P8106 Final Report

Si Li, Weiwei Qi and Qimin Zhang

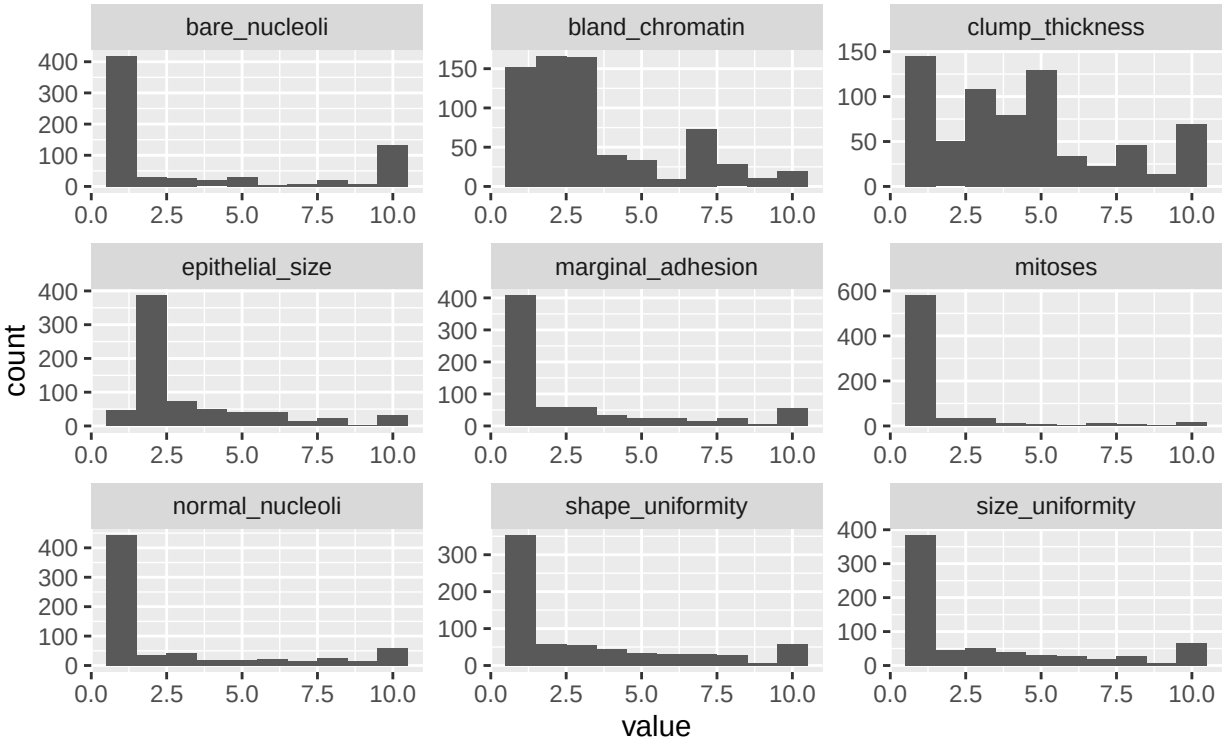
Introduction

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.

