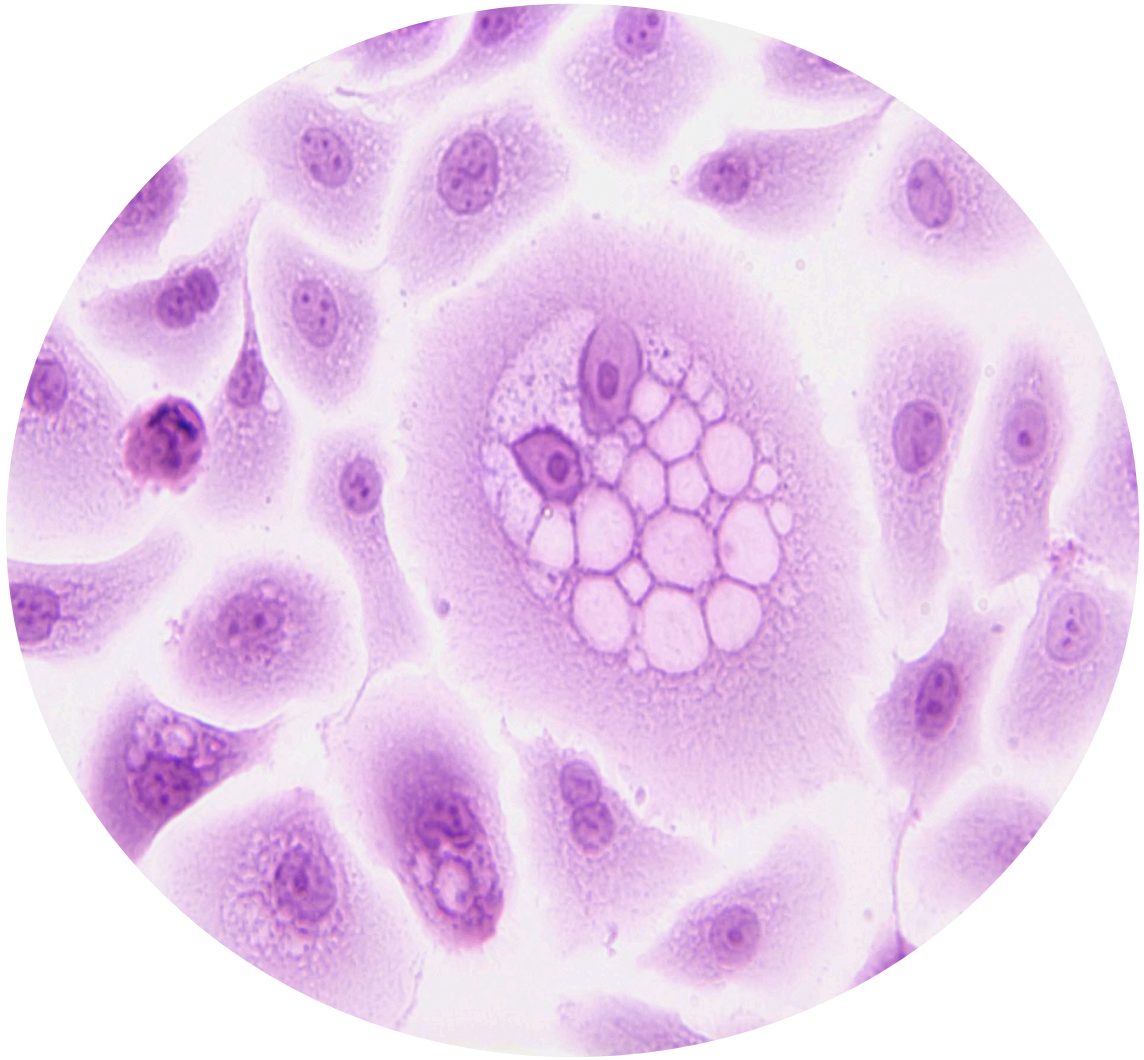
Deloitte.



Predicting Malignant Breast Tumors – Supervised Learning Model

AI Academy Apprenticeship Capstone Project – Rachel Day

Agenda

1 Project Goals

2 Data Understanding

3 Methods

4 Results

Background and Purpose

Background

Breast imaging techniques can detect suspicious areas, but can't tell a patient whether they have cancer or not*.

An estimated 297,790 women are expected to be diagnosed with breast cancer in 2023**.

Goal

What if there was a way, through existing imaging techniques, to get tumor measurements and be able to predict if it was malignant or benign?

Goal: Create a model that accurately predicts malignant tumors given existing measurements.

*Can a Radiologist Diagnose Breast Cancer from Imaging Tests Alone? (healthline.com)
**Breast Cancer Statistics | How Common Is Breast Cancer?

The Data

Pulled from Kaggle.com, an AirBnb for data scientists.

