Contoso

# BREAST CANCER DETECTION

BY: GHATTA TRIVEDI

DATE: NOVEMBER 14TH 2024

# ABSTRACT


- I used the Breast Cancer Wisconsin dataset for 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 an analysis of which model performed better in classifying breast cancer cases.


Contoso

# INTRODUCTION OF THE DATASET

## DATASET: BREAST CANCER WISCONSIN DATASET

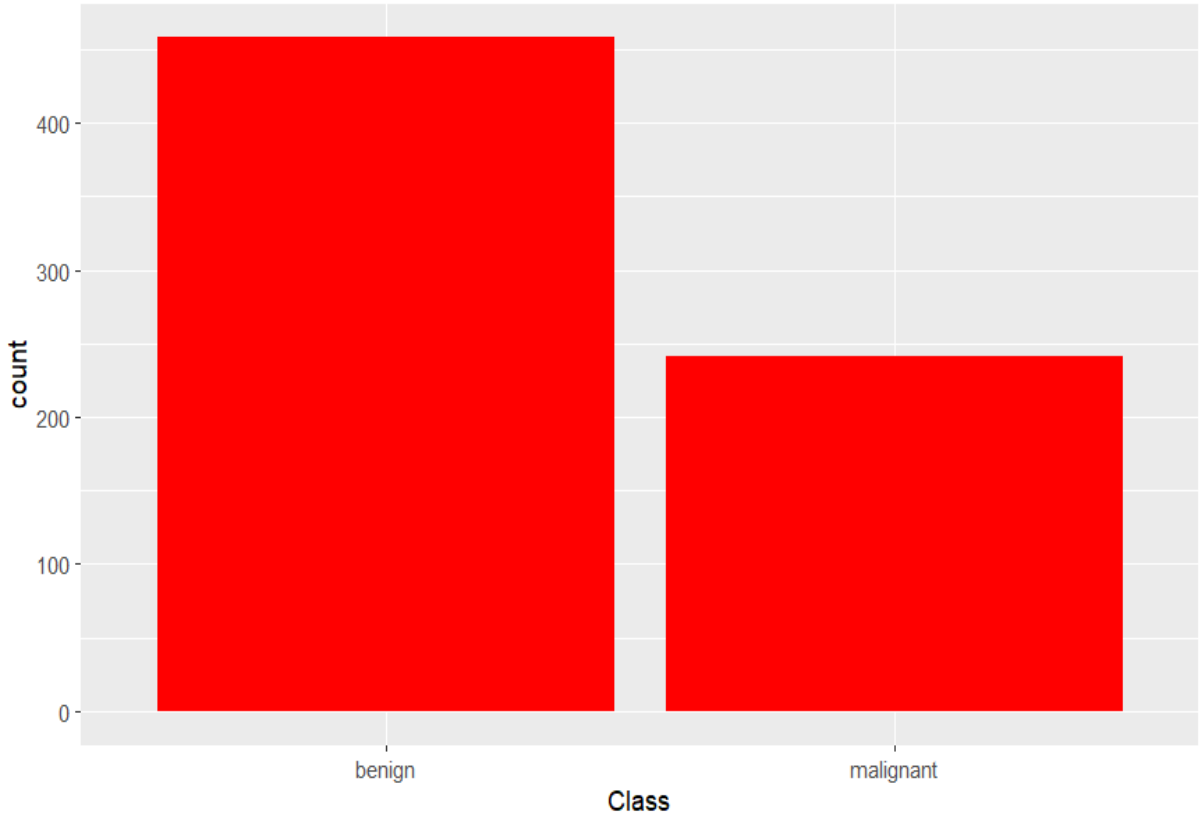
- Some of the features related to the tumor included in the dataset:
  - Clump Thickness
  - Uniformity of Cell Size
  - Uniformity of Cell Shape
  - Marginal Adhesion
  - Single Epithelial Cell Size

