Classification Methods for Predicting Tumor Malignancy

M. de Ferrante, K. Maciejewski, P. Batten

Introduction

For our analysis we used Breast Cancer data from the Wisconsin Breast Cancer Database, compiled by Dr. William Wolberg MD, in the early 1990s, found in the `mlbench` R package. The objective of this dataset is to use characteristics of tumors such as cell size and texture to identify malignant tumors. The goal of this analysis is to identify the best learning method for classifying tumors. Our outcome is binary, with breast cancer tumors being classified as either malignant or benign.

Methods used include logistic regression, Support Vector Classification, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), Random Forests, and K-Nearest Neighbors (KNN). While it will be helpful to identify one specific method for predicting future classifications of tumor, it is still advisable to consider the output from all methods, albeit with weighted considerations based on how effective the models are. Ultimately, while machine learning is a very powerful tool for making highly important medical classifications, user discretion should be taken into account when considering interpretability, such as potentially using a simpler model that has nearly the same power as a more complex model.

Data Cleaning

Cl.thickness	Cell.size	Cell.shape	Marg.adhesion	Epith.c.size
1 :145	1 :384	1 :353	1 :407	2 :386
5 :130	10 : 67	2 : 59	2 : 58	3 : 72
3 :108	3 : 52	10 : 58	3 : 58	4 : 48
4 : 80	2 : 45	3 : 56	10 : 55	1 : 47
10 : 69	4 : 40	4 : 44	4 : 33	6 : 41
2 : 50	5 : 30	5 : 34	8 : 25	5 : 39
(Other):117	(Other): 81	(Other): 95	(Other): 63	(Other): 66

Bare.nuclei	Bl.cromatin	Normal.nucleoli	Mitoses	Class
1 :402	2 :166	1 :443	1 :579	benign :458
10 :132	3 :165	10 : 61	2 : 35	malignant:241
2 : 30	1 :152	3 : 44	3 : 33	NA
5 : 30	7 : 73	2 : 36	10 : 14	NA
3 : 28	4 : 40	8 : 24	4 : 12	NA
(Other): 61	5 : 34	6 : 22	7 : 9	NA
NA's : 16	(Other): 69	(Other): 69	(Other): 17	NA

We load in the data and remove the ID column as it is not needed.

Each of nine variables (exclusing Class) is loaded as a numerical factors with values ranging from 1 through 10. Class is benign or malignant, and this is the variable of interest. There are 16 missing values in bare nuclei, as seen in the summary above.

The variables included in the dataset are Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, Mitoses, and

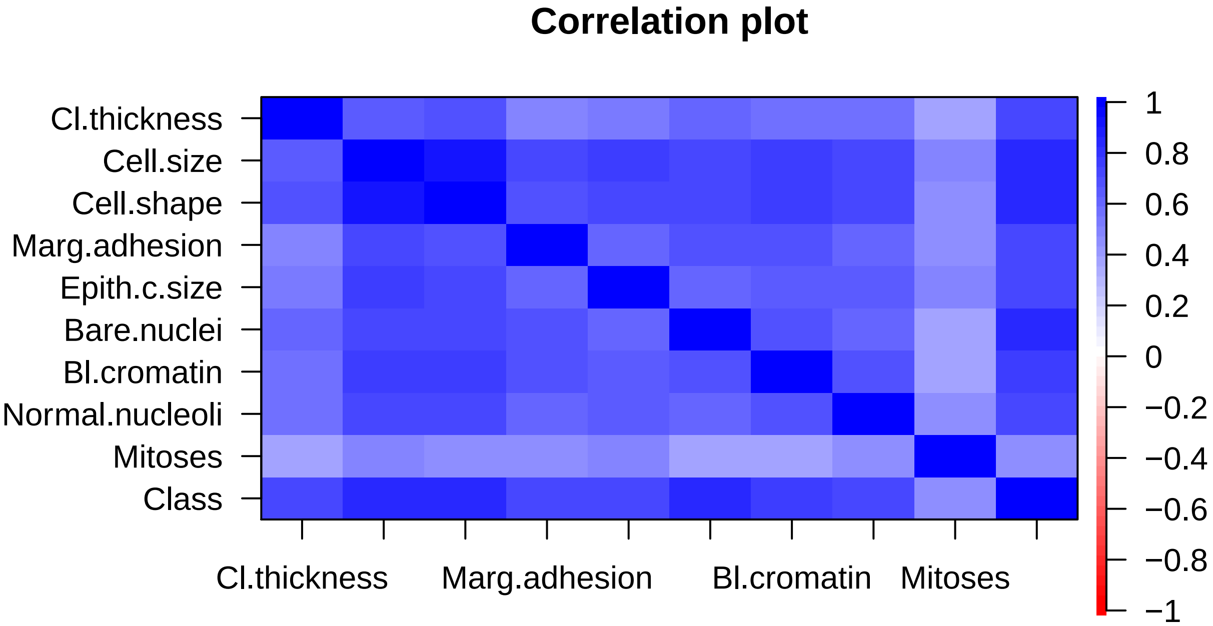
Class. There are 458 subjects with benign growths and 241 with malignant.