Kaggle is a crowd-sourced platform to attract, nurture, train and challenge data scientists.

The Dataset: Contains measurements and **characteristics from breast tumors** coming out of the **State of Wisconsin.**

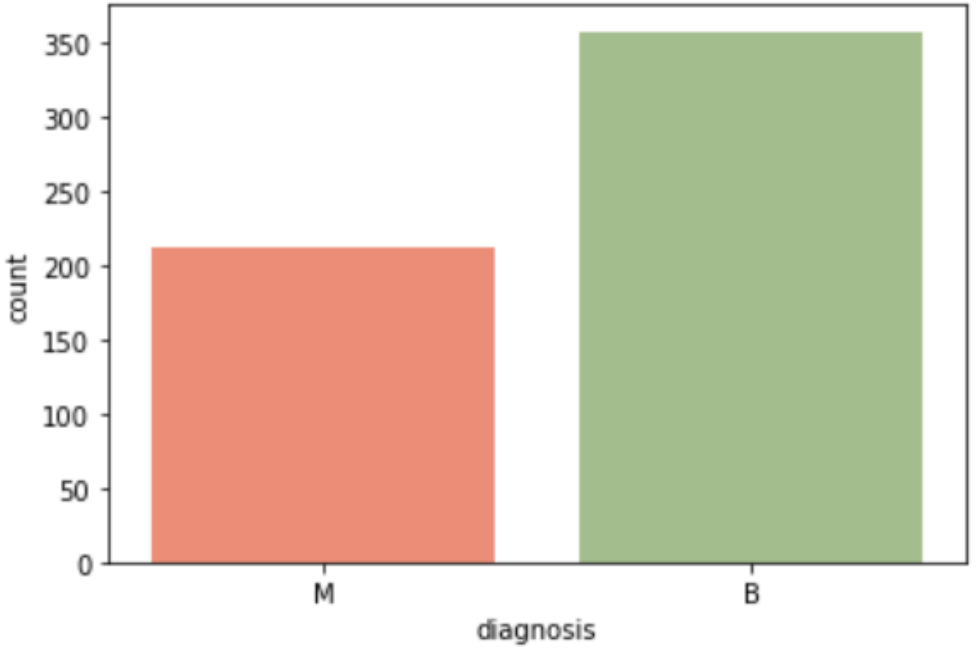
The Data

33 Columns
569 Rows

Containing measurements of tumors: radius, texture, compactness, etc.

