# P8106 Final Report

*Si Li (sl4657), Weiwei Qi (wq2151), and Qimin Zhang (qz2392)*

## Introduction

Breast cancer is considered as one of the most common types of cancer among women all over the world, and machine learning methods for breast cancer classification has been a hot topic for many years. In this report, we want to try multiple classification methods on Wisconsin Breast Cancer Database from the University of Wisconsin Hospitals, Madison from Dr. William H. Wolberg, and compare their performance to see which one is the best for breast cancer classification.

The dataset contains tumor subjects and predictive variables, and the variable 'class' is the type of this tumor, where 'M' is malignant and 'B' is benign. Explanation for some variables: 'clump_thickness': Thickness of clump, from 1 to 10; 'size_uniformity': Uniformity of cell size, from 1 to 10; 'bland_chromatin': Bland chromatin, from 1 to 10.

The variable 'bare_nucleoli' was recorded in the form of character and NA was denoted by '?'. When cleaning the data, we replace '?' by NA and fill it with median. The data is split into two parts, where 2/3 of the original data is used for training the models and the left out data will be used for testing and evaluating model performance.

## Exploratary Analysis

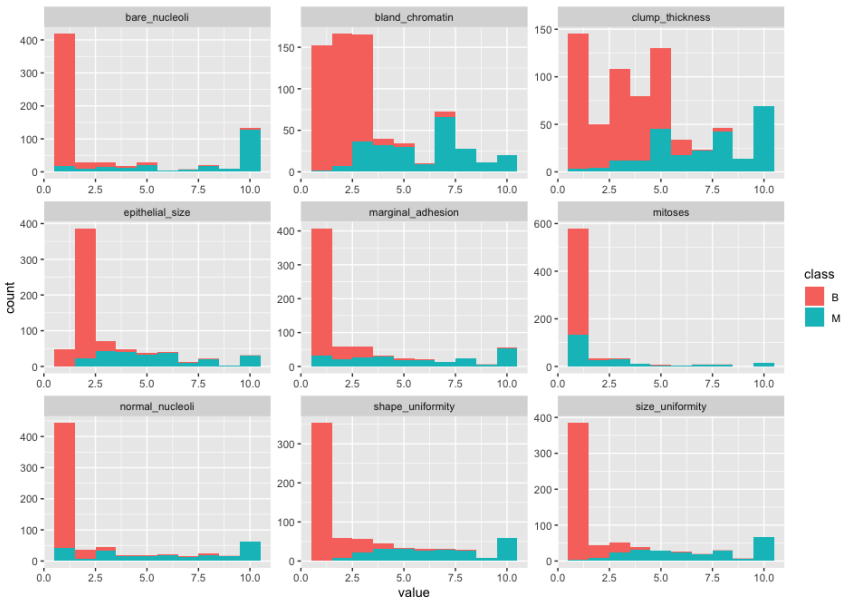
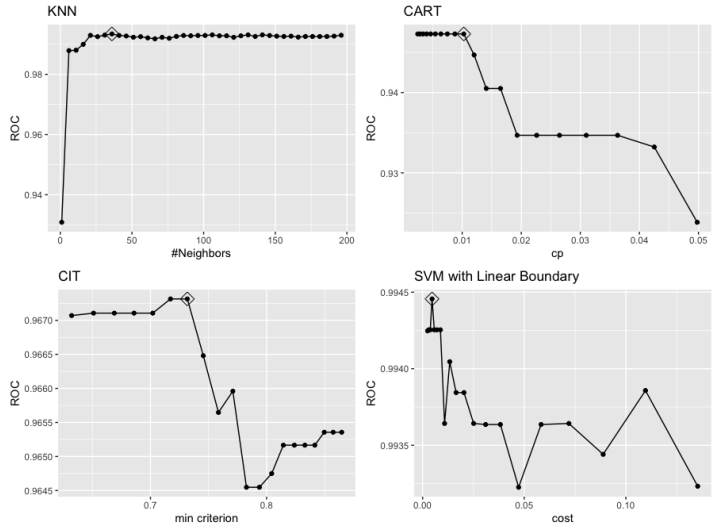



