Applications of Machine Learning in the Diagnosis of Breast Cancer.

Bobby Smith
Department of Electrical and Computer Engineering
University of British Columbia
Vancouver, BC V6T 1Z4
bobby.smith1994@gmail.com

Abstract

Doctors use cancer imaging tests to answer a range of questions, like: Is it cancer or a harmless lump? If it is cancer, how fast is it growing? How far has it spread? Is it growing back after treatment? Studies suggest that AI has the potential to improve the speed, accuracy, and reliability with which doctors answer those questions (1). Doctors believe AI can automate assessments and tasks that humans currently can do but take a lot of time. After the AI gives a result, a radiologist simply needs to review what the AI has done—did it make the correct assessment?. That automation is expected to save time and costs, but that still needs to be proven. The topic of this report is looking at the application of a variety of classification algorithms to diagnose breast cancer via data obtained from a pathology slide. Several machine learning classification algorithms are analyzed in google colab with an additional focus on theory for logistic regression. Confusion matrix data is generated to evaluate the models and supports the idea of the decision tree classifier as a diagnostic model due to its low false negativity rate, preventing many patients from slipping through the cracks.

1 Introduction

Cancer is a major public health issue in the modern world. One of the most common types of cancer that kill women is breast cancer. Throughout the last few decades, investigators have demonstrated the ability to automate the initial level identification and classification of breast cancer tumors. This could allow for more efficient treatments and potentially save lives. All of this can be done by applying machine learning classification algorithms to histopathological data gathered through analysis of slide images from tumors. The classification algorithms utilized and compared in this report include logistic regression, the K neighbours classifier (KNC), Gaussian Naïve Bayes (GNB), Decision tree classifier (DTC), and support vector classifier (SVC). The methods employed will utilize supervised learning since the dataset being used has each sample labeled with whether it was actually cancerous or not. The algorithms will be employed on a labeled dataset consisting of 569 samples containing features computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. The features describe characteristics of the cell nuclei present in the image.

2 Methods

Classification algorithms are used for this problem because we're simply interested in classifying the samples as one of two different categories, malignant (cancerous) or benign (non-cancerous). Some pre-processing is necessary before the data can be fed into a machine learning algorithm. The csv datasheet is read in as a pandas data-frame, any entries containing Nan (nothing) must be removed. This was done on the very last column. The labels then had to be encoded into numbers, malignant is represented as a 1 and benign as a 0 in these models. The data-frame was then split into two datasets, a feature dataset consisting of independent variable data such as the radius of the tumors, and a target dataset of the dependent variable which was the diagnosis of the tumor (malignant or benign). These datasets were then split as well into training and testing sets each. After this the feature datasets had to be scaled to the same level of magnitude. Transform instead of fit_transform is used for the testing feature set to ensure that no extra information is given to the model to match or assumptions.

After pre-processing is finished, the data can be fed into different machine learning algorithms. The simplest of the algorithms employed is logistic regression, this is a form of regression that's used when the dependent variable is dichotomous (e.g. present or not present) and the independent variables are of any type (discrete or continuous). The independent variables are related to the dependent via the equation below, where p is the probability of cancer being present, X is the features matrix containing attributes of the tumor cell nuclei (radius_mean, symmetry_mean, etc.) and y is the label/dependent variable (1 for cancerous, 0 for non-cancerous).

$$P(\hat{y} = 1|X) = \text{logit}^{-1}(X^T\Theta) = \frac{1}{1 + e^{-(X^T\Theta)}}$$

Another algorithm used is the support vector machine. This classification algorithm is well suited to this problem as it excels with classification of complex but small or medium sized datasets, this one having less than 1000 samples definitely fits into this category by machine learning standards. It also excels in higher dimensional spaces, a good thing considering each sample vector would have 32 dimensions. Another advantage being that the decision boundary is explicitly constructed to minimize generalization error.

The K neighbors classifier is another algorithm used. Big advantages to this are it does not rely on any assumptions of linearity about the underlying data, and it can handle multi-class cases. This model can also be used for regression as well besides classification which could aid in investigating further trends with the cancer data.

Gaussian Naïve Bayes is also used as a classifier in this report. This algorithm may not be the best as it relies on strong assumptions that the features are independent, which is unlikely in this case given many of them are probably correlated. The advantages of it are efficiency and ease of implementation. The Naïve Bayes formula is .

The last algorithm being used is the decision tree classifier, the advantages of this method are it does not require scaling and normalization which could be of benefit to efficiency when dealing with much larger datasets, missing values also does not affect the process of building the tree considerably.

3 Results

Five machine learning models were trained on the data and their accuracy scores with the training data and test data were determined.

Table 1: Accuracy of several models when applied to training data.