357 Benign
212 Malignant

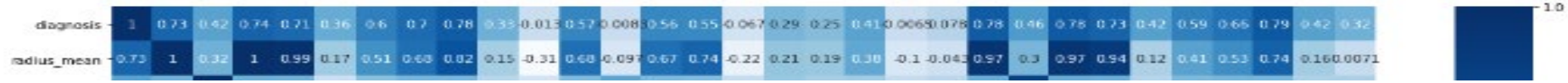
Classified as B-Benign, or M-Malignant



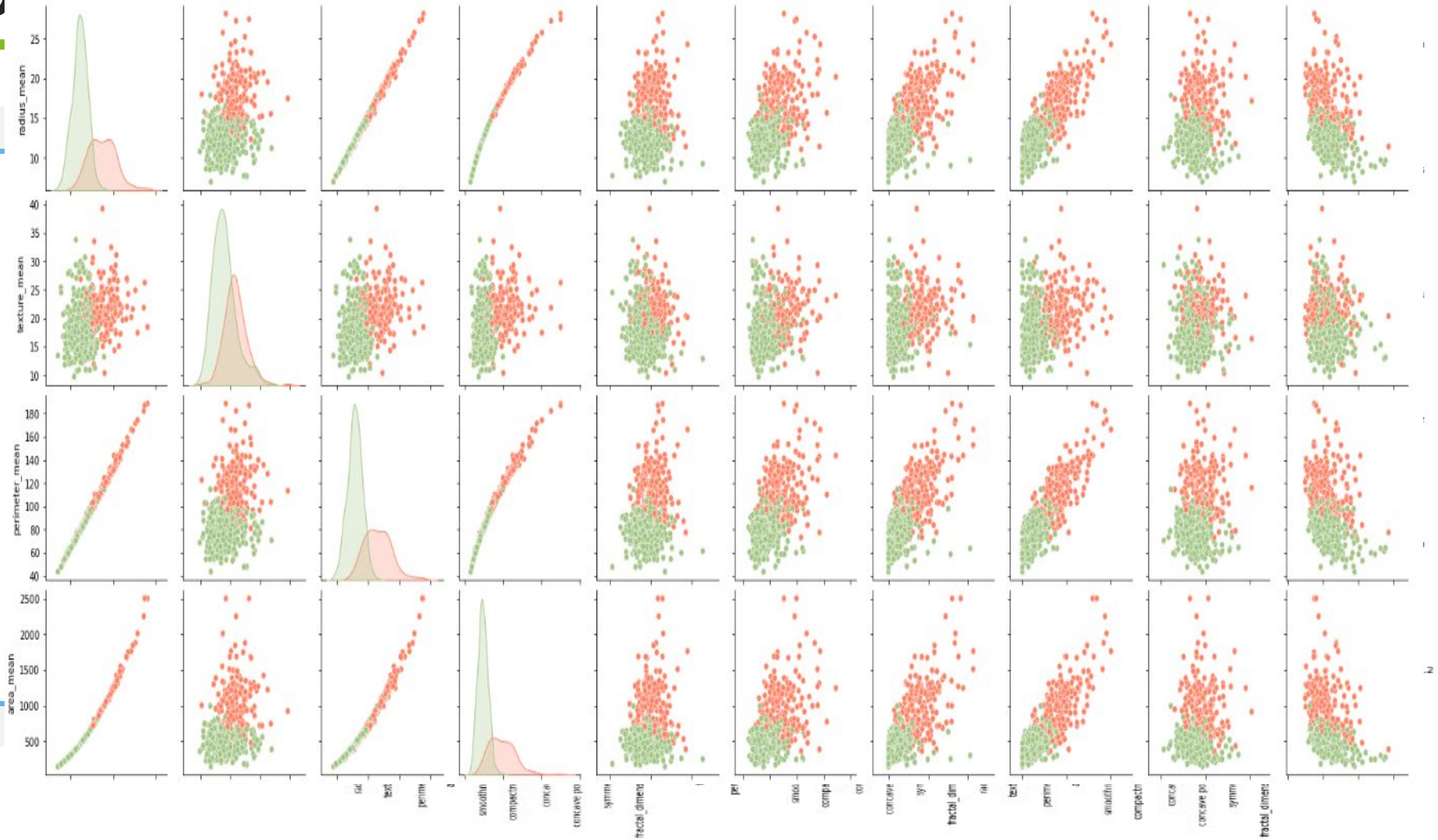
symmetry_worst	fractal_dimension_worst
0.4601	0.11890
0.2750	0.08902

diagnosis	radius_mean	texture_mean	perimeter_mean
M	17.99	10.38	122.80
M	20.57	17.77	132.90

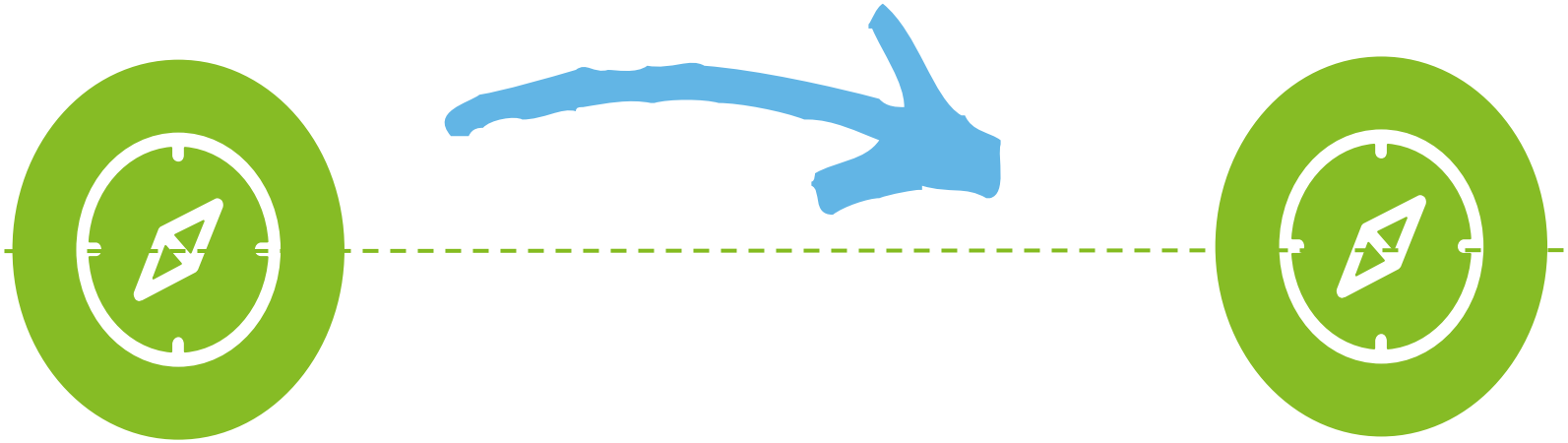
area_mean	smoothness_mean	compactness_mean	concavity_mean
1001.0	0.11840	0.27760	0.3001
1326.0	0.08474	0.07864	0.0869