Figure 1: Distribution of Predictors

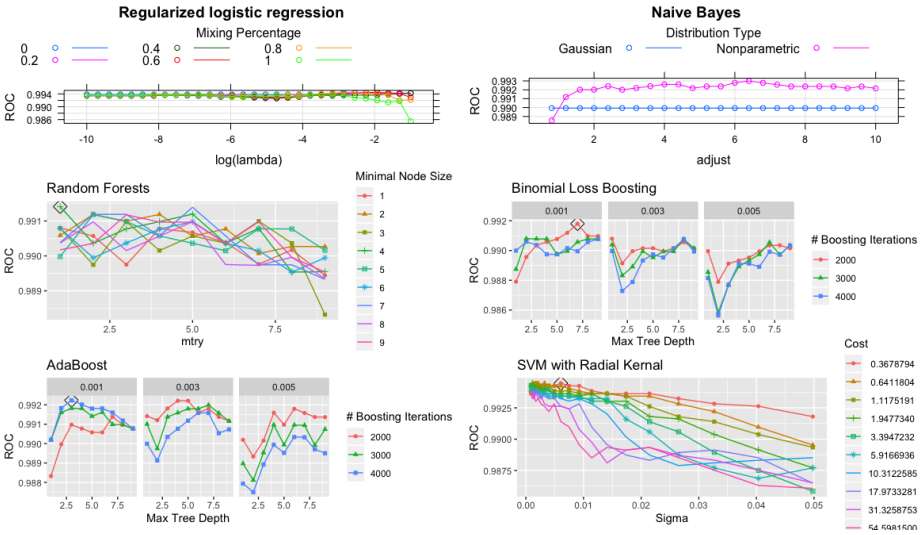
We visualize the distributions of all predictive variables stratified by class (Malignant or Benign). The distributions are different across the class.

# Models

For classification tasks, multiple models are fitted to test against the data for evaluating performance. For linear methods, Logistic, Regularized Logistic regression and LDA are fitted; For discriminant analysis, QDA and Naive Bayes methods are used; Tree-based methods are implemented as well, including Random Forest, Boosting; Support Vector Machines are considered as well. All models are built using the caret package, and the optimal model is selected based on maximizing ROC with 10-fold cross-validation, repeated five times. And below is the visualization of cross-validation:




(a)




(b)

Figure 2: Cross-validation

Below is the ROC comparison:

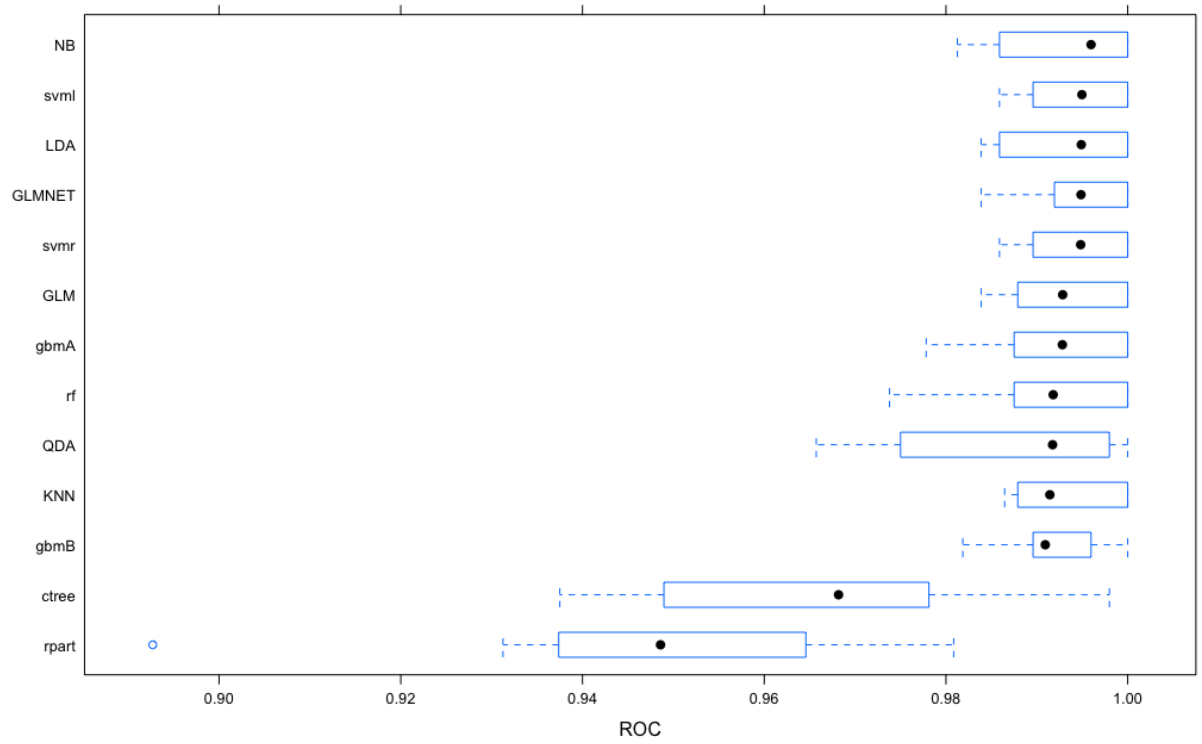



Figure 3: ROC Comparison

Among linear methods, LDA produces the highest AUC for ROC curve (0.9949), and Regularized Logistic Regression has almost the same performance.

The best tuned quadratic discriminant model is Naive Bayes which produces the highest AUC (0.996) among the 3 non-linear models.

The results of classification trees showes the lowest AUC among all other models, as flexibility lower the level of predictive accuracy. By aggregating decision trees, the predictive performances of trees are substantially improved, AUC are up to 0.99.