We convert the factors into numeric values and remove the NA's. Since there are few missing values compared to the number in the dataset, removing them should not effect our overall analysis power.

Exploratory data analysis

Summary plots

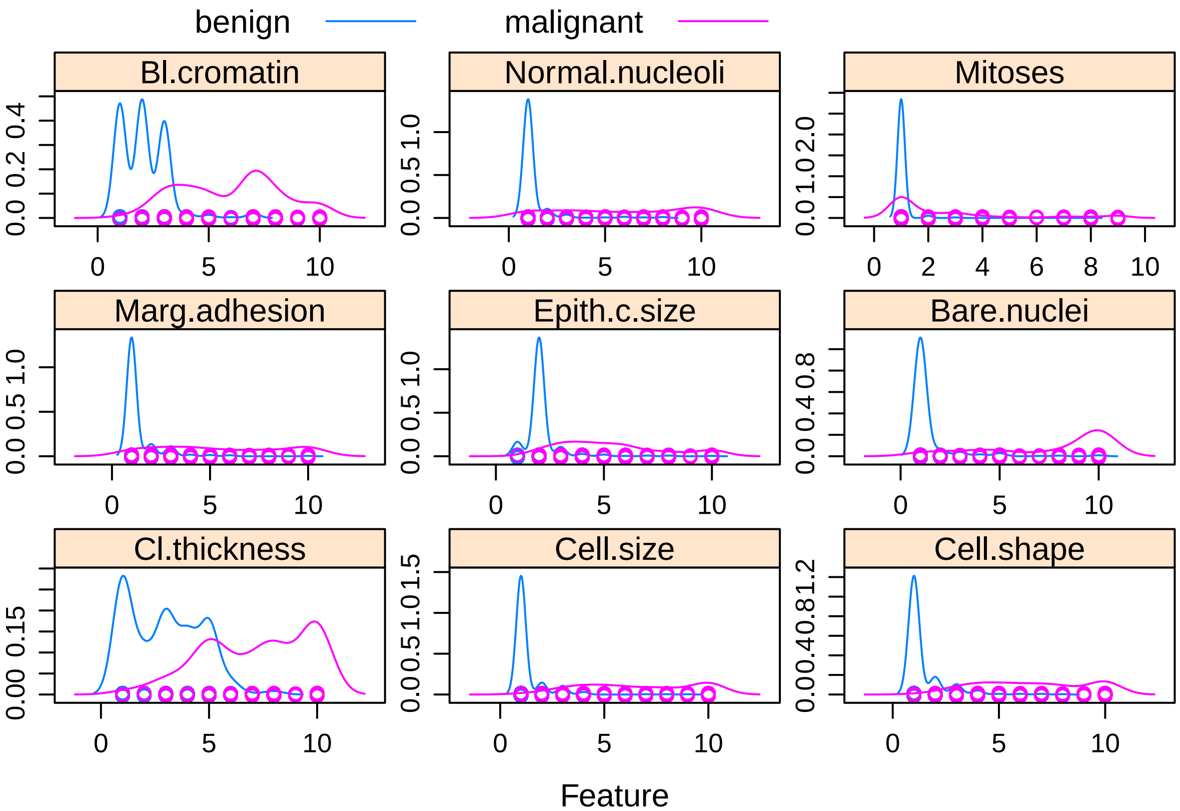
The correlation plot below does not give much information. As expected, most measures have correlation with the outcome. Mitoses, however, has almost no correlation.