Model	**Training Accuracy**	**Testing Accuracy**
Logistic regression	0.991	0.944
K nearest neighbor	0.977	0.958
Support vector machine	0.988	0.965
Gaussian Naïve Bayes	0.951	0.923
Decision Tree Classifier	1.0	0.944

Precision, recall, F1 score, and other performance metrics were computed for each model.

	precision	recall	f1-score	support
0	0.96	0.96	0.96	90
1	0.92	0.92	0.92	53
accuracy			0.94	143
macro avg	0.94	0.94	0.94	143
weighted avg	0.94	0.94	0.94	143

Figure 1: Logistic regression classification report on test data.

	precision	recall	f1-score	support
0	0.95	0.99	0.97	90
1	0.98	0.91	0.94	53
accuracy			0.96	143
macro avg	0.96	0.95	0.95	143
weighted avg	0.96	0.96	0.96	143

Figure 2: K nearest neighbor classification report on test data.

	precision	recall	f1-score	support
0	0.98	0.97	0.97	90
1	0.94	0.96	0.95	53
accuracy			0.97	143
macro avg	0.96	0.96	0.96	143
weighted avg	0.97	0.97	0.97	143

Figure 3: Support vector machine classification report on test data.

	precision	recall	f1-score	support
0	0.93	0.94	0.94	90
1	0.90	0.89	0.90	53
accuracy			0.92	143
macro avg	0.92	0.92	0.92	143
weighted avg	0.92	0.92	0.92	143

Figure 4: Gaussian Naïve Bayes classification report on test data.

Decision Tree Classifier testing performance metrics

	precision	recall	f1-score	support
0	0.99	0.93	0.96	90
1	0.90	0.98	0.94	53
accuracy			0.95	143
macro avg	0.94	0.96	0.95	143
weighted avg	0.95	0.95	0.95	143

Figure 5: Decision tree classification report on test data.