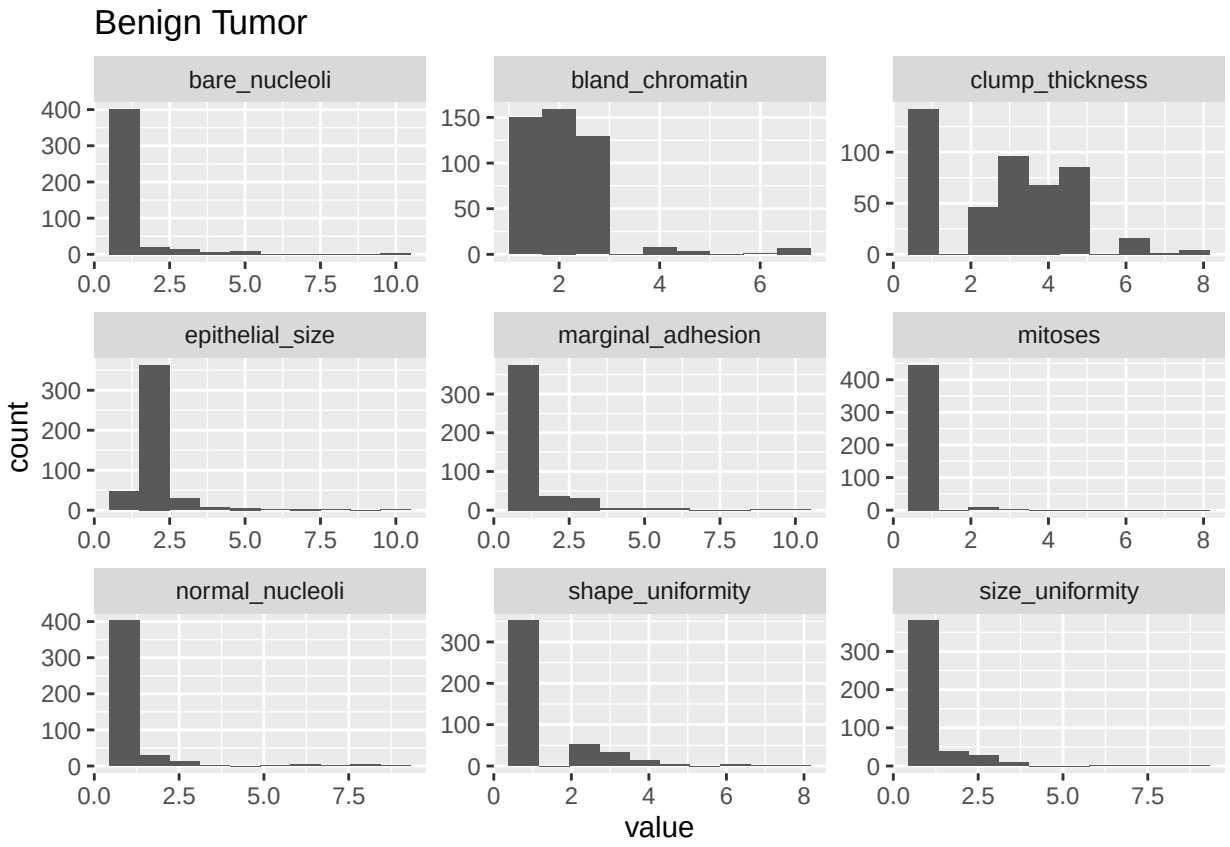
The dataset contains 699 tumor subjects and 9 predictive variables, and the variable 'class' is the type of this tumor, where 'M' is malignant and 'B' is benign. Explanation for some variables: 'clump_thickness': Thickness of clump, from 1 to 10; 'size_uniformity': Uniformity of cell size, from 1 to 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.

