Contoso

BREAST CANCER DETECTION

BY: GHATTA TRIVEDI

DATE: NOVEMBER 14TH 2024

ABSTRACT

- I used the Breast Cancer Wisconsin dataset to attempt to automate the classification of tumors as malignant or benign.
- Initial data exploration involved visualizing feature distribution using density plot, scatterplot, box plot, and bar plot.
- Then, I built a decision tree using the rpart function and evaluated its performance using a confusion matrix.
- Next, I applied a Naïve Bayes classification model to the same dataset.
- I compared the accuracy, sensitivity, and specificity of both models.
- The project concluded with a check of previous understanding using k fold cross validation.

Contoso

INTRODUCTION OF THE DATASET

DATASET: BREAST CANCER WISCONSIN DATASET

- Some of the features related to the tumor included in the dataset:
  - Clump Thickness
  - Uniformity of Cell Size
  - Uniformity of Cell Shape
  - Marginal Adhesion
  - Single Epithelial Cell Size

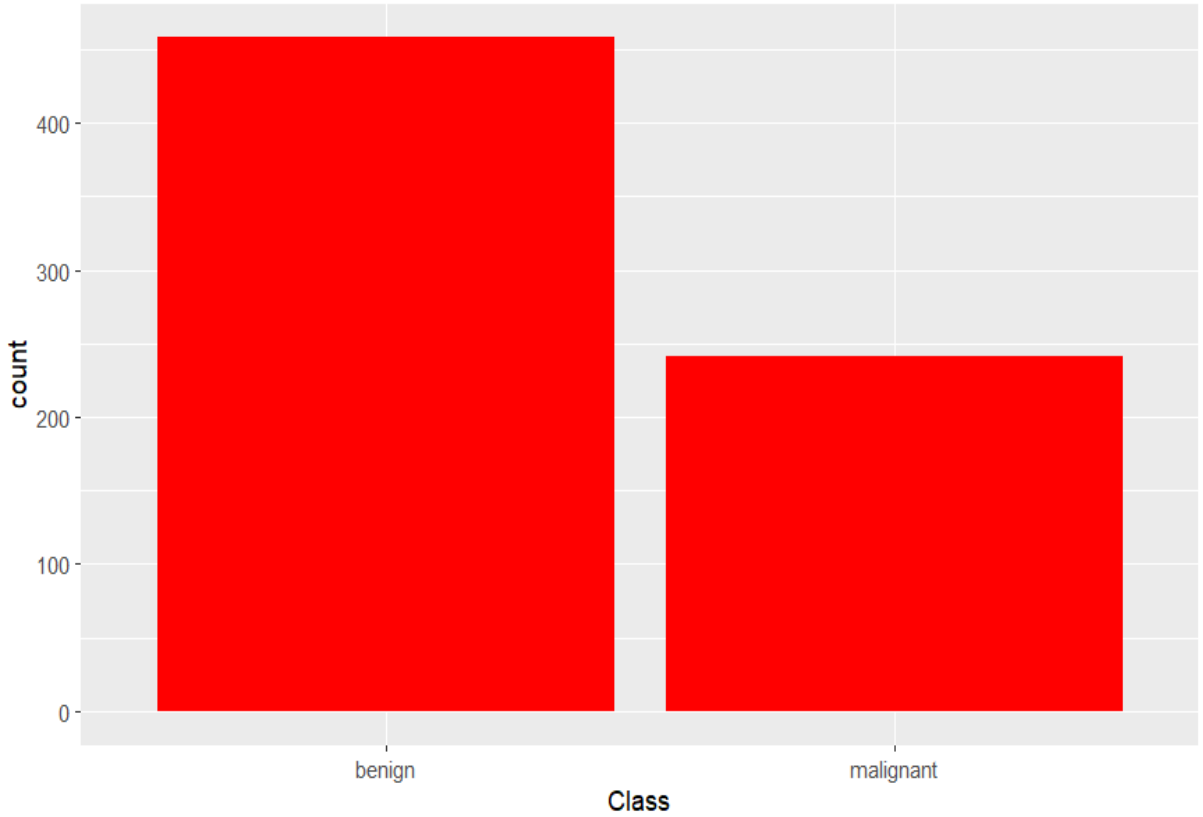
- Goal: To classify the tumor samples based on the features into either Malignant or benign categories.
- This type of classification is vital for early detection of breast cancer, which can significantly impact treatment outcomes and patient survival.

