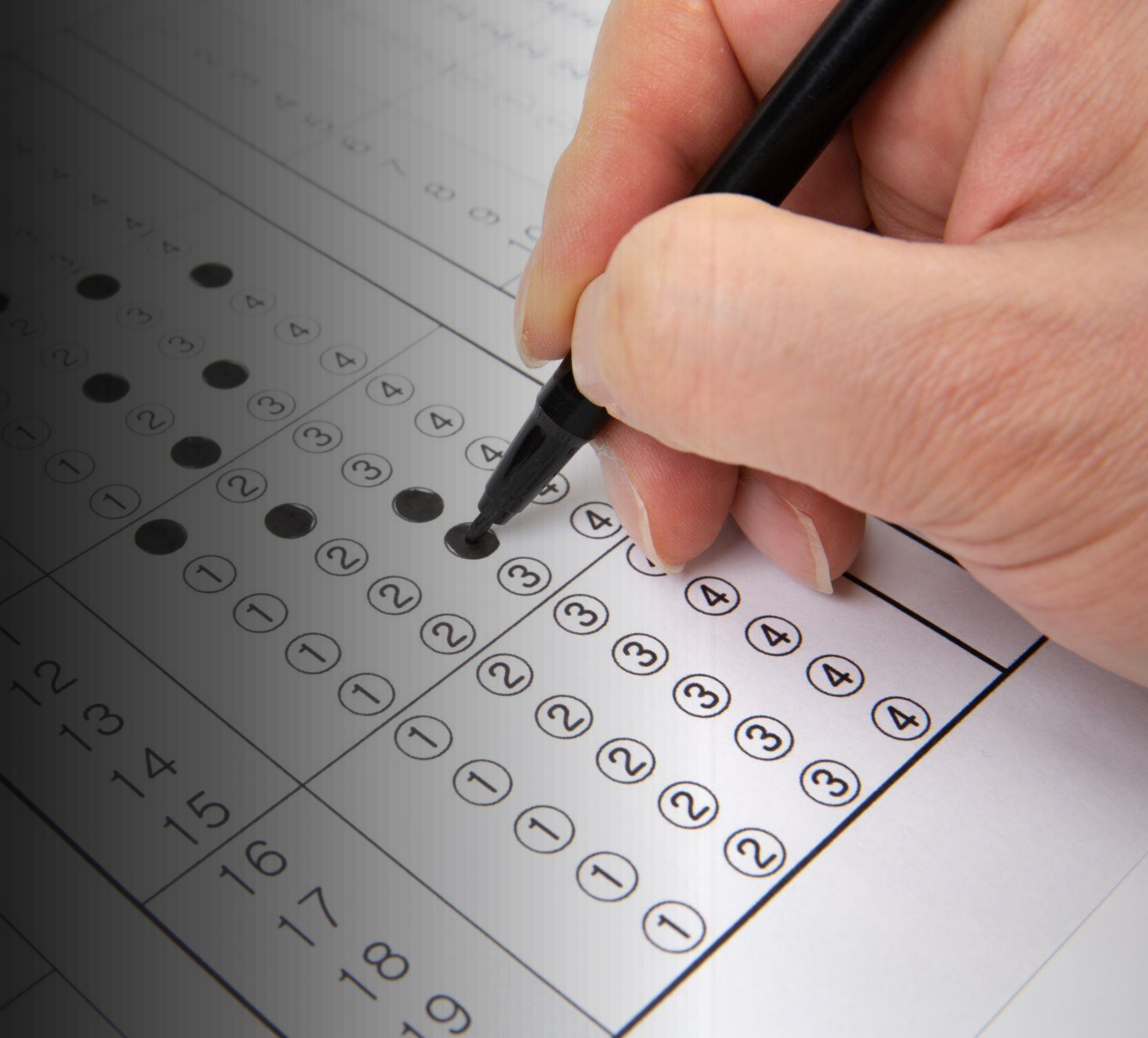
# Mammographic Prediction with BI-RADS Assessments



Milestone Project 3

Wisconsin Data Science and Analytics

Donald Taggart

January 15, 2023

## “The” Conversation

The ultrasound data predicted that you had a 96.58% probability of having malignant breast cancer and the biopsy confirmed it.

That’s an awful specific number Doc. How on earth did you arrive at that?

We take measurements during the ultrasound and use a statistical model with a more than 98% accuracy to predict the probability that a tumor is malignant. Basically, if the model says you have cancer, you have cancer.

# Mammographic Mass Data Set

M. Elter, R. Schulz-Wendtland and T. Wittenberg (2007)
The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process.
Medical Physics 34(11), pp. 4164-4172

UCI Machine Learning Repository

- Dataset
- Abstract
- Data Dictionary

- “The” Conversation
- Dataset
- **Question**
- Background
- Data
- EDA
- Modeling
- Results
- Conclusion

## Question

“Mammography is the most effective method for breast cancer screening available today. However, the low positive predictive value of breast biopsy resulting from mammogram interpretation leads to approximately 70% unnecessary biopsies with benign outcomes.”

Can you improve upon the 30% precision by applying a machine learning model to BI-RADS assessment data and reduce the number of unnecessary breast biopsies?

