

Model Evaluation Report

I. Model Evaluation Scenario: Cross-Validation

Cross-validation was selected as the model evaluation scenario for several reasons:

- A. Reliable Estimation of Model Performance: By averaging the model performance across various folds, cross-validation provides a more reliable estimation of how the model will perform on unseen data
- B. Robustness: Models can be overly fitted to a particular subset of data – cross-validation ensures that the model is tested across different subsets, reducing the risk of biased evaluations. For the data set, there were no missing values

II. Metrics Reported

- A. Accuracy: percentage of test set tuples that are correctly classified
 - Accuracy: 65%
- B. Precision (Positive Predictive Value): exactness – what % of tuples that the classifier labeled as positive are actually positive
 - Precision: 56%
- C. Recall (Sensitivity or True Positive Rate): completeness – what % of positive tuples did the classifier label as positive?
 - Recall/Sensitivity: 41%
- D. F-Measure (Harmonic Mean of Precision and Recall): harmonic mean of precision and recall
 - F-Measure: 47%
- E. Specificity (True Negative Rate): True Negative recognition rate
 - Specificity: 34%

III. Analysis

A. Average Accuracy: 0.65 (65%)

- a. The model correctly predicts the target 65% of the time. For a task like labeling heart disease, this is likely to be considered low. The model would not be a good tool for identifying if a patient has heart disease or not.

B. Average Precision: 0.56 (56%)

- a. The model is only correct about positive cases 56% of the time, meaning the other 44% is false positives. This is a low precision resulting in a person receiving a (problematic) false positive a significant proportion of the time.

C. Average Recall: 0.41 (41%)

- a. A recall of 41% is critically low for a heart disease classifier, as it implies that the model fails to detect approximately 59% of the true heart disease cases.

D. Average F1-Score: 0.47 (47%)

- a. An F1-score of 47% suggests that the model is neither efficient at correctly identifying patients with heart disease nor at avoiding misclassification of healthy individuals as having heart disease.

E. Average Specificity: 0.34 (34%)

- a. The low specificity of 34% means the model incorrectly labels many healthy individuals as having heart disease. This can lead to unnecessary anxiety, further testing, increased healthcare costs, and potential overtreatment.