ANA680 National University

Kirsten Mattingly

2/8/2025

GitHub Link: <https://github.com/kmattingly8/ANA680.git>

Assignment 2: Building Classification Models

|  |  |  |
| --- | --- | --- |
| A blue squares with white text  AI-generated content may be incorrect.  Accuracy = 96.57% | A blue squares with white text  AI-generated content may be incorrect.  Accuracy = 97.14% | A blue squares with white text  AI-generated content may be incorrect.  Accuracy = 96.00% |
| A blue squares with white text  AI-generated content may be incorrect.  Accuracy = 97.71% | A blue squares with white text  AI-generated content may be incorrect.  Accuracy = 97.14% | A blue squares with white text  AI-generated content may be incorrect.  Accuracy = 94.28% |
| A blue squares with white text  AI-generated content may be incorrect.  Accuracy = 96.57% | A blue squares with white text  AI-generated content may be incorrect.  Accuracy = 96.57% |  |

**Summary**: In the case of predicting breast cancer, both precision and recall are critical evaluation statistics for any model. Regarding precision, it is important that the model minimizes instances of false positives because telling someone they have cancer when that is not actually the case could be quite traumatic. That being said, the cost of false negatives, i.e. telling someone they don’t have cancer when they actually do, is even greater, so a model with high recall is even more crucial.

After building prediction models using 8 different classification techniques, we see in the charts above that the RBF SVM model has the highest overall accuracy (97.71%). This model only misses 2 instances of actual malignancy (i.e. 2 false negatives), which means it has a high level of recall (tied with the Naïve Bayes and XGBoost models). It also only incorrectly predicts malignancy 2 times (i.e. 2 false positives), which means it has a high level of precision (tied with the Logistic Regression, KNN, and Linear SVM models) as well. Because the RBF SVM model maximizes recall while also maximizing precision, it is definitely the strongest overall model among all the models built for this assignment.

If there had been a model that further maximized recall (i.e. 0-1 false negatives) but perhaps had lower precision, we would have had to weigh the cost of a false negative to that of false positives in order to conclude which model is best. Again, I would argue that the cost of a false negative is much greater in the long run than a false positive in this case, so would likely choose the model with the highest recall, even if it meant having slightly lower precision. Ideally, as is the case above, we’d want to maximize both statistics.