- “The” Conversation
- Dataset
- Question
- **Background**
- Data
- EDA
- Modeling
- Results
- Conclusion

## Background

- Breast Imaging Reporting and Data System (BI-RADS)
- Project started in the late 1980s
- American College of Radiologists
- Goal to develop a common lexicon
- Updated several times
- Covers the areas of
  - **Mammography**
  - Ultrasound
  - MRI

# Background

BI-RADS Assessment Categories

0 – Inconclusive

1 – Negative

2 – Benign

3 – Probably Benign (< 2%)

4 – Suspicious (50 – 95%)

5 – Highly Suggestive (> 95%)

6 – Known Biopsy-proven

- “The” Conversation
- Dataset
- Question
- Background
- Data
- EDA
- Modeling
- Results
- Conclusion

# Data

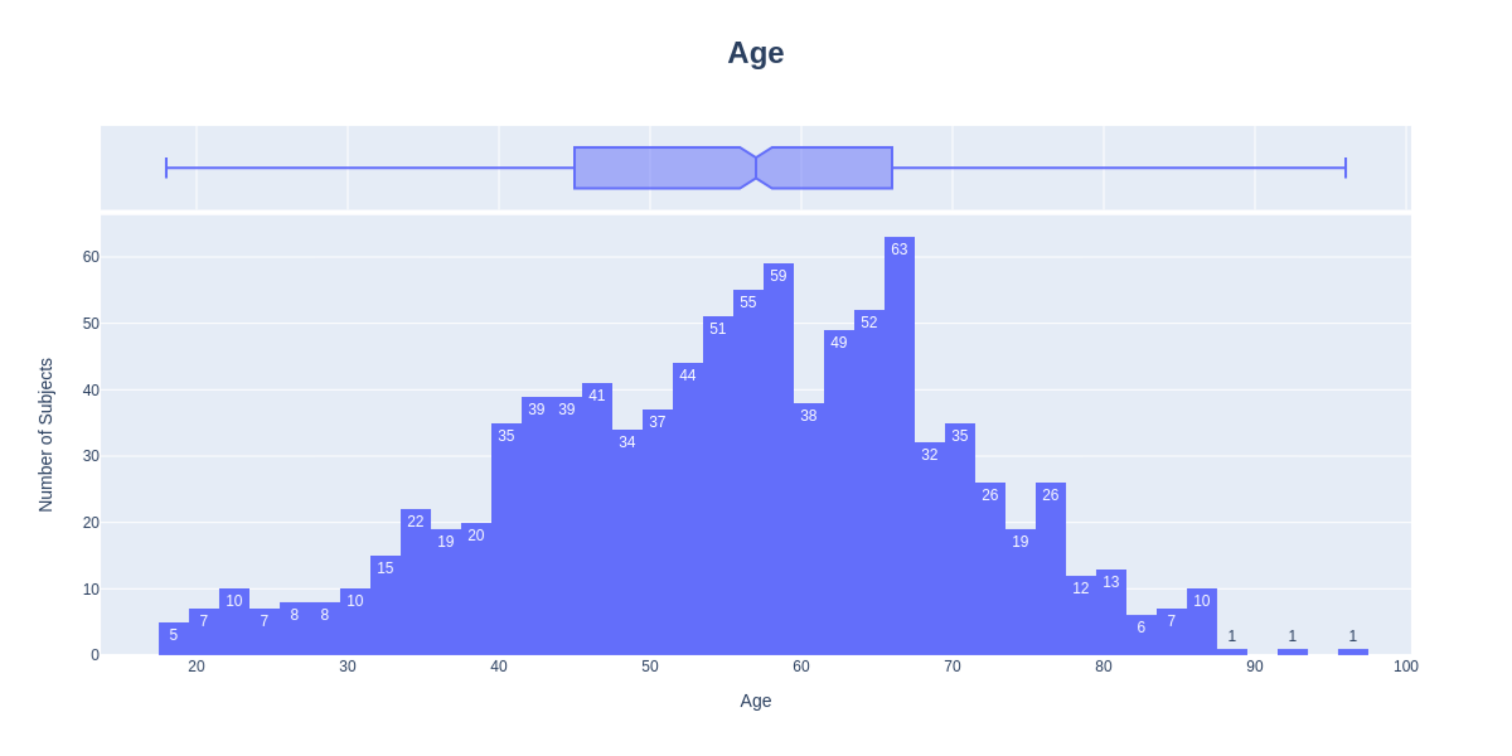
“This dataset can be used to predict the severity (benign or malignant) of a mammographic mass lesion from BI-RADS attributes and the patient's age.”

- BI-RADS Assessment (non-predictive)
- Patient’s Age
- Three BI-RADS Attributes
  - Shape
  - Margin
  - Density
- Ground Truth
  - 516 Benign
  - 445 Malignant

Institute of Radiology of the University Erlangen-Nuremberg between 2003 and 2006

- “The” Conversation
- Dataset
- Question
- Background
- Data
- **EDA**
- Modeling
- Results
- Conclusion

| | |
|---|---|
| Count | 956 |
| Mean | 55.49 |
| Std | 14.48 |
| Min | 18 |
| 25% | 45 |
| 50% | 57 |
| 75% | 66 |
| Max | 96 |



