^2. Consequently, it was not possible to draw meaningful comparisons against other approaches in that branch of the analysis.

Subsequent introduction of L2 regularization and feature scaling offered an additional layer of refinement. Both methods delivered measurable enhancements over the unregularized forward stepwise baseline, further reducing the MSE and shift R^2 values in a more favorable direction. Regularization alone, however, outperformed feature scaling in isolation, and the combination of the two did not exceed the performance of regularization by itself, though it remained superior to scaling alone. Altogether, a forward stepwise model supplemented with L2 regularization proved to be the strongest configuration under these conditions, as it minimized MSE while simultaneously yielding the highest (or least negative) R^2 values.