diagnosis
radius_mean

D


radius_mean
texture_mean
perimeter_mean
area_mean

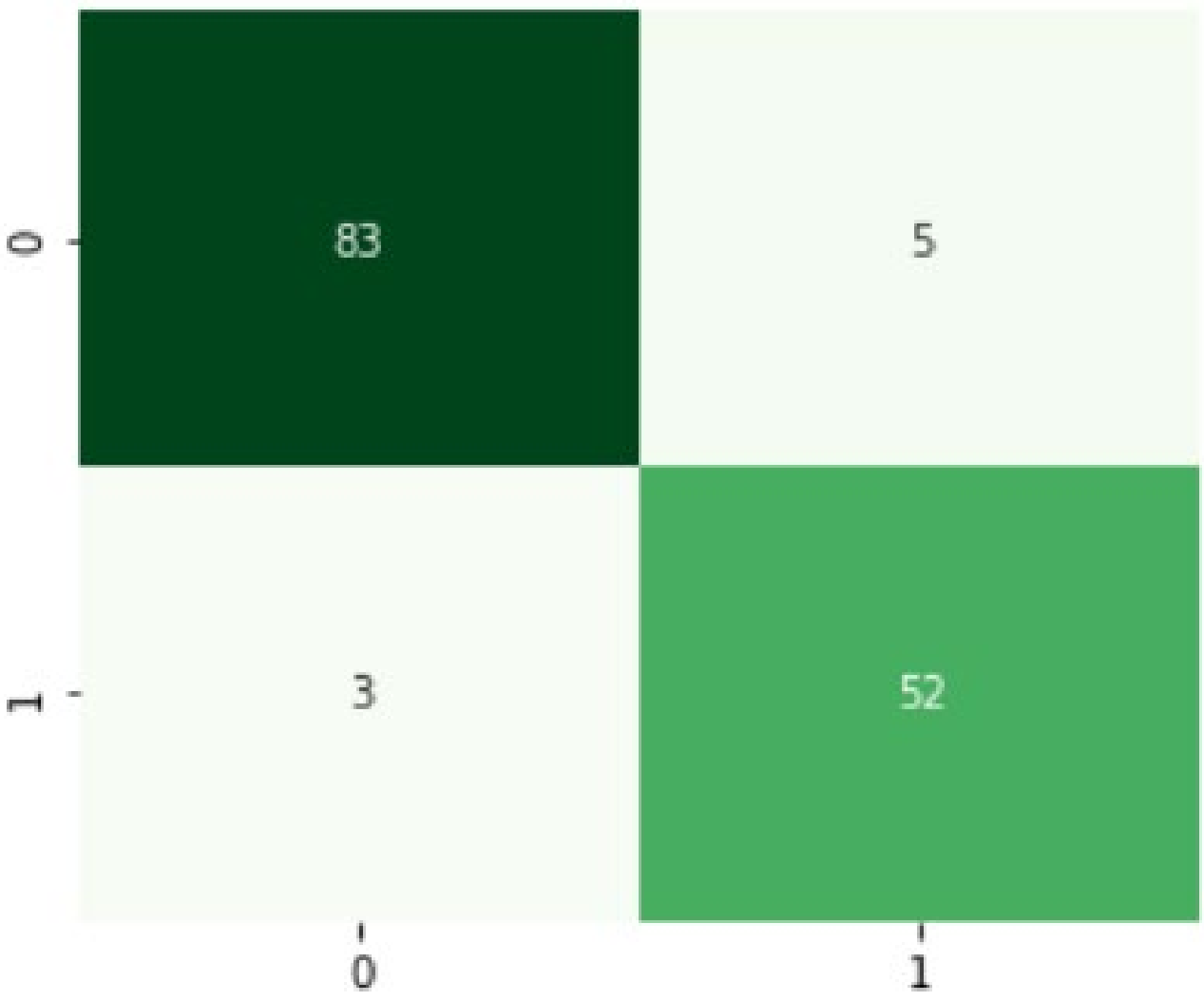
Pre-processing