TARGET VARIABLE: CLASS MALIGNANT OR BENIGN

Contoso

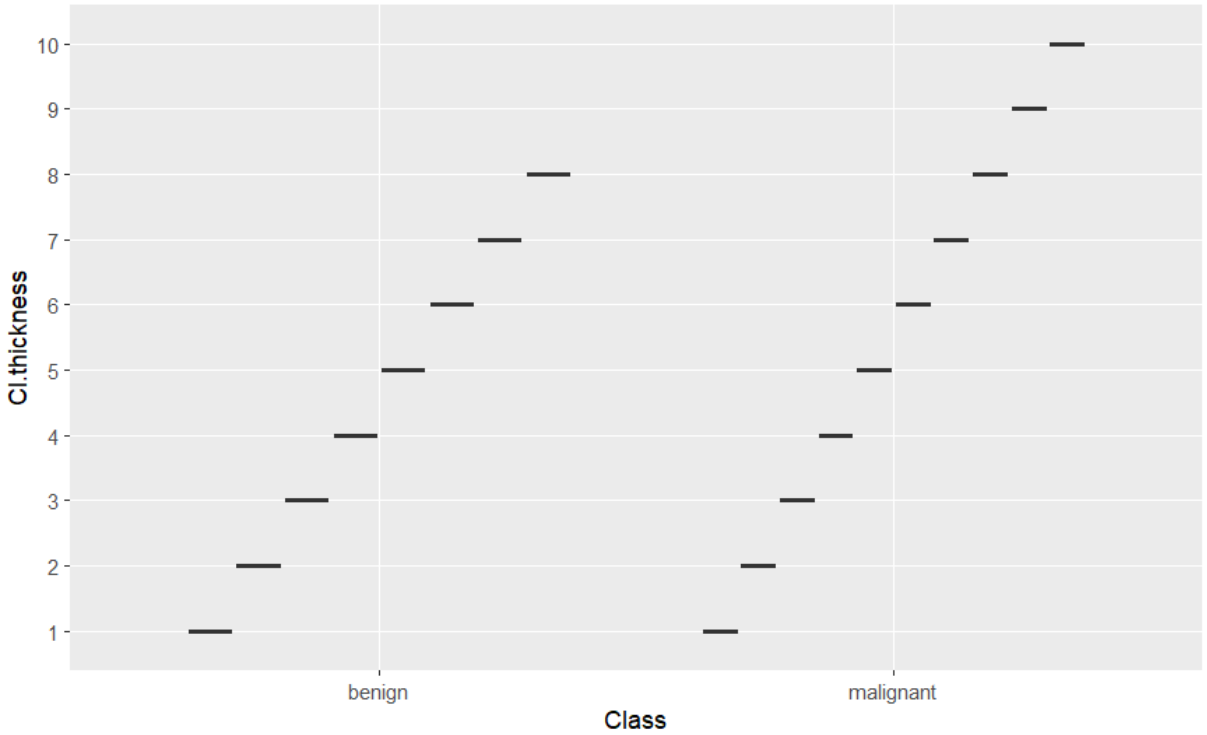
DATA EXPLORATION AND VISUALIZATION (METHODS)

Bar Plot of Class vs. Count



The bar plot reveals that the data set is imbalanced.