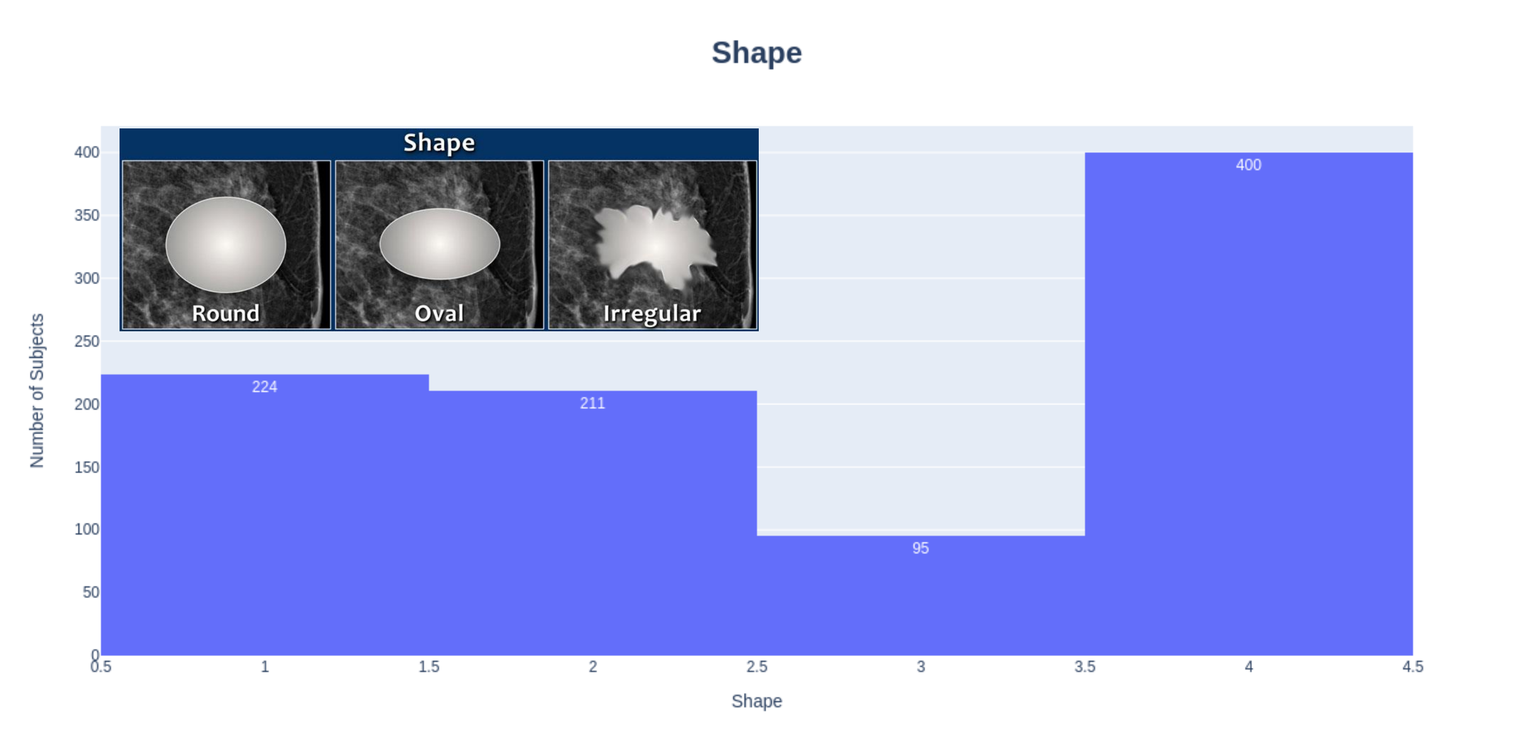

Data Wrangling

- Type: Integer
- Null Values: 5 imputed to the mean
- Standardized

- "The" Conversation
- Dataset
- Question
- Background
- Data
- **EDA**
- Modeling
- Results
- Conclusion