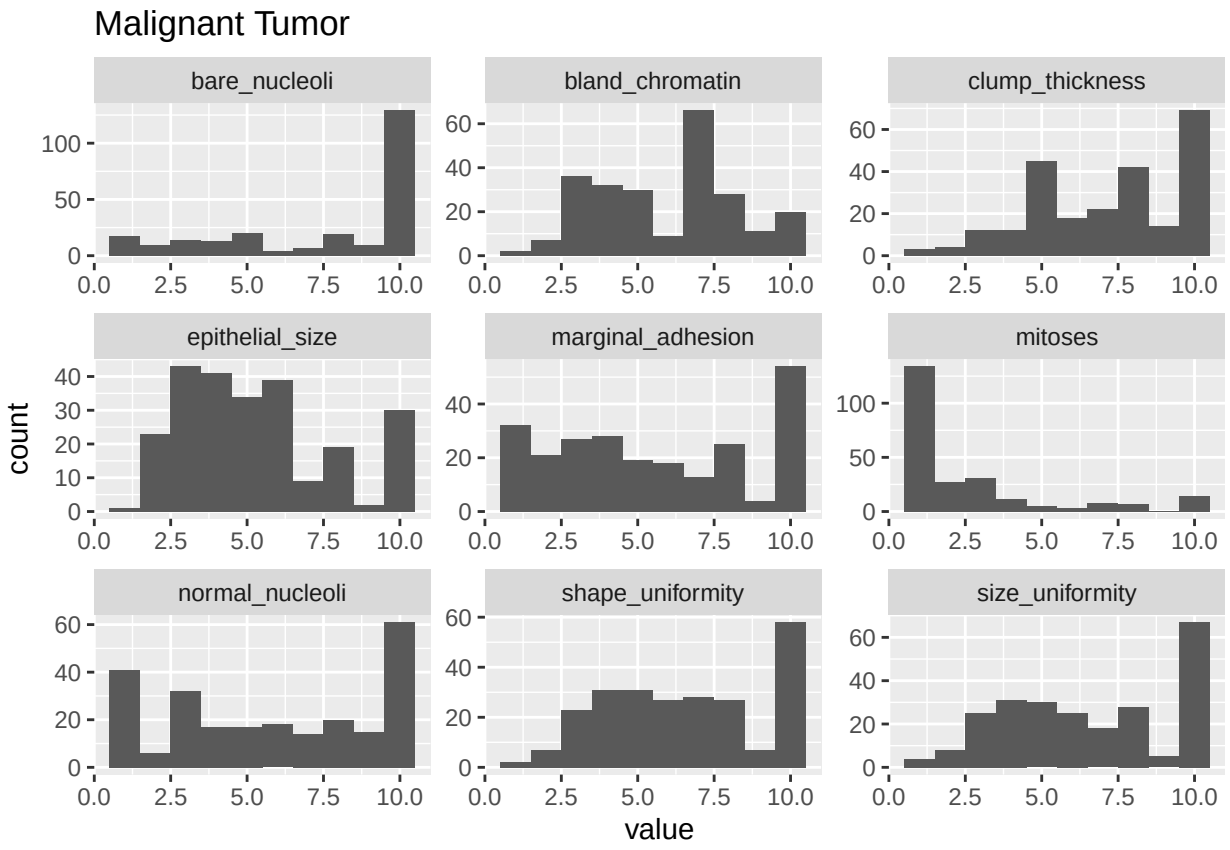
Exploratary Analysis



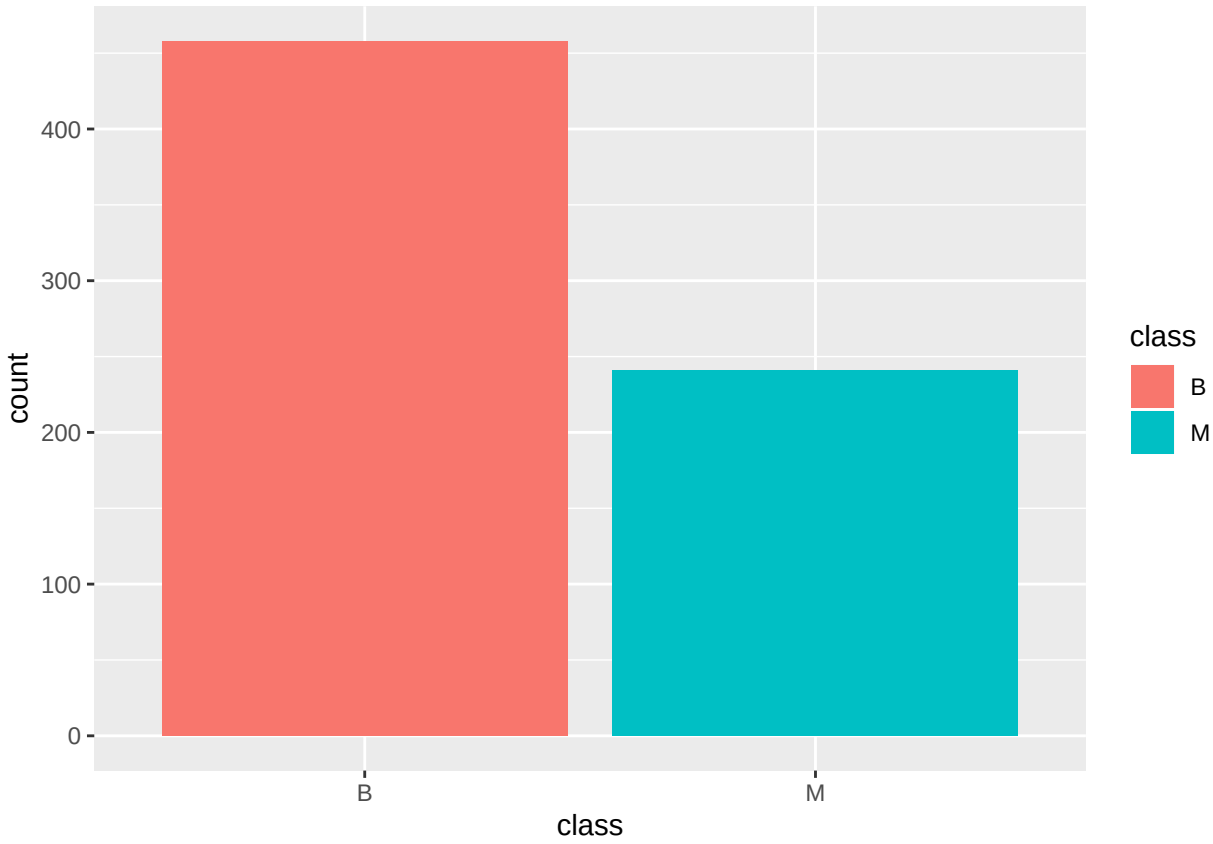
```
data %>%
  filter(class == "B") %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
  facet_wrap(~ key, scales = "free") +
  geom_histogram(bins = 10) +
  ggtitle("Benign Tumor")
```




```
data %>%
  filter(class == "M") %>%
  keep(is.numeric) %>%
  gather() %>%
  ggplot(aes(value)) +
  facet_wrap(~ key, scales = "free") +
  geom_histogram(bins = 10) +
  ggtitle("Malignant Tumor")
```




```
data %>%
  ggplot(aes(x = class, fill = class)) +
  geom_bar()
```



Models

Conclusions