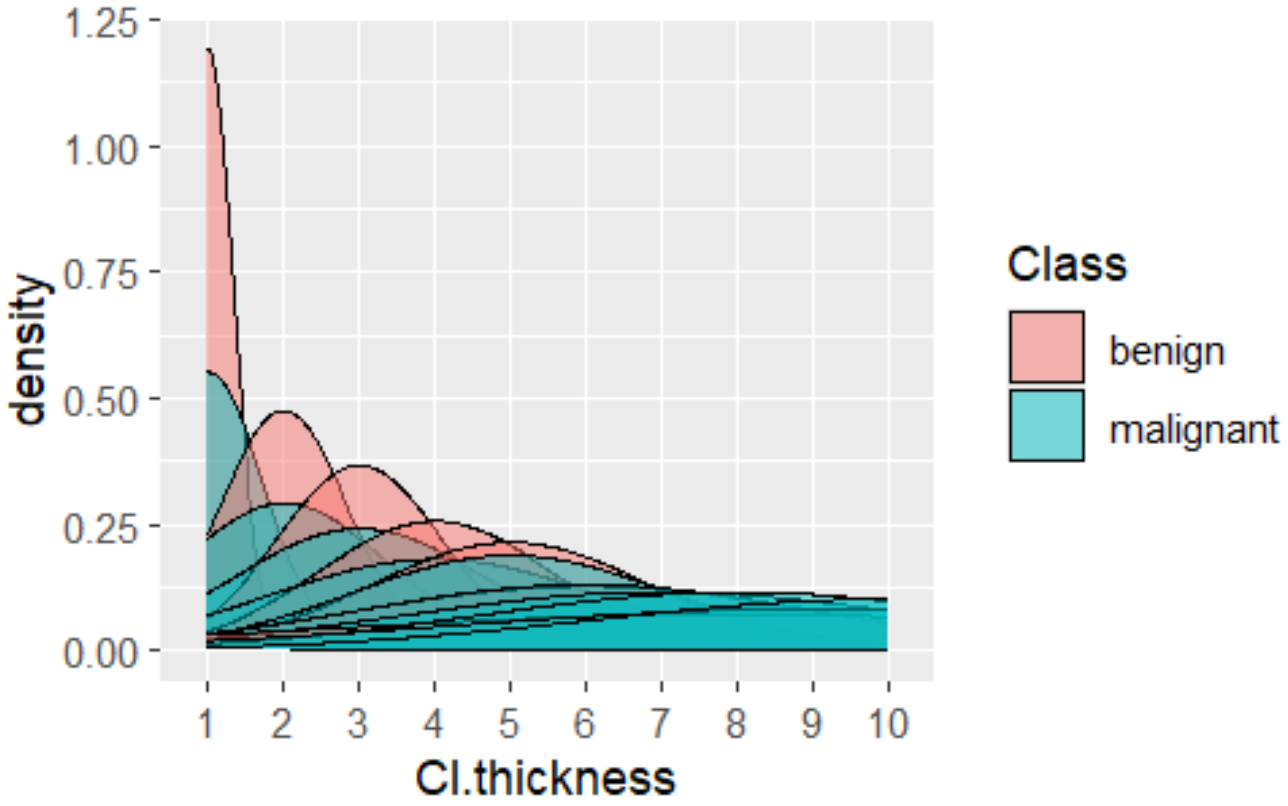
Box Plot of Clump Thickness vs. Class



The box plot shows that clump thickness may not be used as a predictor.