1. Round
2. Oval
3. Lobular
4. Irregular

Data Wrangling
- Type: Nominal
- Null Values: 31
- One-hot encoded



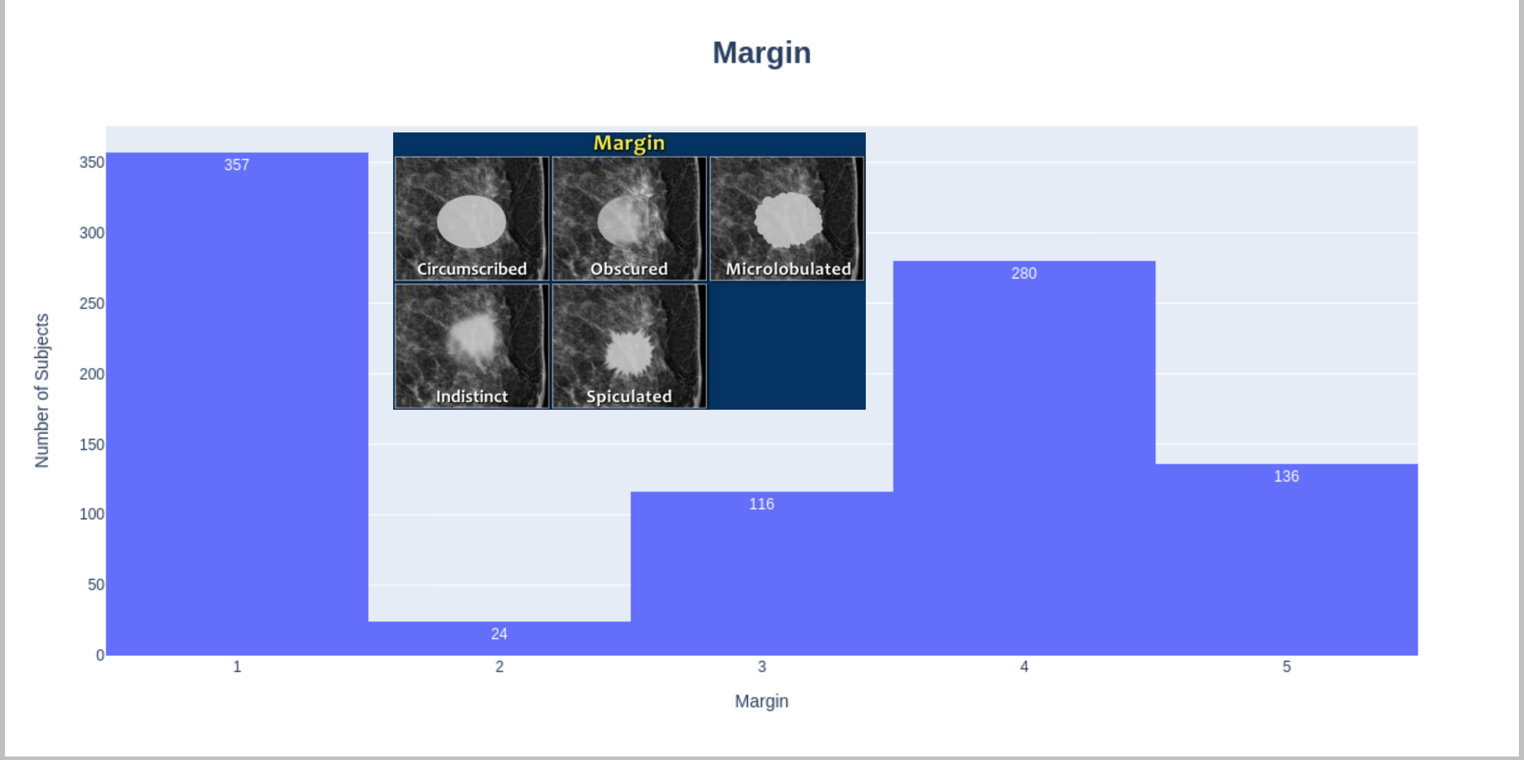

Radiology Assistant: Accessed 1-13-2023

- “The” Conversation
- Dataset
- Question
- Background
- Data
- **EDA**
- Modeling
- Results
- Conclusion