- 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

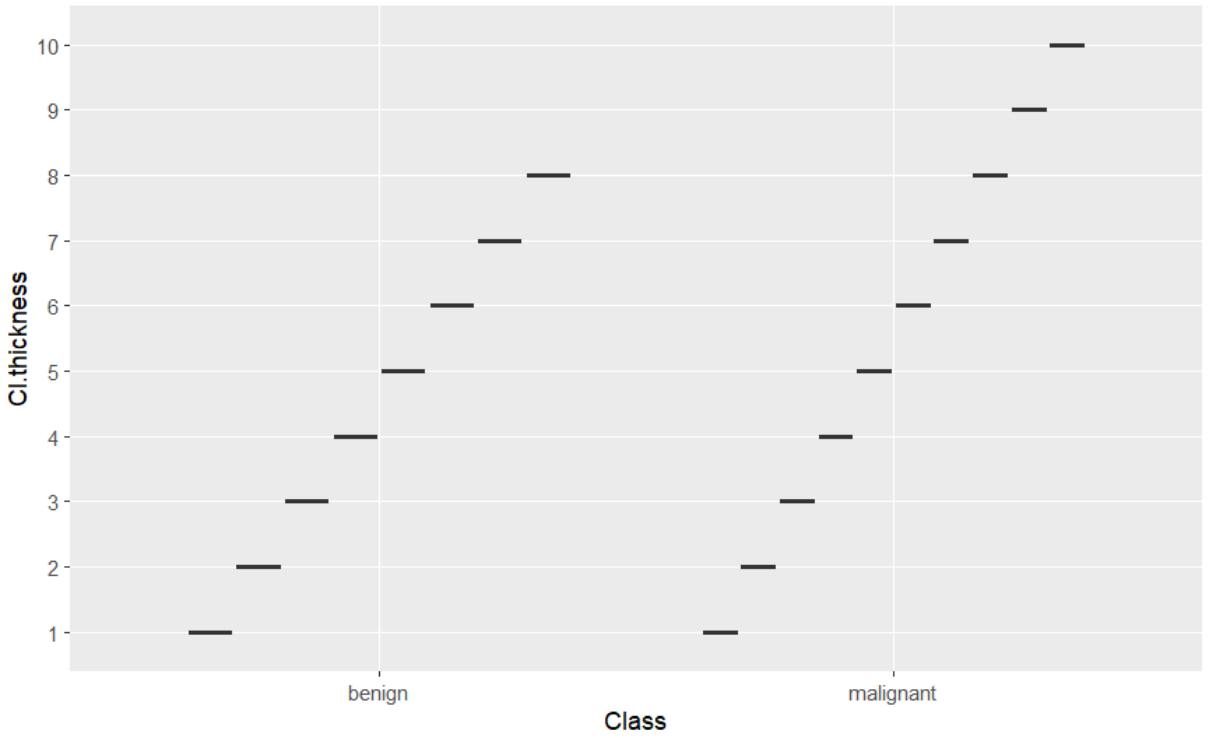
# DATA EXPLORATION AND VISUALIZATION (METHODS)

Bar Plot of Class vs. Count




The bar plot reveals that the data set is imbalanced.