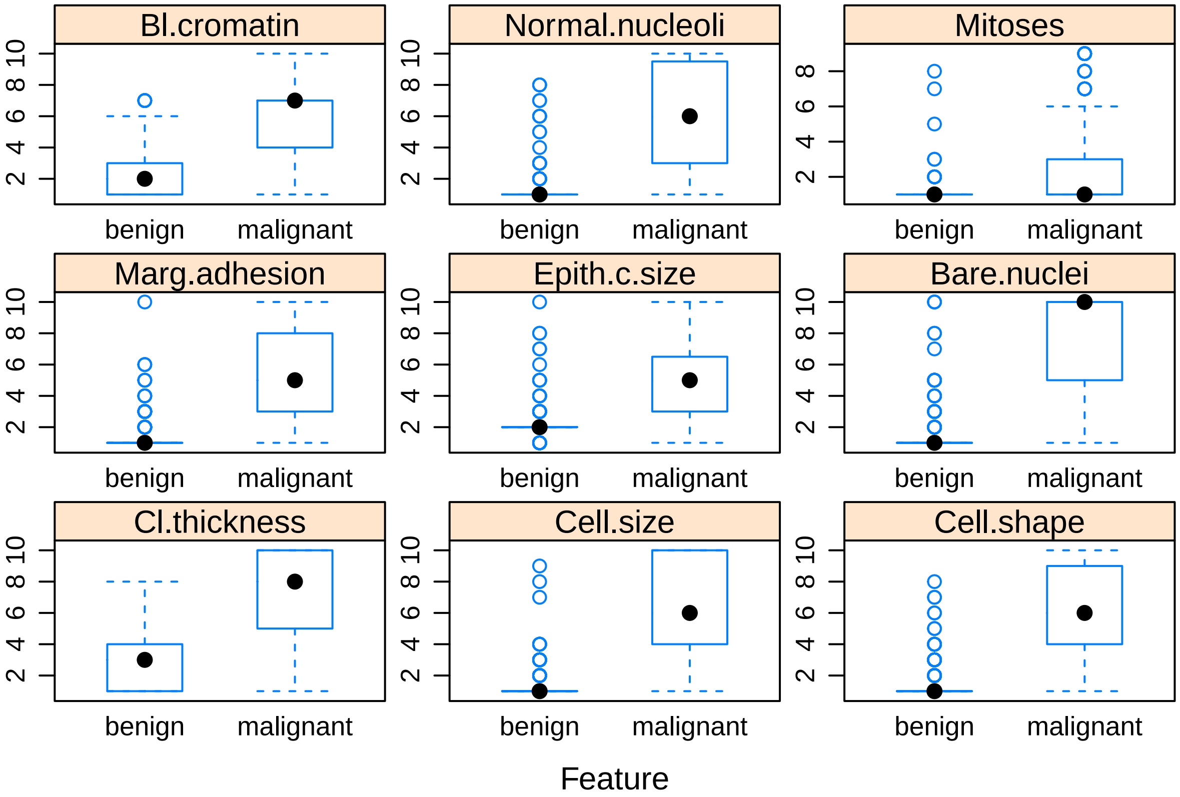


Below are the density plots for the variables. The blue line signifies benign and pink are malignant. We see that there are few malignant subjects with the following: normal nuclei, mitoses, marginal adhesion, single epithelial cell size, uniformity of cell size, uniformity of cell shape.

There are differences in the density plots of the following variables: bland chromatin, bare nucleoli, clump thickness.