1. Circumscribed
2. Microlobulated
3. Obscured
4. Indistinct
5. Spiculated

Data Wrangling

- Type: Nominal
- Null Values: 48
- One-hot encoded



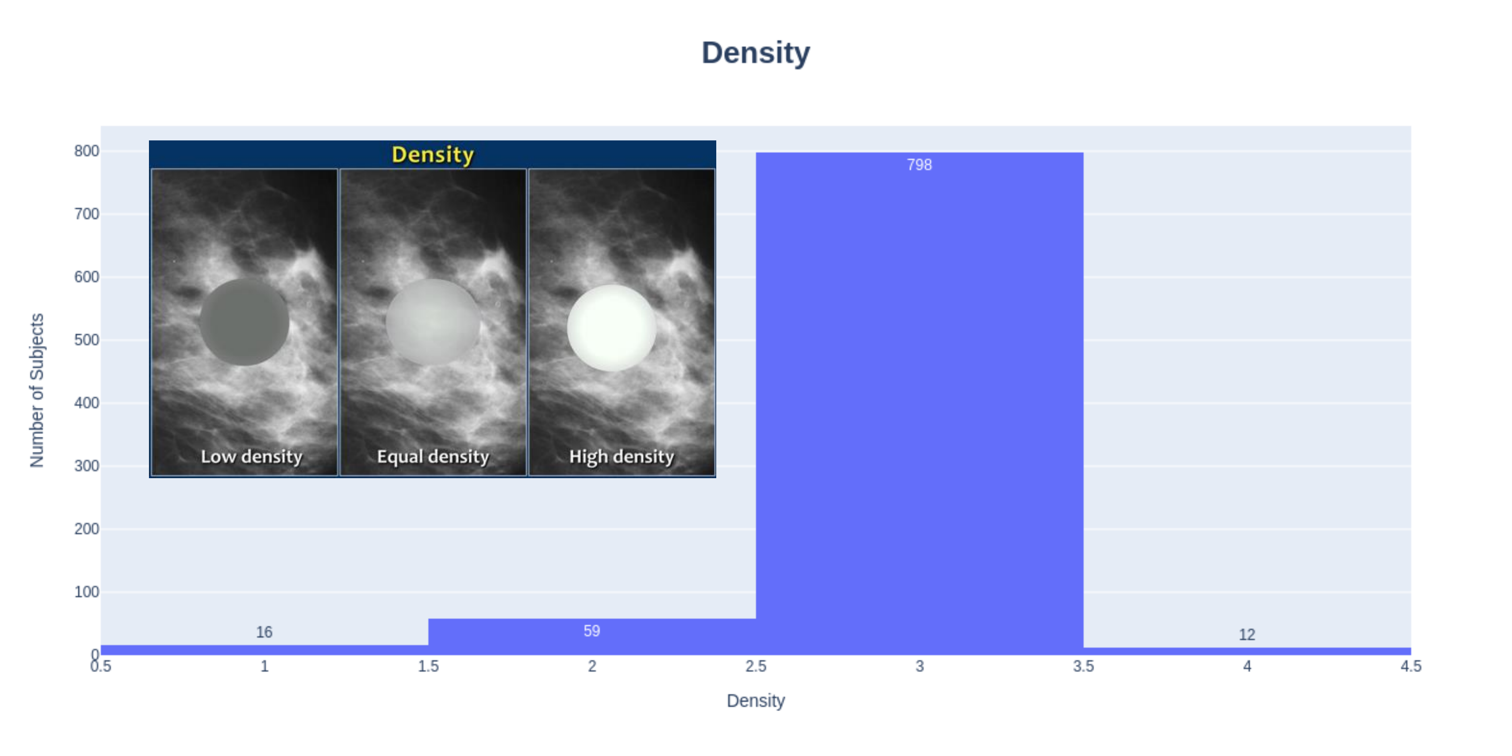

Radiology Assistant: Accessed 1-13-2023

- “The” Conversation
- Dataset
- Question
- Background
- Data
- **EDA**
- Modeling
- Results
- Conclusion

# Density

1. High
2. Equal
3. Low
4. Fat-containing

Data Wrangling
- Type: Ordinal
- Null Values: 76 imputed to the mean
- Standardized




Radiology Assistant: Accessed 1-13-2023

# Modeling

Binary classification problem

Models tested

- Logistic Regression
- K-Nearest Neighbors
- Support Vector Machine
- Decision Tree
- Random Forest
- AdaBoost
- Gradient Boost