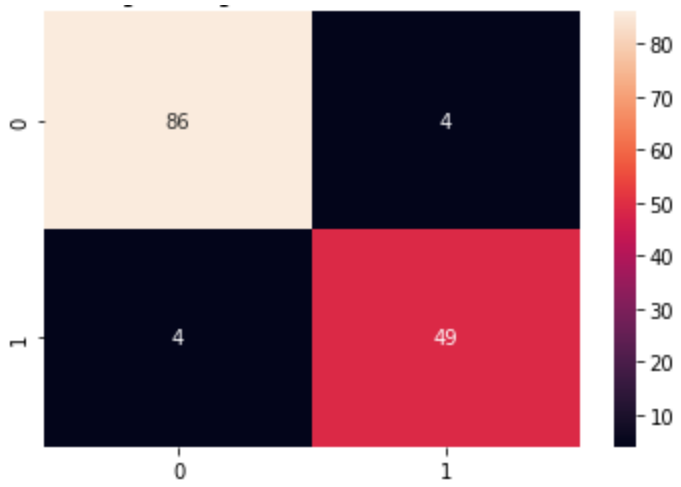

Figure 6: Logistic regression classification confusion matrix heatmap

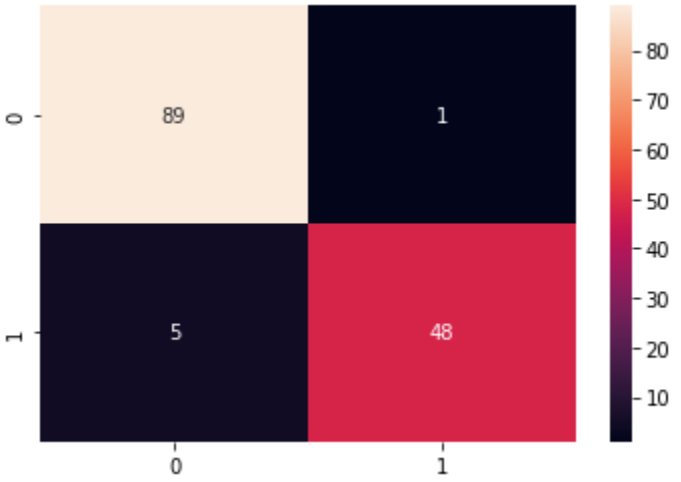

Figure 7: K nearest neighbor classification confusion matrix heatmap

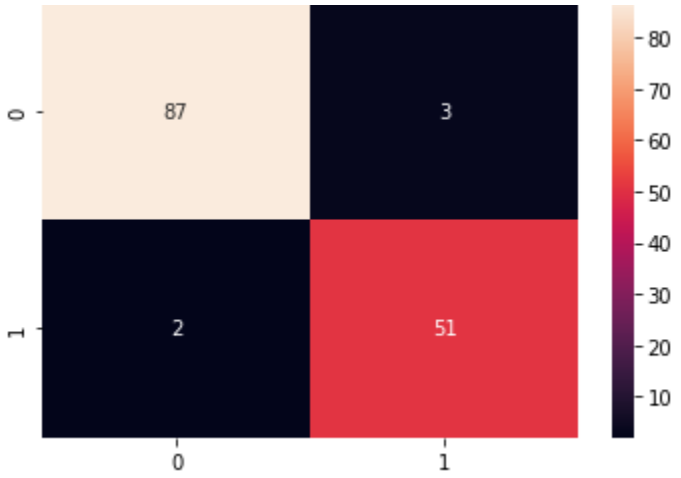

Figure 8: Support vector machine classification confusion matrix heatmap

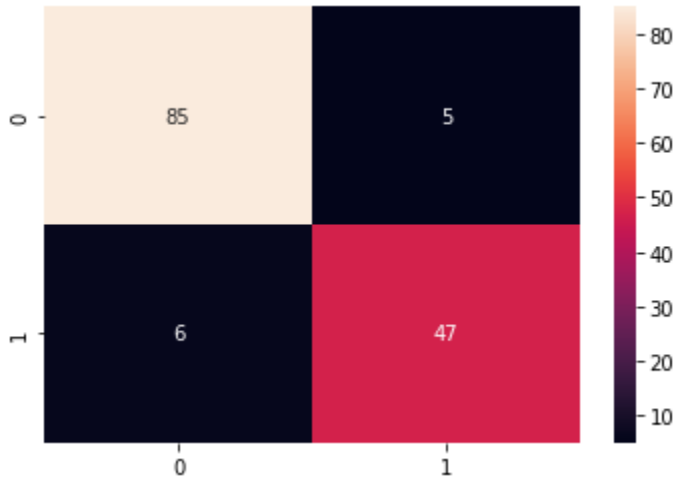

Figure 9: Gaussian Naïve Bayes classification confusion matrix heatmap

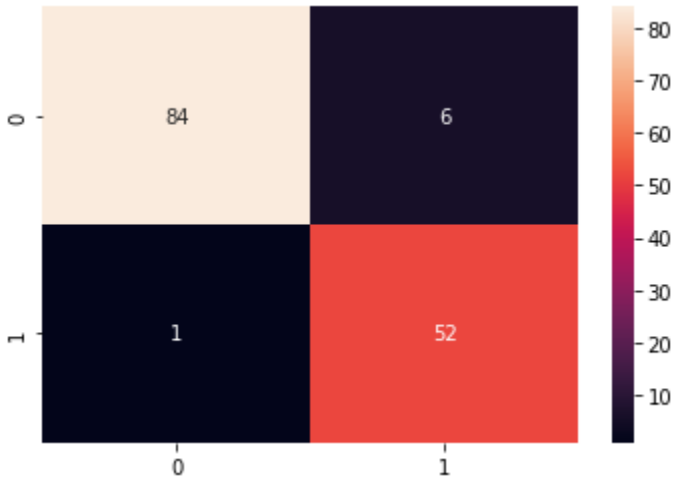

Figure 10: Decision tree classification confusion matrix heatmap

4 Discussion

Overall the decision tree classifier has performed the most accurately on the training data with an accuracy of 100%, the support vector machine was shown to have the highest accuracy on the testing data-set with an accuracy of 96.5%, this is likely owed to the high dimensionality of the feature space.

The highest true positive from the confusion matrix and highest recall of label 1 was on the decision tree classifier indicating that it was the most sensitive model, meaning it had a high probability of detection and was unlikely to register a false negative. Since the consequences of missing breast cancer are significantly higher than miss-labeling a healthy person as having breast cancer, this indicates this model is the best.

The K nearest neighbor classification had the highest specificity as seen from the upper left of the confusion matrix and the precision level on label 1. This indicates this algorithm did the best job of correctly diagnosing non-cancerous tumors, avoiding false positive is less important in this problem though so the decision tree is still the better option, because missing a cancer diagnosis could cost somebody their life.

5 Conclusion

In this paper, a shallow comparison of several classification methods is given for breast cancer, with a focus on theory with regards to logistic regression. The performance metrics from several models was analyzed to glimpse at what could be the most advantageous model to use for this type of problem. With a low false negative rate being the most desirable trait with time sensitive cases like breast cancer where an early catch could allow a patient to avoid metastasis and likely death, this report gives evidence for the decision tree classifier as the best model to use.

References

1) Breast cancer prediction using machine learning approaches. (2019). *JOURNAL OF MECHANICS OF CONTINUA AND MATHEMATICAL SCIENCES*, *14*(6). https://doi.org/10.26782/jmcms.2019.12.00012