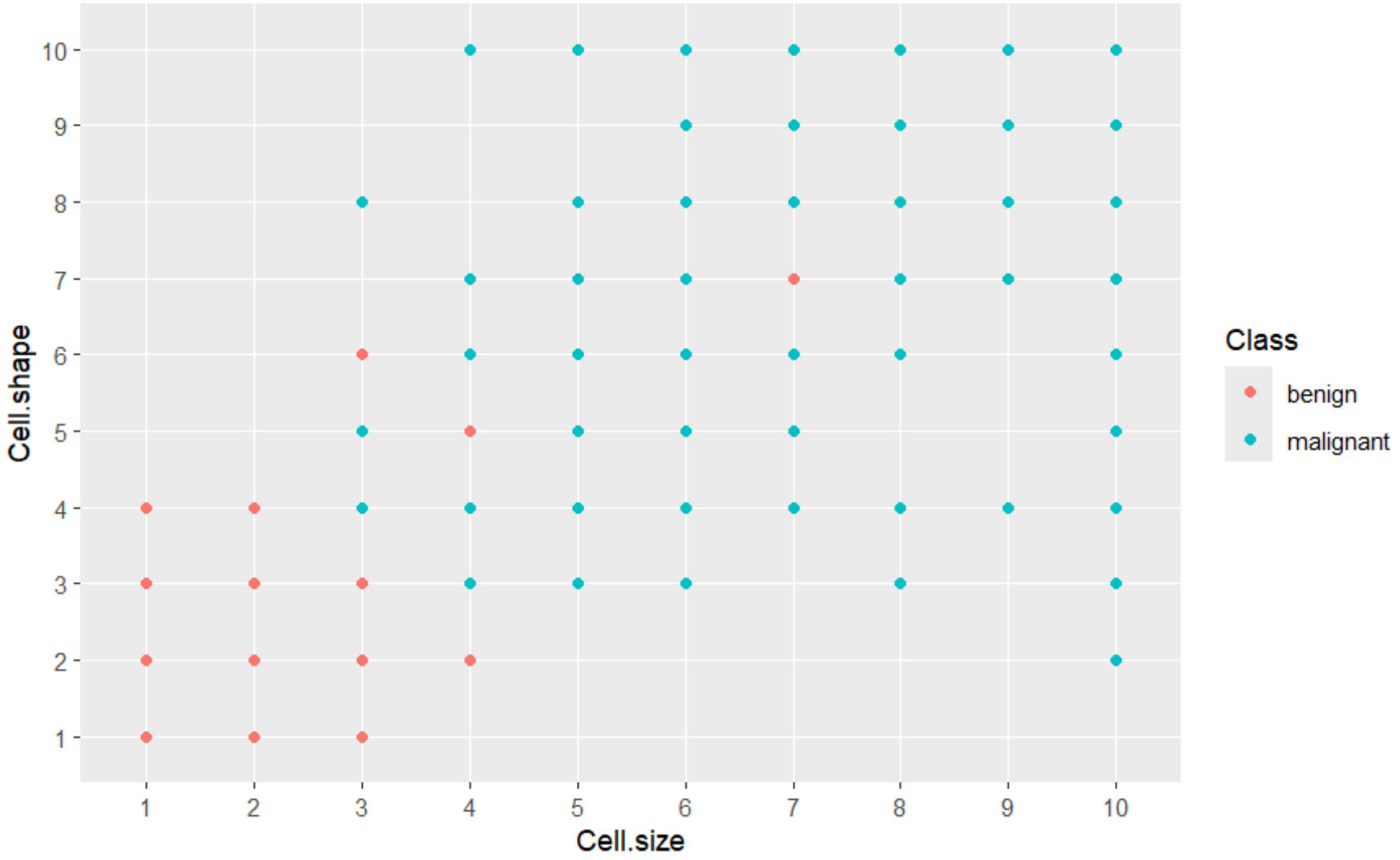
DENSITY PLOT OF CLUMP THICKNESS (METHOD)

- Previously, clump thickness seemed to be a good predictor for classification.
- Reinforced idea:
  - Malignant tumors often have a higher clump thickness
- But, there is little overlap in the middle of the plot that suggest that clump thickness should not be used for classification.



SCATTERPLOT OF CELL SIZE VS CELL SHAPE (METHOD)