Train-Test split

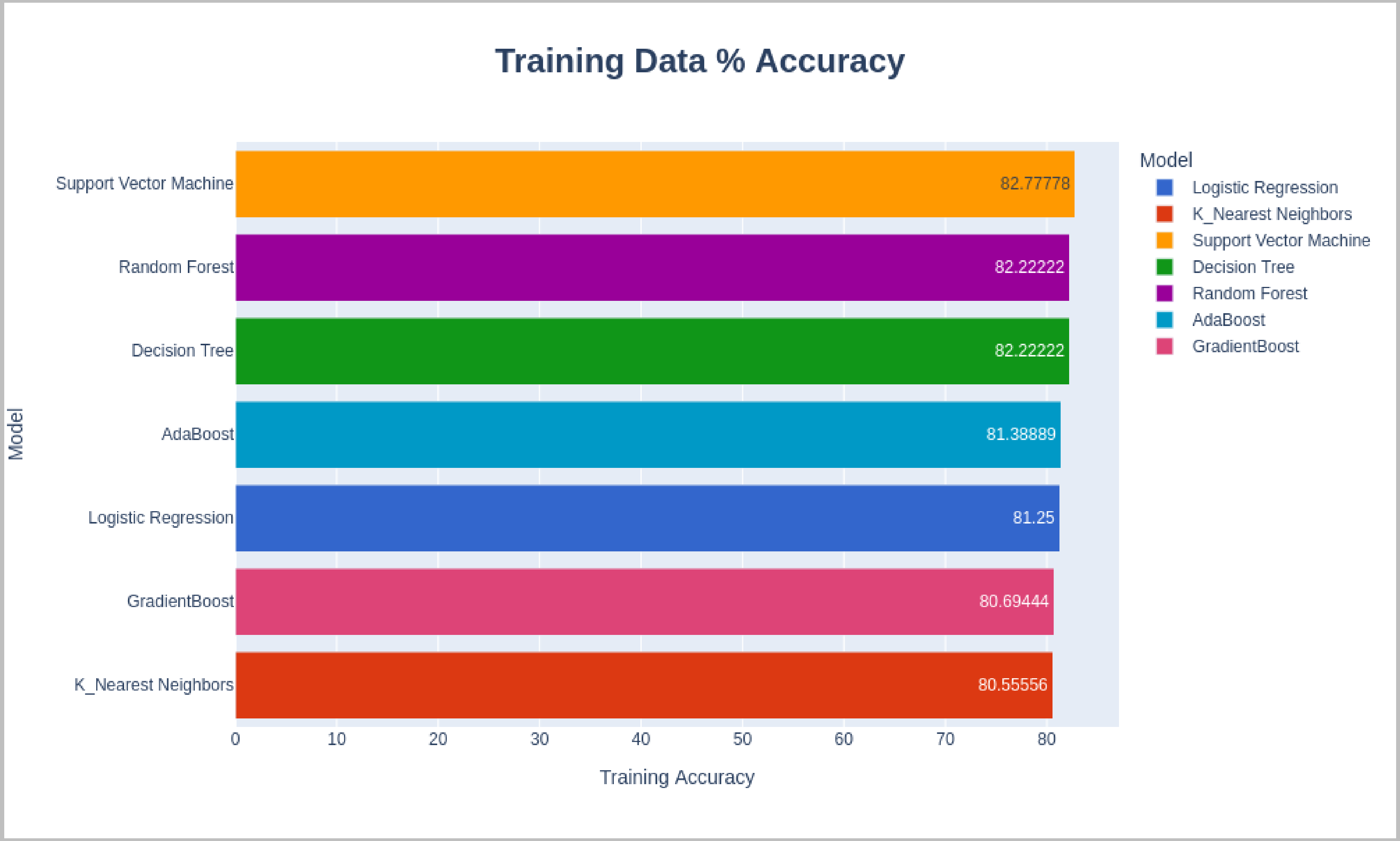
- 75% / 25%
- Stratified

Tuning

- GridSearchCV hyperparameter tuning
- 10-fold cross validation

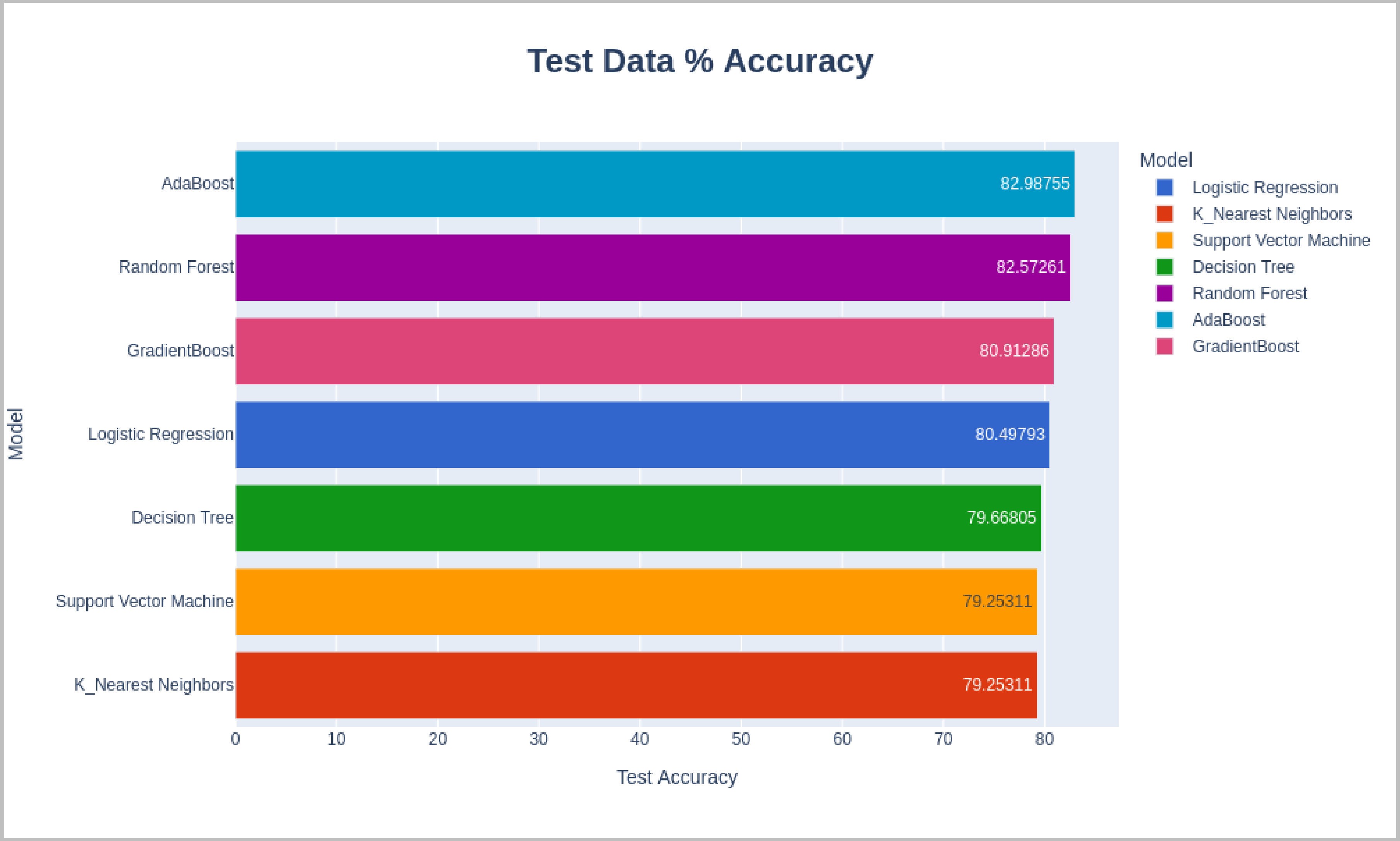
- “The” Conversation
- Dataset
- Question
- Background
- Data
- EDA
- Modeling
- **Results**
- Conclusion

