# P8106 Final Report

Si Li, Weiwei Qi and Qimin Zhang

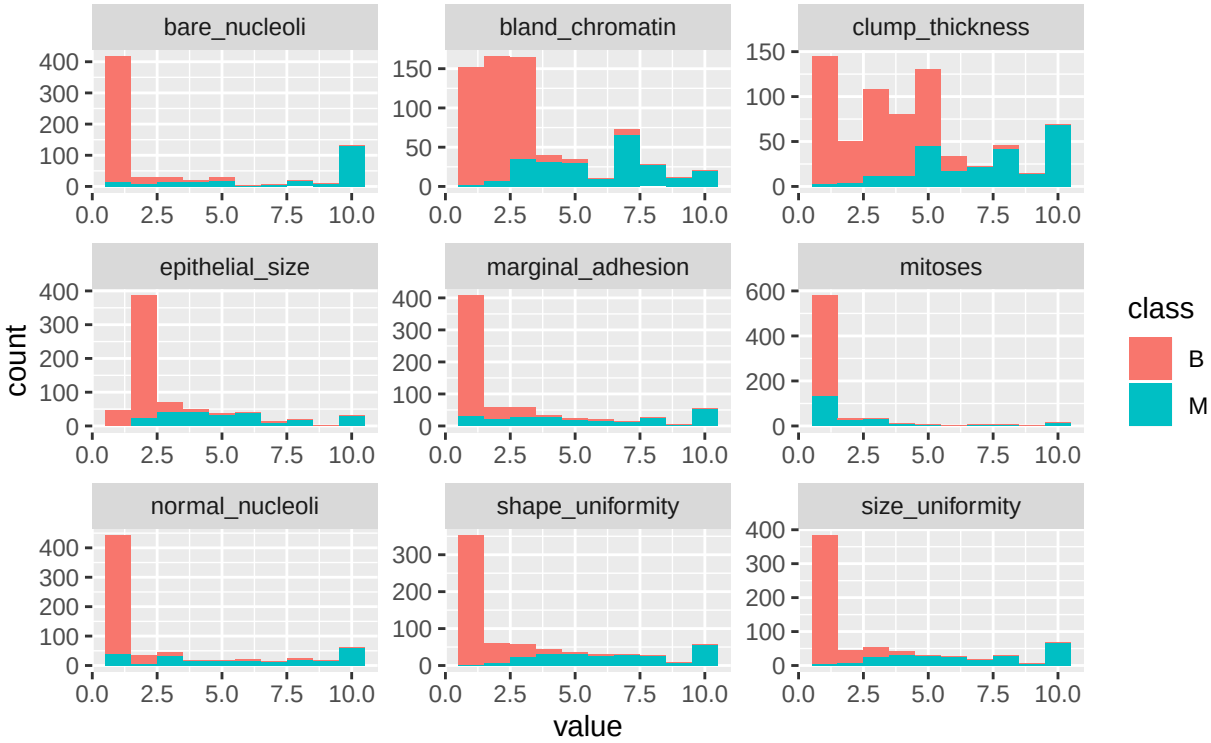
## Introduction

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 699 tumor subjects and 9 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.

## Exploratary Analysis



## Models

For classification task, mulyiple models are fitted to test against the data for evaluating performance. For linear methods, Logistic and Regularized Logistic regression are fitted; For discriminant analysis, LDA, QDA, and Naive Bayes methods are used; Tree-based methods are implemented as well, including Random Forest, Boosting. All models are builed using the caret package, and the optimal model is selected based on maximizing ROC with 10-fold cross-validation, repeated five times.

The data is split into two part, 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.

## Results

## Conclusions