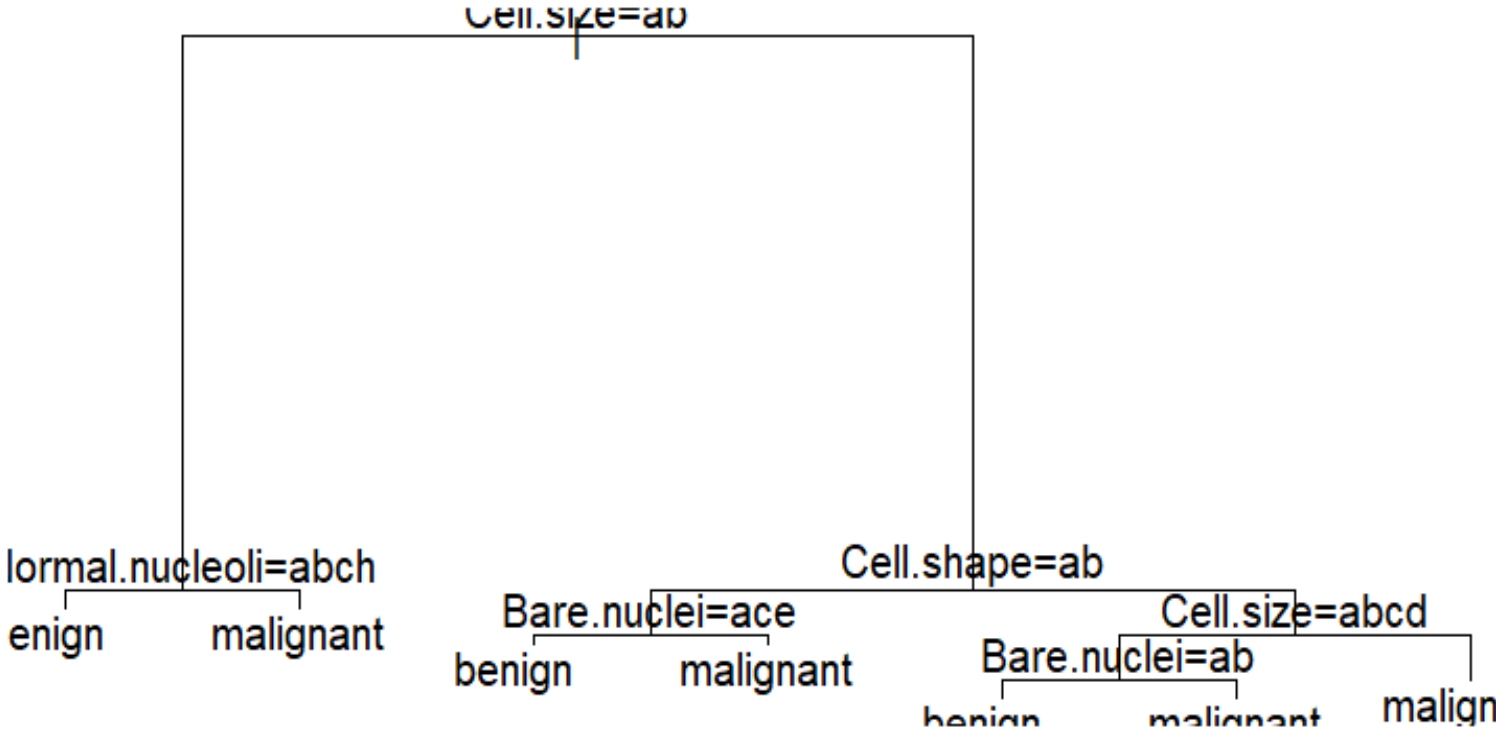
- Cell size and cell shape is good predictor for classification.
  - The malignant samples tend to be more spread out and appear in a separate region from the Benign samples.



Contoso

R PART (MATERIAL)

- Certain features such as cell shape, were critical for differentiating between benign and malignant tumors.
  - As predicted, clump thickness is not the best predictor.
  - Instead, cell size seems to be the top predictor of classification.
- The decision tree demonstrated a reasonable classification performance, though tuning may be required to reduce overfitting.



R PART CONFUSION MATRIX (MATERIAL)

Confusion Matrix and Statistics

Prediction	Reference: Benign	Reference: Malignant
Benign	442	9
Malignant	16	232

Accuracy : 0.9642
95% CI : (0.9477, 0.9767)
No Information Rate : 0.6552
P-Value [Acc > NIR] : <2e-16

- The model achieved an accuracy of 96.42% which indicates good performance.
  - Even though the data is imbalanced, the accuracy is significantly higher than No Information Rate
    - Suggests that the model is effective in distinguishing between classes
- Additionally, high recall and high precision
- P value is extremely small
- This confusion matrix suggests that the model preforms well with very few misclassifications

NAÏVE BAYES CONFUSION MATRIX (MATERIAL)

Confusion Matrix and Statistics

Prediction	Reference: benign	Reference: malignant
benign	443	3
malignant	15	238

Accuracy : 0.9742
95% CI : (0.9596, 0.9847)
No Information Rate : 0.6552
P-Value [Acc > NIR] : < 2.2e-16

- Again, the overall conclusions are same.
- NB accuracy is slightly higher than Rpart .
- Their No Information rates are the same.
- Both models perform well, but the Naïve Bayes shows slightly better overall performance compared to rpart.
- However, both models demonstrate strong predictive capabilities, making them both suitable for this dataset

RESULTS AND DISCUSSION

RESULTS

- Various visualization techniques helped explore the relationship between features and the target class
- Naïve bayes had a slightly higher accuracy and better performance in terms of sensitivity and specificity

DISCUSSION

- Despite the class imbalance both models managed to achieve high performance.
  - However, the NB model handled imbalance better, providing a higher specificity.