# Results



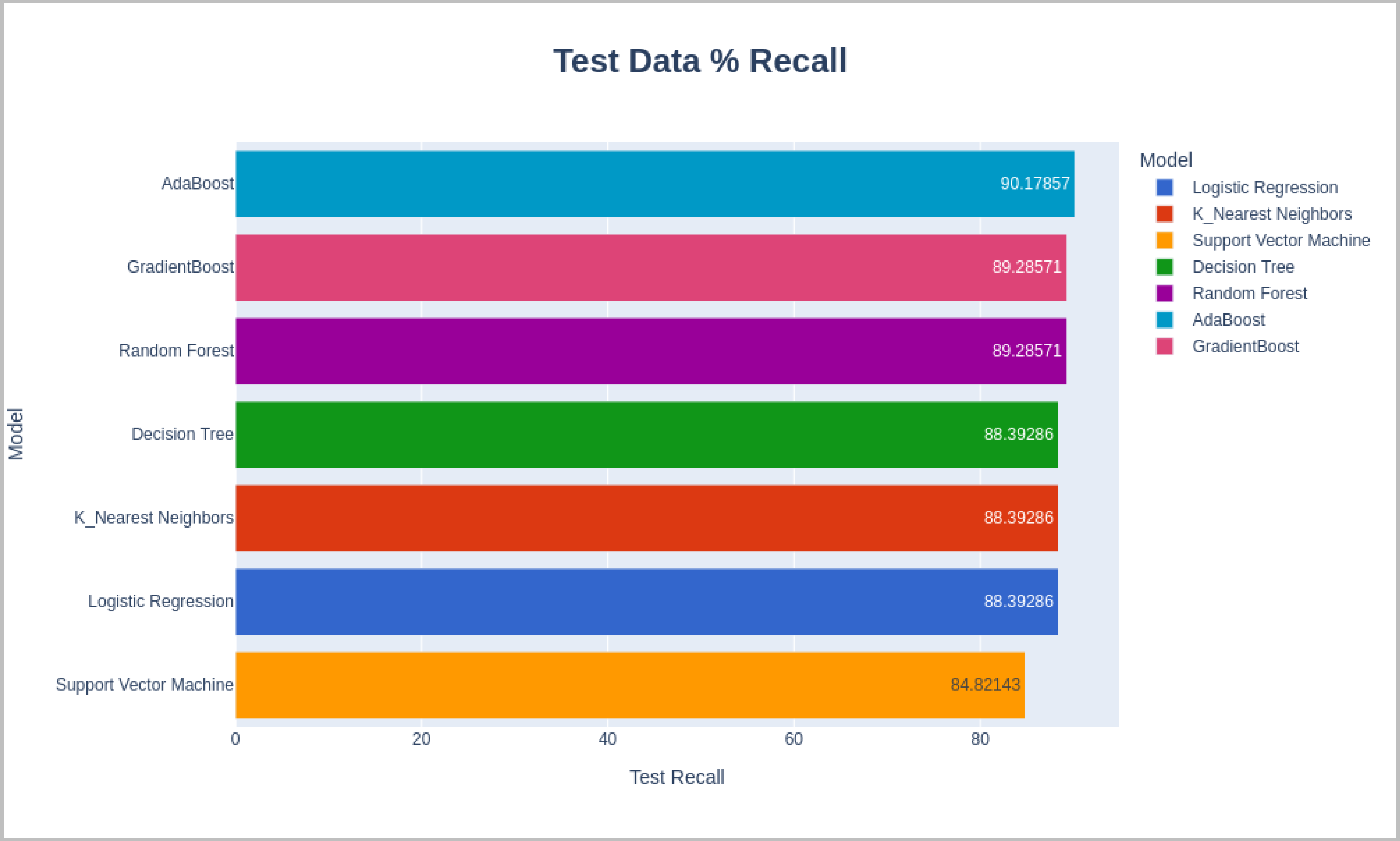
- “The” Conversation
- Dataset
- Question
- Background
- Data
- EDA
- Modeling
- **Results**
- Conclusion

# Results



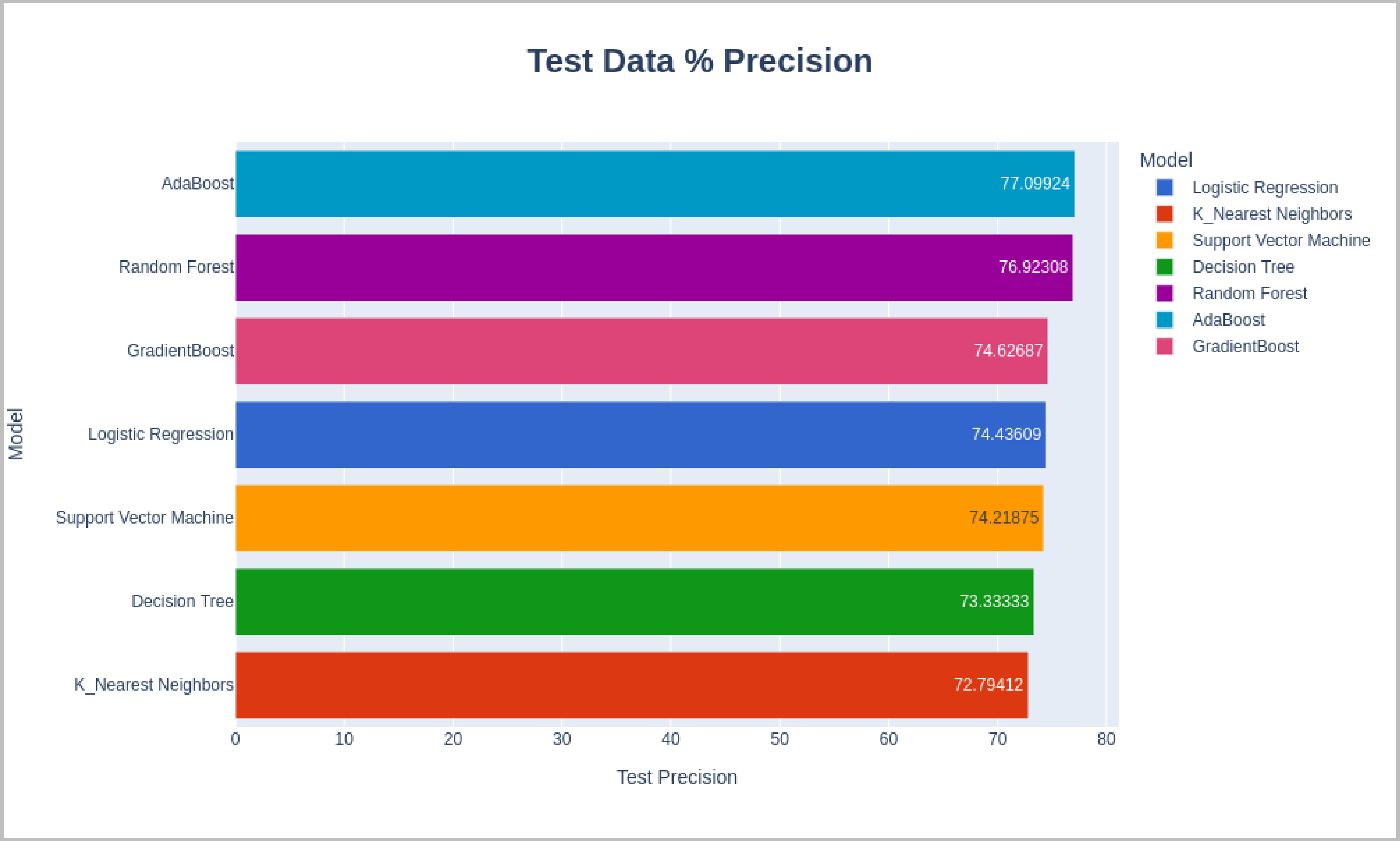
- “The” Conversation
- Dataset
- Question
- Background
- Data
- EDA
- Modeling
- **Results**
- Conclusion