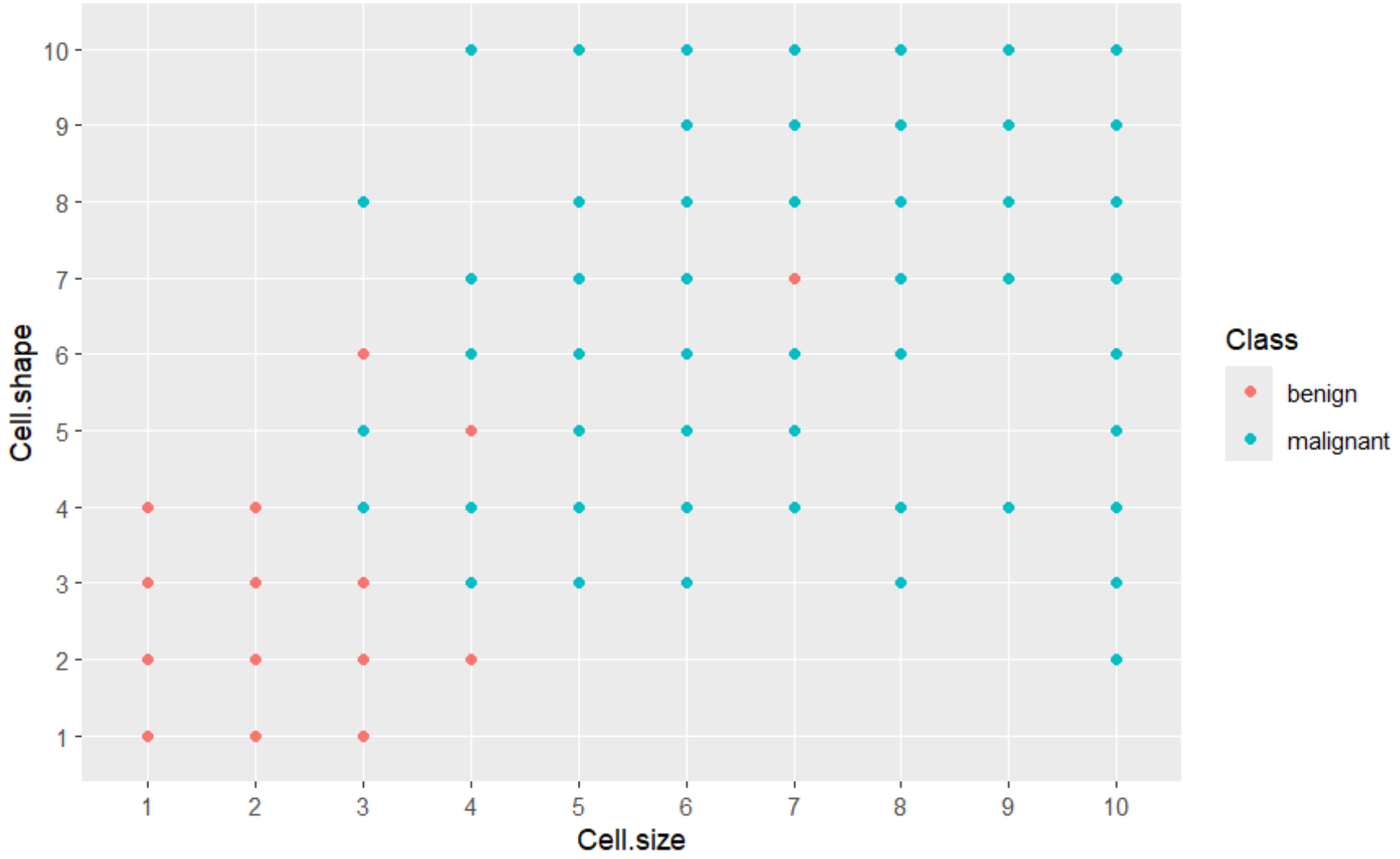
Box Plot of Clump Thickness vs. Class




The box plot shows that clump thickness may be used as a predictor.

# SCATTERPLOT OF CELL SIZE VS CELL SHAPE (METHOD)

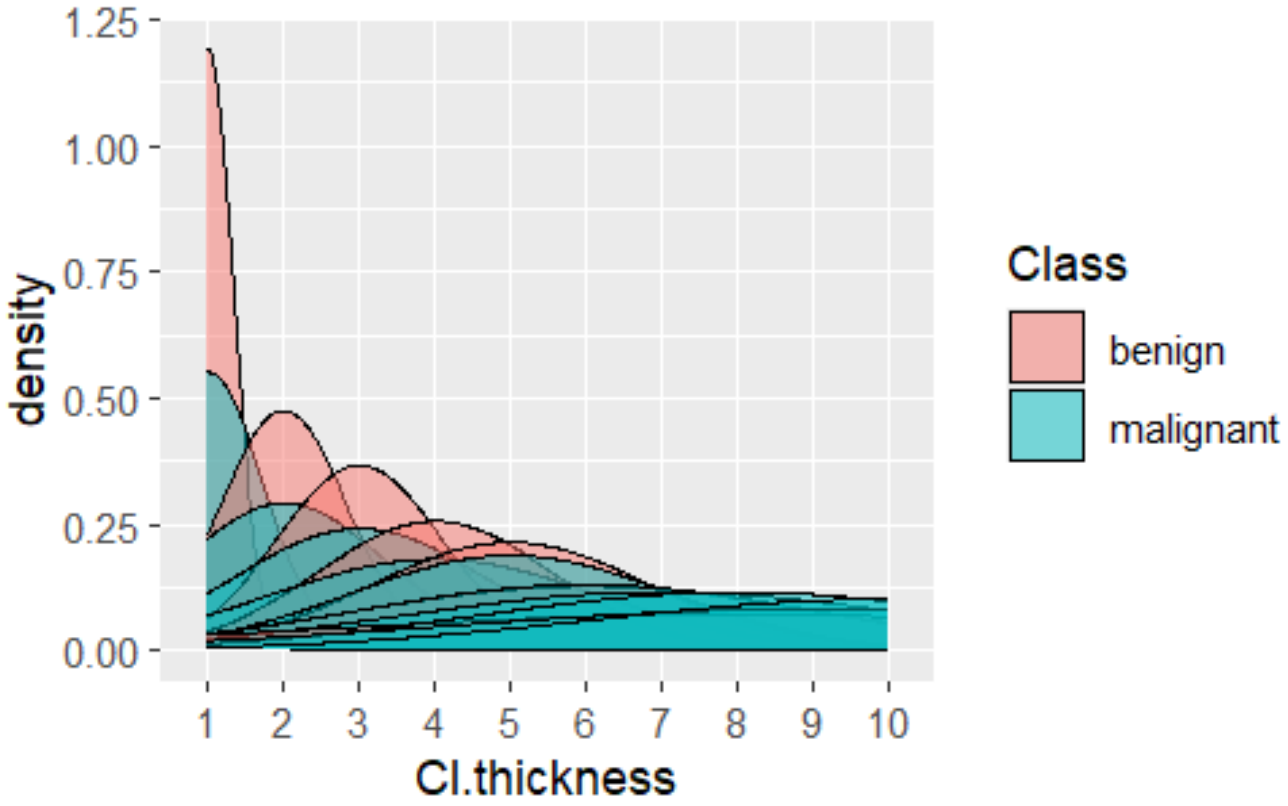
- Cell size and cell shape is good predictor for classification.
  - The malignant samples tend to be more spread out and appear in a separate region from the Benign samples.