The box plots below also show differences in the distributions for bland chromatin, bare nuclei, clump thickness.


benign
malignant
Bl.cromatin
Normal.nucleoli
Mitoses
Marg.adhesion
Epith.c.size
Bare.nuclei
Cl.thickness
Cell.size
Cell.shape
0.0 0.2 0.4
0.0 0.5 1.0
0.0 1.0 2.0
0.0 0.4 0.8
0.00 0.15
0.0 0.5 1.0 1.5
0.0 0.4 0.8 1.2
0 5 10
0 2 4 6 8 10
Feature


Bl.cromatin
Normal.nucleoli
Mitoses
Marg.adhesion
Epith.c.size
Bare.nuclei
Cl.thickness
Cell.size
Cell.shape
2 4 6 8 10
2 4 6 8
benign
malignant
Feature

Supervised Analyses

For all analyses we used the R predict function on the testing set (or training set if looking for train set accuracy) and compared the predictions with the test set classifications.

Logistic Analysis

When we use a generalized linear model with all variables included, at $\alpha = 0.05$ the following appear significant: `Cl.thickness`, `Marg.adhesion`, `Bare.nuclei`, `Bl.cromatin`

Now we split our data into a training and test set so we can make predictions.Since our research area is highly regulated and in the clinical setting, we chose to use an untouched testing set to evaluate the performance of our methods. The training set was generated randomly using about 60% of the original data. The same training and testing data was used throughout the analysis.

Prediction on the training set was 96.83% accurate. 6 benign tumors were misclassified as malignant and 7 malignant tumors were misclassified as benign.

Using the test set, there were only 13 incorrect predictions; 5 benign were predicted to be malignant and 8 that were truly malignant were predicted to be benign. Logistic has 95.24% correct response for test, which is pretty good. Sensitivity is 91.30%, specificity is 97.24%, positive predictive value is 94.38% and negative predictive value is 95.65%, where malignant is the positive class. These are all very high measures and would be good values of interest to doctors and patients.

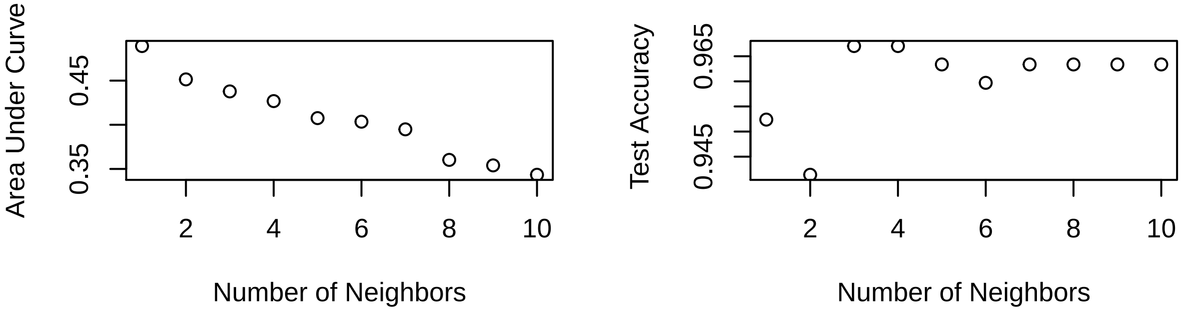
The area under the test ROC curve is 99%.

LDA and QDA

Linear Discriminant Analysis (LDA) resulted in 95.6% accuracy and a kappa statistic of .9 when comparing predictions from the model generated with the training data to the test data classes. Additionally sensitivity was 90.22% and specificity was 98.34% for LDA, and the area under the ROC curve was .9935. For Quadratic Discriminant Analysis (QDA), there was also 95.6% prediction accuracy for the test data and the kappa statistic was slightly improved with a statistic of .9032. QDA had sensitivity of 96.74% and specificity of 95.03%, and the area under the ROC curve was .9856. With higher sensitivity, we have a higher probability of the model accurately predicting malignancy (versus a tumor being benign) given that someone has a malignant tumor. This is important for being able to treat patients using chemotherapy or performing surgery if necessary. With high specificity, we have a higher probability of predicting that someone has a benign tumor given that they have a benign tumor. This is important because we do not want to subject patients to unnecessary risk by performing unnecessary surgery, or putting them on chemotherapy if they don't need to be since it can be very damaging to a patients health. Since both of these values are extremely important for patient health and safety, QDA would be a better choice for a model since we sacrifice too much in sensitivity with LDA. QDA ends up maximizing these values almost equally. The training prediction accuracy for LDA was 96.59%, and the training prediction accuracy for QDA was 95.85%.