# Results



- “The” Conversation
- Dataset
- Question
- Background
- Data
- EDA
- Modeling
- **Results**
- Conclusion

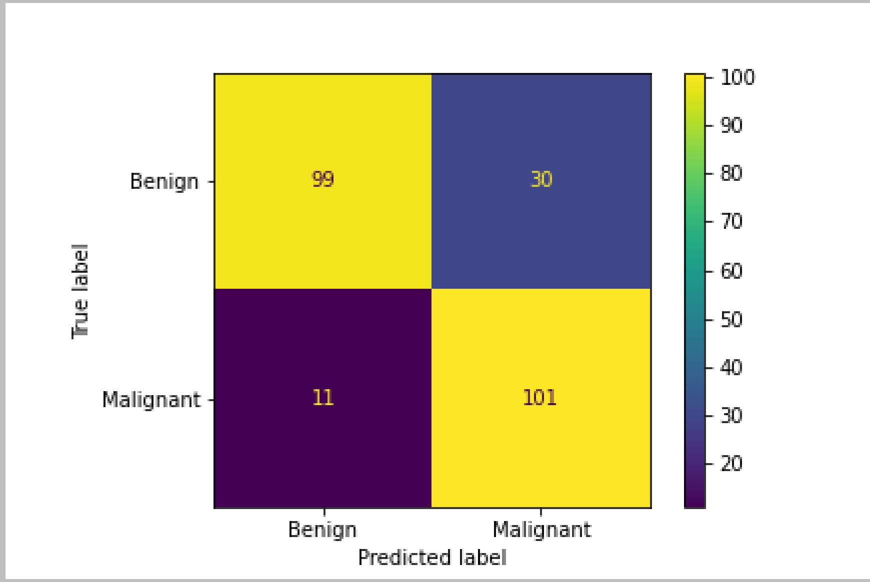
# Results

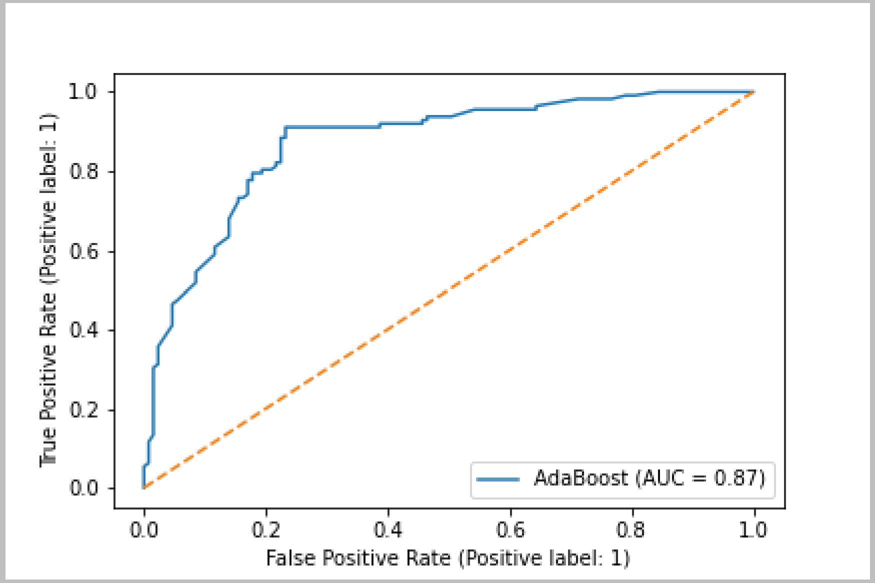


- "The" Conversation
- Dataset
- Question
- Background
- Data
- EDA
- Modeling
- **Results**
- Conclusion

# Results

| Model | AdaBoost |
|---|---|
| Accuracy | 83.0% |
| Recall | 90.2% |
| Precision | 77.1% |
| AUC | 0.87 |



- “The” Conversation
- Dataset
- Question
- Background
- Data
- EDA
- Modeling
- Results
- **Conclusion**

# Conclusion

Can you improve upon the 30% precision by applying a machine learning model to BI-RADS assessment data and reduce the number of unnecessary breast biopsies?

Yes!  By applying an AdaBoost machine learning model to BI-RADS assessment data, you can achieve a 77% precision with a 90% recall.

## Final Note

92% 5-year survival probability

We are 4.5 years in, and all is well.