Contoso

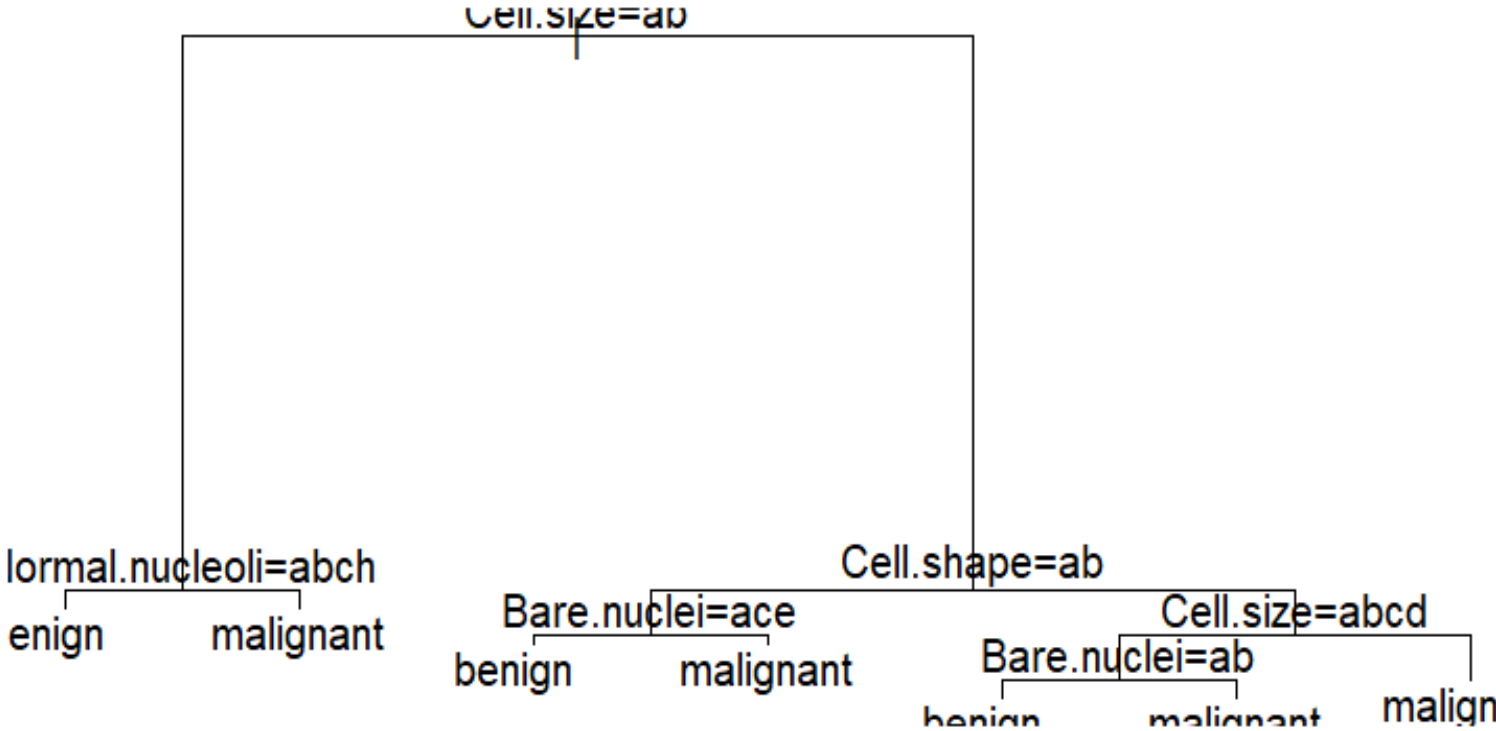
# DENSITY PLOT OF CLUMP THICKNESS (METHOD)

- 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.



# 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

# ACKNOWLEDGMENTS AND LITERATURE CITED


- ACKNOWLEDGE: USED CHATGPT TO SELECT A DATASET AND IN WRITING THE BULLET POINTS
- LITERATURE: USED PROFESSOR'S WORKS AND MY PREVIOUS HOMEWORKS TO CREATE CONFUSION MATRIX

# R MARKDOWN FILE