KNN

Choosing Number of Neighbors



The plots above demonstrate the test set prediction accuracy and area under the ROC curve for different number of neighbors. The goal is to choose the number of neighbors that maximizes both of these values. It appears the best number of neighbors to achieve the highest of both is 3 nearest neighbors. This value appears to maximize accuracy without losing too much in area under the curve.

Number of Neighbors	Area Under Curve	Accuracy
1	0.4891304	0.9523810
2	0.4513872	0.9413919
3	0.4377853	0.9670330
4	0.4267956	0.9670330
5	0.4075486	0.9633700
6	0.4034951	0.9597070
7	0.3948475	0.9633700
8	0.3602570	0.9633700
9	0.3540115	0.9633700
10	0.3434122	0.9633700

Results

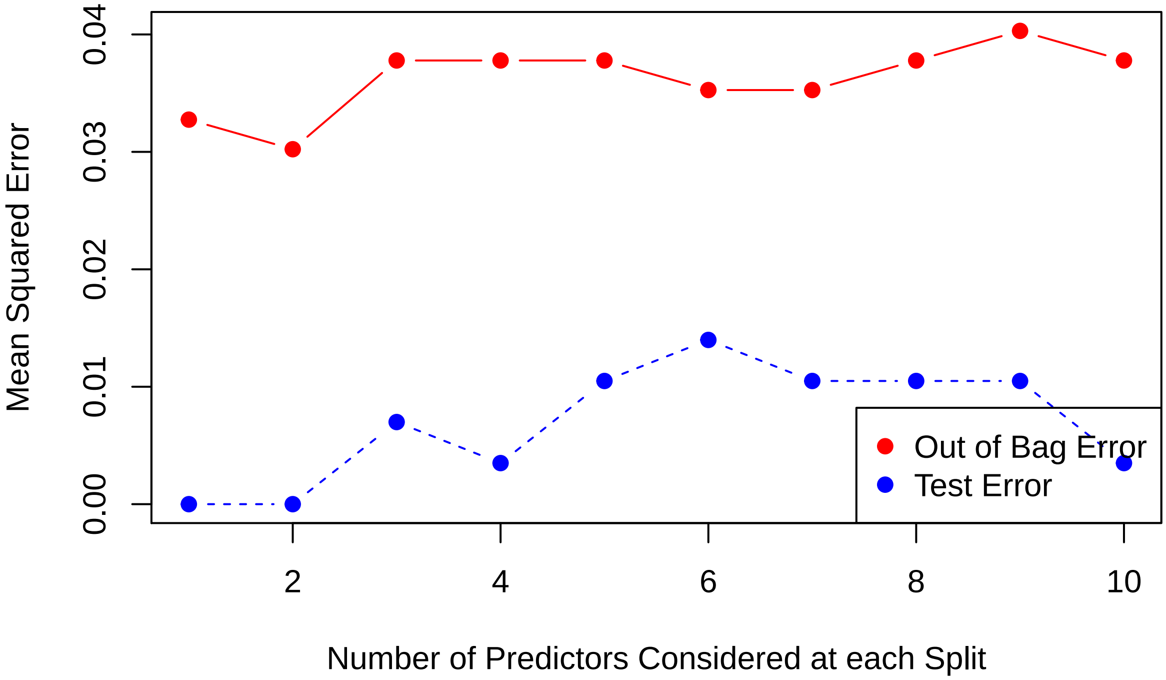
After comparing the area under the ROC curve and the percentage of accuracy in predicting tumor type, the best number of neighbors was chosen to be 3. The chosen number of neighbors has area under the curve equal to .4378 and prediction accuracy against the test set of 96.7%. The kappa statistic is .9256. The sensitivity and specificity are 93.48% and 98.34%, respectively. The train data accuracy for KNN is 97.56%. K nearest neighbors is one of the simplest machine learning algorithms, and it is relatively interpretable. The prediction accuracy KNN with 3 neighbors is also pretty high making this a good model for the data.