Split Dataset into Training and Testing Subsets

Scale the Data

Modeling



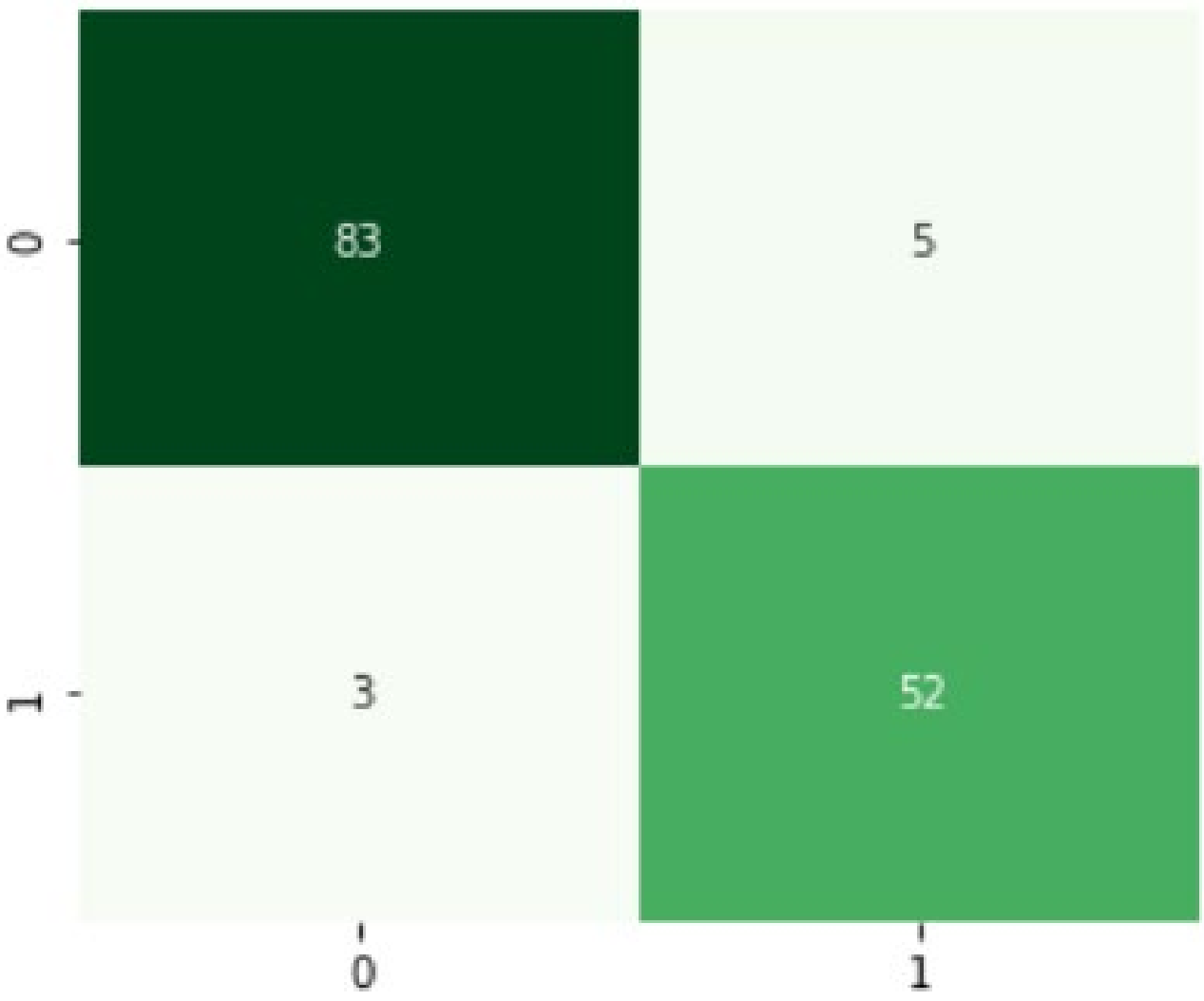
Baseline Model

Results:

95% Recall

94% Accuracy

Modeling



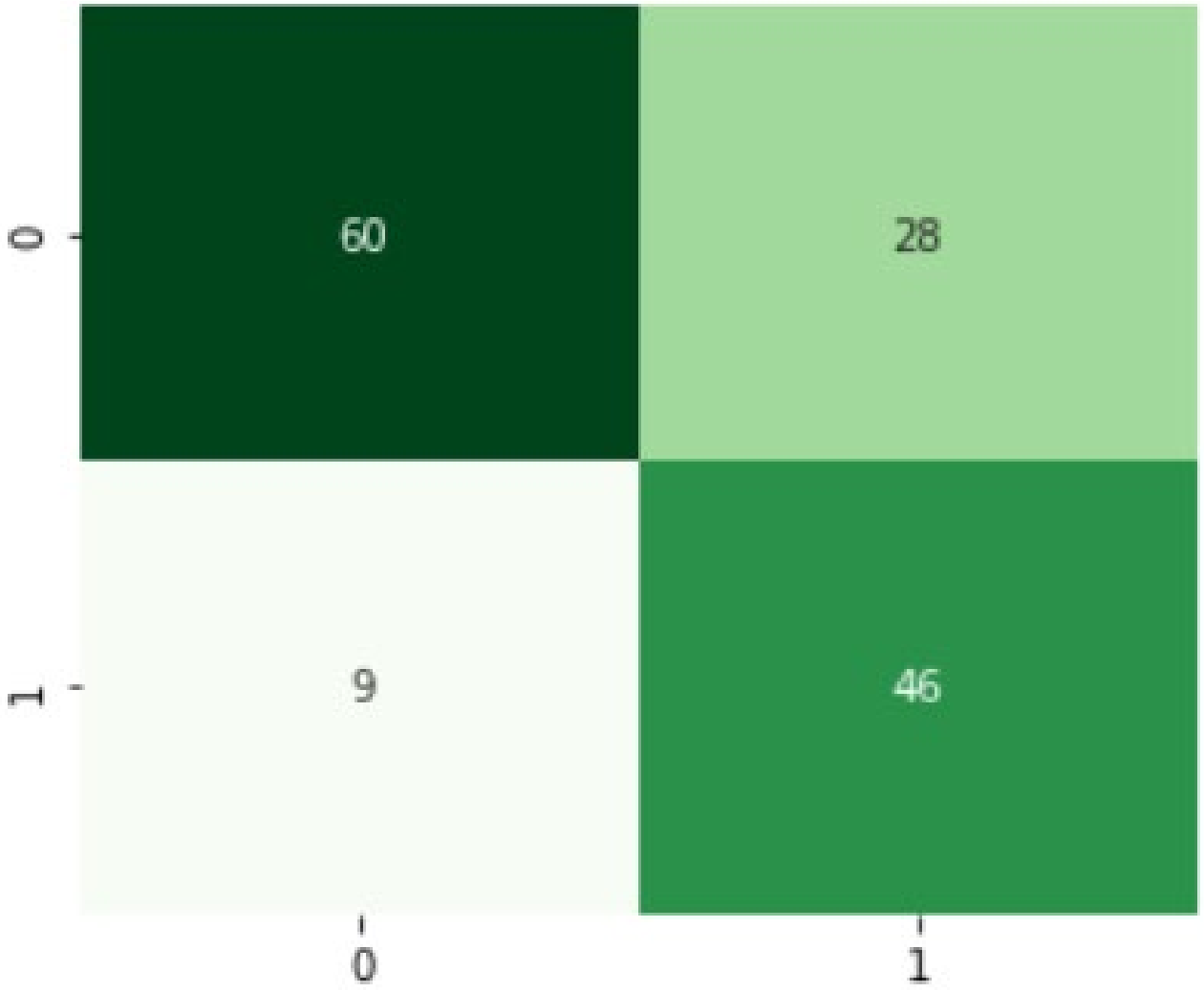
Tuned Logistic Regression Model

Tuned with a GridSearch

Results:
95% Recall

94% Accuracy

Modeling



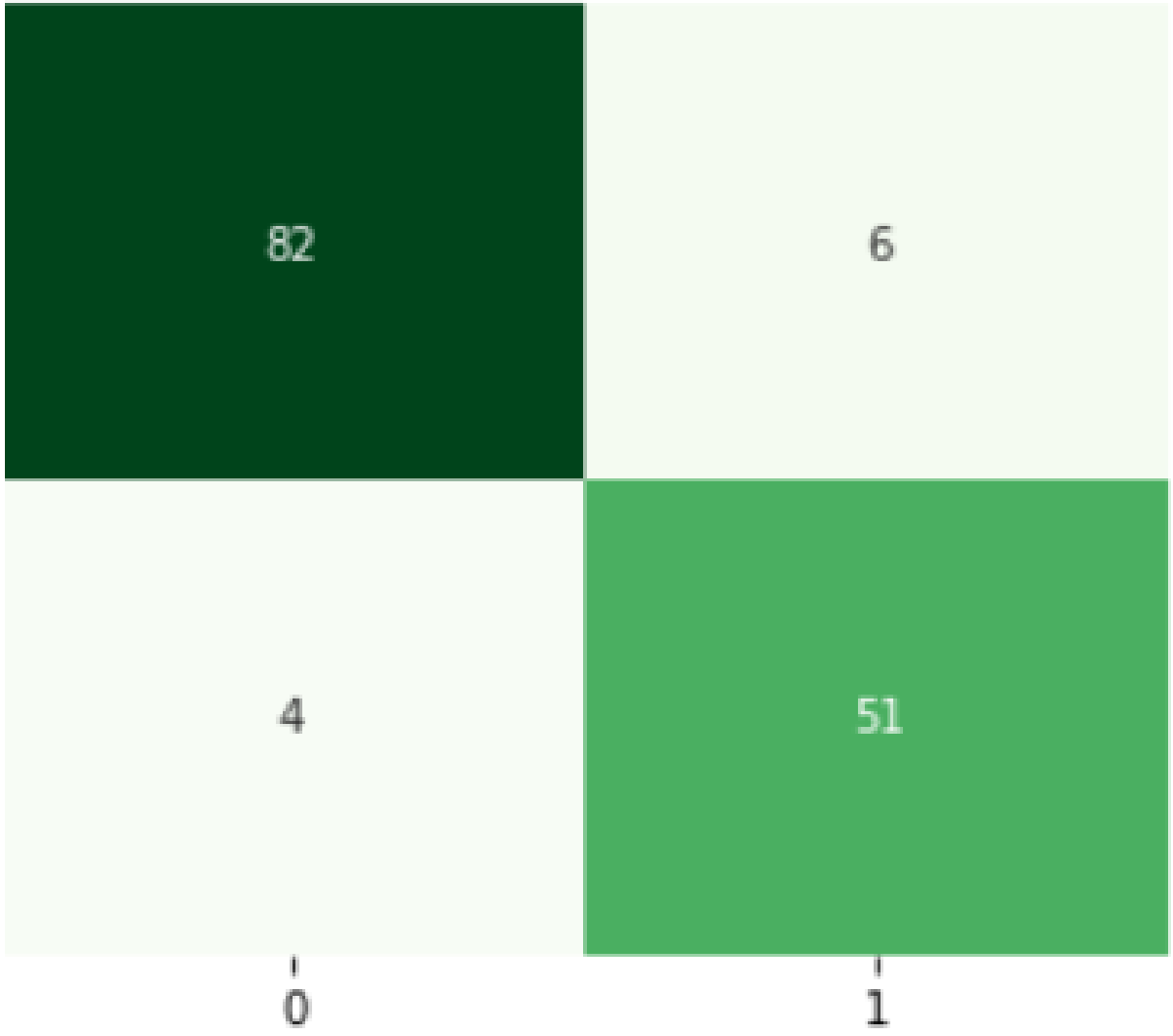
Tuned Decision Tree Model