For Support Vector Machines, we fit 2 models using linear kernel and radial kernel respectively. The AUC of linear kernel is 0.9949 and the AUC of radial kernel is 0.9948.

Therefore we select Naive Bayes as our final model.

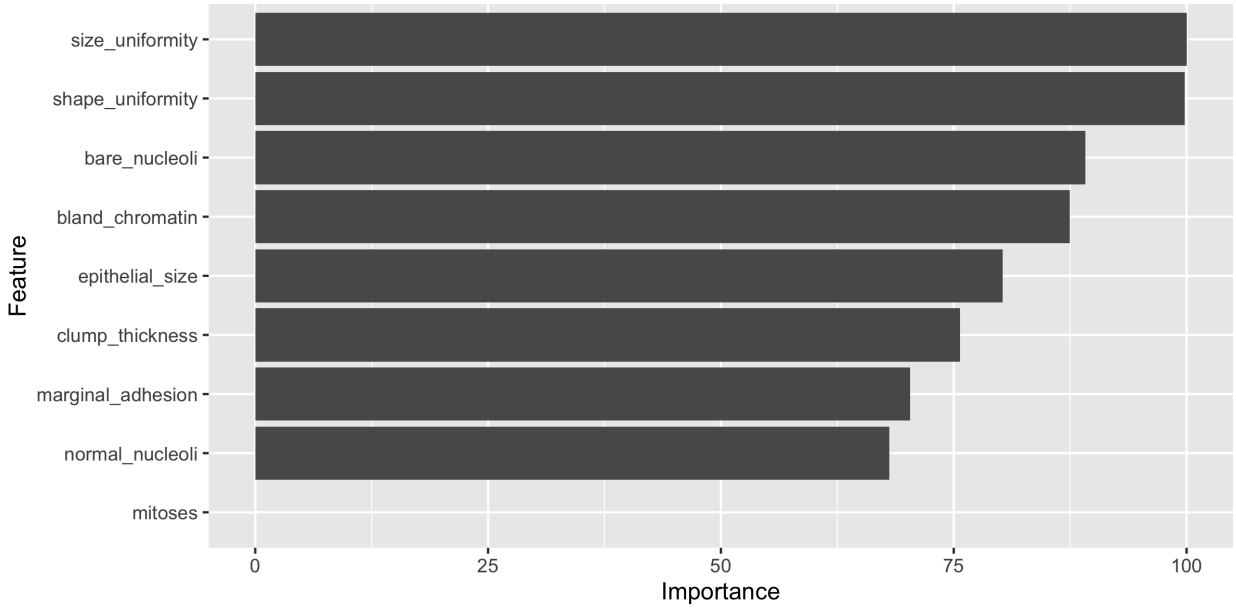



Figure 4: Variable Importance from Naive Bayes Model

Naive Bayes model is used to explore the variable importance. According to the plot above, uniformity of cell size, uniformity of cell shape and bare nuclei are the top 3 important features in predicting whether the cancer is benign or malignant.

As for test data performance, all but unstable single trees are quite constant with cross-validation performance. We evaluate the final model on test set and yield the confusion matrix below:

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   B   M
##          B 151  10
##          M   1  70
##
##                Accuracy : 0.9526
##                  95% CI : (0.9167, 0.9761)
##     No Information Rate : 0.6552
##     P-Value [Acc > NIR] : < 2e-16
##
##                   Kappa : 0.8922
##
##  Mcnemar's Test P-Value : 0.01586
##
##             Sensitivity : 0.8750
##             Specificity : 0.9934
##          Pos Pred Value : 0.9859
##          Neg Pred Value : 0.9379
##              Prevalence : 0.3448
##          Detection Rate : 0.3017
##    Detection Prevalence : 0.3060
##       Balanced Accuracy : 0.9342
##
```

```
##        'Positive' Class : M
##
```

The overall accuracy is 0.9526, which is quite satisfying. We can see the sensitivity is rather low and specificity is high, indicating that this classification model is suitable for ruling out almost everyone who doesn't have the disease and won't generate many false-positive results. For example, this method can serve as the second test for patients who are positive for the first test.

Naive Bayes assumes features are independent in each class while the truth may not be like that. It's also useful when the number of predictors is large, but in this case the number of predictors is not large.

# Conclusions

Based on cross-validation training AUC, Naive Bayes is the best-performing model among all, and SVM with linear kernal, Adaboost and LDA are the best-performing models in their own categories respectively. All fitted models had good AUC (>0.9) and most of them were as high as 0.99, this meets our expectation, since from the density plot we could see that the distributions of most predictors under each category are well-separated.

The variable importance provided by most models suggest that uniformity of cell size, uniformity of cell shape and bare nuclei are the most influential predictors in breast cancer type of classification. This result also meets our expectation as in exploratory analysis we observed clusters and the particularly salient difference in Logistic regression is perhaps due to collinearity.