Support Vector Classification

When training a support vector machine using classification, we must also make the choice between using a linear kernel and a radial kernel and then tune the cost of constraints violation and gamma parameter (just for radial). The linear kernel with tuning resulted in better performance using cost = .01, and with this model the prediction accuracy on the test data was 95.6% with a kappa statistic of .9011. The sensitivity for the linear kernel was 94.44% and the specificity was 96.17%. Support vector classification with a linear kernel had a train data accuracy of 97.8%. Support vector classification with a radial kernel had test set accuracy of 94.87% and a kappa statistic of .8858 (after tuning the parameters and choosing cost = 10 and gamma = .1). The sensitivity for the radial kernel was 91.49% and the specificity was 96.65%, and the training data prediction accuracy was 99.27%. In comparing test set prediction accuracy of support vector classification using a linear kernel versus a radial kernel after tuning parameters, the linear kernel performed better on this data. Although the training data prediction accuracy was higher with a radial kernel, we want a model with better testing accuracy.

Random Forest

	MeanDecreaseGini
Id	1.835143
Cl.thickness	3.233681
Cell.size	131.547742
Cell.shape	6.833064
Marg.adhesion	3.342785
Epith.c.size	1.113545
Bare.nuclei	10.231893
Bl.cromatin	9.176162
Normal.nucleoli	9.904904
Mitoses	1.026219



Applying random forests techniques to predict tumor classification results in 3 variables being deemed important in contributing to this classification, as suggested by the out-of-bag error being minimized at 3 predictors and the relatively low test error at 3 predictors. Restricting the number of variables used at each split per tree to three yields the best out-of-bag error, as well as the (tied-for) best test error. 3 variables is also relatively simple, which is important to consider when building classifier models. The 3 variables that most effectively decrease impurity at each decision node are cell size, bland chromatin (which measures texture of the nucleus), and bare nuclei, a term used to describe the composition of the nucleus. The accuracy of the random forests in predicting tumor type are as follows: Accuracy = 95.45%, Kappa statistic = 0.9012, Sensitivity = 92.31%, Specificity = 97.25% (where sensitivity predicts Malignant tumors). Comparisons to other learning methods can be seen in the table in the Conclusion section.

Conclusions

Overall Comparison of Methods

	Test Accuracy	Train Accuracy	Kappa Statistic	Sensitivity	Specificity
Logistic Regression	95.24%	96.83%	0.8926	91.30%	97.24%
Random Forest	95.45%	100%	0.9012	92.31%	97.25%
KNN	96.7%	97.56%	0.9256	93.48%	98.34%
LDA	95.6%	96.59%	0.9000	90.22%	98.34%
QDA	95.6%	95.85%	0.9032	96.74%	95.03%
Support Vector Classification	95.6%	97.8%	0.9011	94.4%	96.17%

All of our models had a very high test prediction accuracy. We also examined the kappa statistic as a measure of prediction accuracy since our outcome is unbalanced. We expected more complicated models to be a better fit, and were surprised to find KNN to be such a good fit for the data. We were also surprised that prediction accuracy was so high in all of the models. Taking into account that it could be harmful to patients if both sensitivity and specificity are not maximized while also wanting to maximize overall prediction accuracy, we decided the best models for predicting tumor malignancy would be KNN or QDA. We also want to choose a model that is the most interpretable, thus our conclusion for the best model to predict tumor malignancy is KNN. KNN had an overall test prediction accuracy of 96.7% [CI 95.56%, 98.82%] with a kappa statistic of .9256, which was the highest kappa statistic out of all the models.