Tuned with a GridSearch

83.6% Recall

74% Accuracy

Modeling

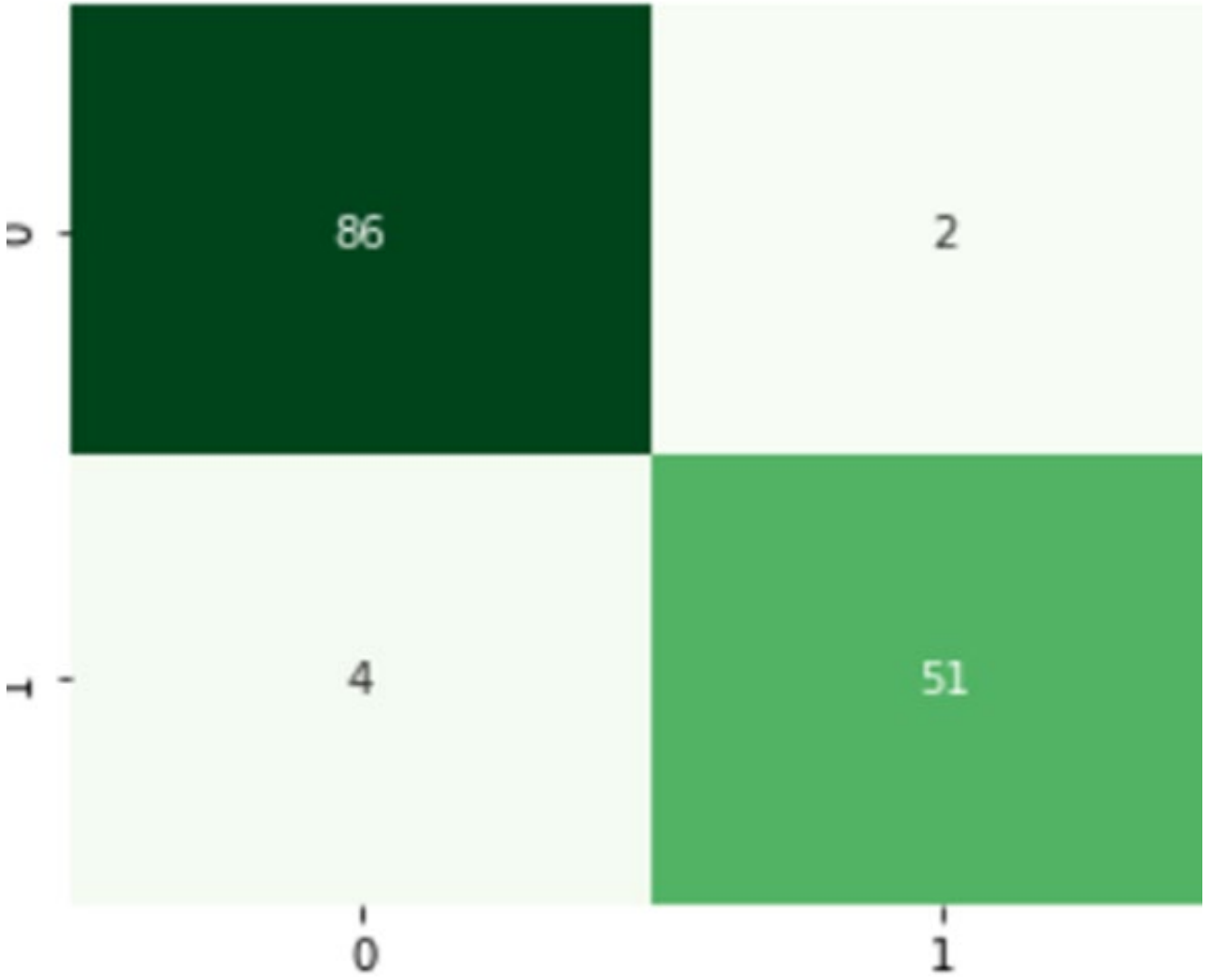


Decision Tree Model

Results:
92.7% Recall

93% Accuracy

Modeling

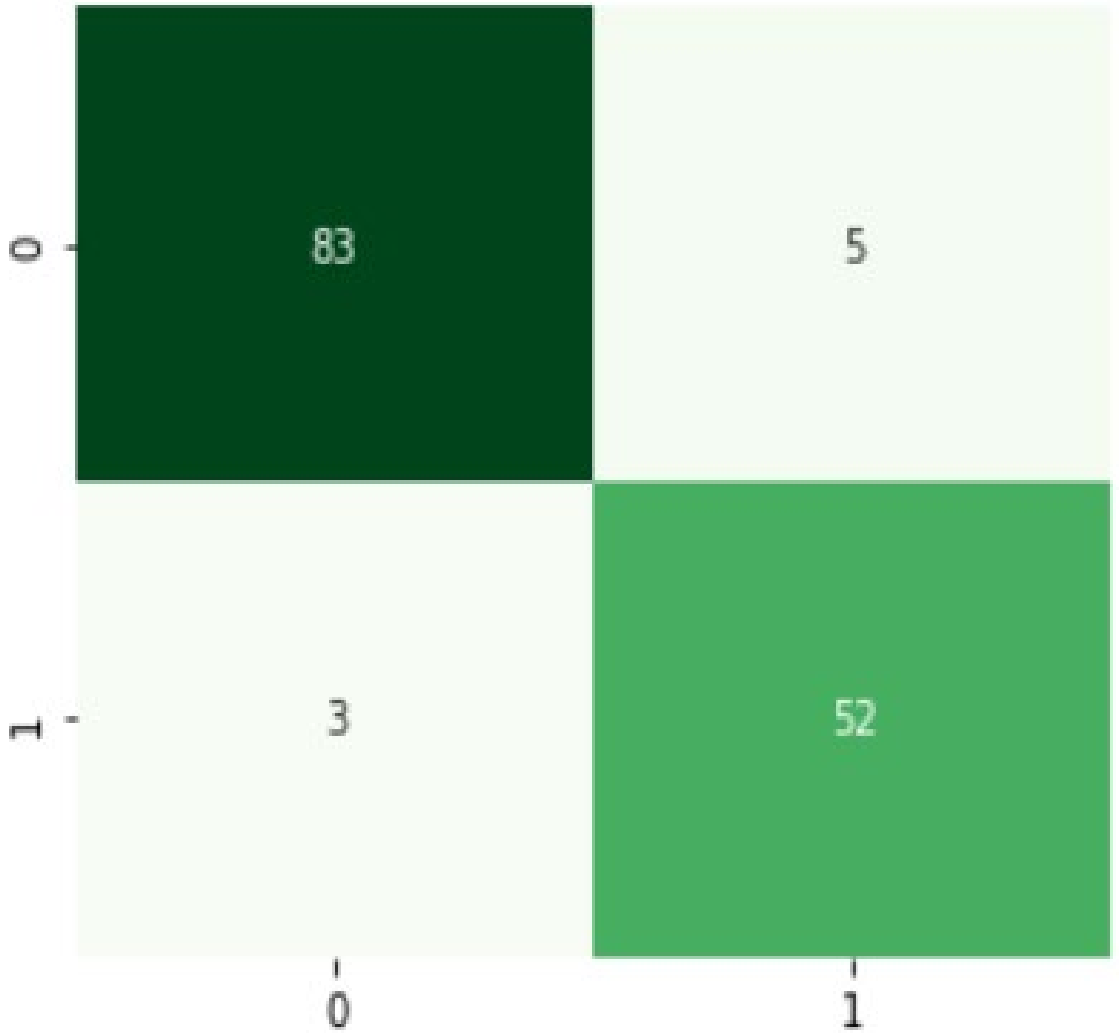


Random Forest Model

Results:
92.7% Recall

93% Accuracy

Optimal Model - Logistic Regression Model



Results

Out of 100 tumors observed, approximately 2 malignant tumors will be incorrectly classified as benign.

95% recall, 94% accuracy

Next Steps: tune so that false negatives are minimized.

Thank You!

Questions?