Contoso

HOWEVER.. UPON USING K FOLD CROSS VALIDATION...(METHOD)

```
Call:
summary.resamples(object = results)

Models: rpart, naive_bayes
Number of resamples: 5

Accuracy
                Min.  1st Qu.
rpart       0.9197080 0.919708
naive_bayes 0.6470588 0.649635
              Median      Mean
rpart       0.9264706 0.9385144
naive_bayes 0.6496350 0.6500751
             3rd Qu.      Max. NA's
rpart       0.9558824 0.9708029    0
naive_bayes 0.6496350 0.6544118    0
```

- Upon cross validation to check accuracy of our results, different results popped up.
  - CART/Decision tree was the highly accurate method.
    - Cp= 0.0209 (complexity parameter)
    - Kappa = .8646 which indicates strong agreement between predicted and actual values as opposed kappa(nb)=0.
  - As the complexity parameter increases, the accuracy and Kappa score decrease, which shows that higher complexity can lead to overfitting
  - Additionally, rpart received the highest accuracy (93.85%) compared to nb (65%).

FINAL RESULTS AND DISCUSSION

RESULTS

- CART performed better than NB over multiple samples
- NB struggles likely due to its assumption of feature independence
- NB also risks overfitting the data
- K-fold helped determine optimal tuning parameters

- Since decision tree was the better performing model, it should be used to automate cancer diagnosis.
- However, we should still try to check and make this model work better using boosting and bootstrapping

DISCUSSION

Contoso

BOOSTING (METHOD)

Confusion Matrix and Statistics

Prediction	Reference 0	1
0	109	3
1	1	58

Accuracy : 0.9766
95% CI : (0.9412, 0.9936)
No Information Rate : 0.6433
P-Value [Acc > NIR] : <2e-16

Kappa : 0.9487

Mcnemar's Test P-Value : 0.6171

Sensitivity : 0.9909
Specificity : 0.9508
Pos Pred Value : 0.9732

- The model has an accuracy of 97.66%, meaning it correctly classified almost 98% of the instances in the dataset
- The 95% CI for accuracy is (0.9412, 0.9936), indicating that the accuracy is statistically reliable
- High Information Rate and outperforms the baseline.

Contoso

RANDOM FORESTS (METHODS)

```
randomForest(formula = Class ~ ., data = train_data, nodesize = 10,
ntrees = 500, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3
      OOB estimate of  error rate: 2.09%
Confusion matrix:
          benign malignant
benign       303         8
malignant      2       166
          class.error
benign     0.02572347
malignant  0.01190476
```

- The random forest model achieved a high accuracy meaning it correctly predicted the class for 96% of the observations
- Oob error rate is 2.09%, suggesting that the model generalized well to unseen data, as the error on the out of bag samples is low

BOOTSTRAPPING (METHOD)

Bootstrap Statistics :

	original	bias	std. error
t1*	0.3499268	0.0002079063	0.01858928

2.5%	97.5%
0.3133236	0.3879941

- Small bias and the standard error indicate that the bootstrap resampling method produced stable and consistent results. This suggests that the bootstrap method is an appropriate choice for estimated variability and C in this context.

FINAL RESULTS AND DISCUSSION

RESULTS

- High accuracy across models, indicating strong predictive capabilities in distinguishing between classes.
- Bootstrap resampling of 5000 iterations showed a stable and robust estimate for the malignant class proportion.

DISCUSSION

- With the models effectively distinguishing between classes, they could assist health care professionals in prioritizing patients
- The bootstrap analysis and low error rates indicate that the models are stable and generalizable

ACKNOWLEDGMENTS AND LITERATURE CITED

- ACKNOWLEDGE: USED CHATGPT TO SELECT A DATASET, TO CHECK BOOSTING, BOOTSTRAPPING AND RANDOM FORESTS AND IN WRITING THE BULLET POINTS
- LITERATURE: USED PROFESSOR'S WORKS, ACTIVE TEXTBOOK, AND MY PREVIOUS HOMEWORKS TO CREATE CONFUSION MATRIX AND K- FOLD CROSS VALIDATION
- USED GOOGLE AND RSTUDIO HELP THINGY TO SEARCH UP SPECIFIC FUNCTIONS, ESPECIALLY K-FOLD CROSS VALIDATION, AND POST PROJECT ONE CODING.

R MARKDOWN FILE