Breast Cancer Prediction

By: Steve Kim

Purpose

- The goal of this project is to interpret the data provided from Kaggle
- Hone my skills on everything I have learned the past few months (Data cleaning, machine learning, streamlit, etc…)
- Creating a prediction model for car prices
- This model can be used for sellers or buyers in Australia for estimating the value of a car depending on different variables
- If selling, one can search for similar vehicles and list theirs for a close price
- If buying, can search for various kinds of different cars to see the prices or see the cars they can afford in their given budget

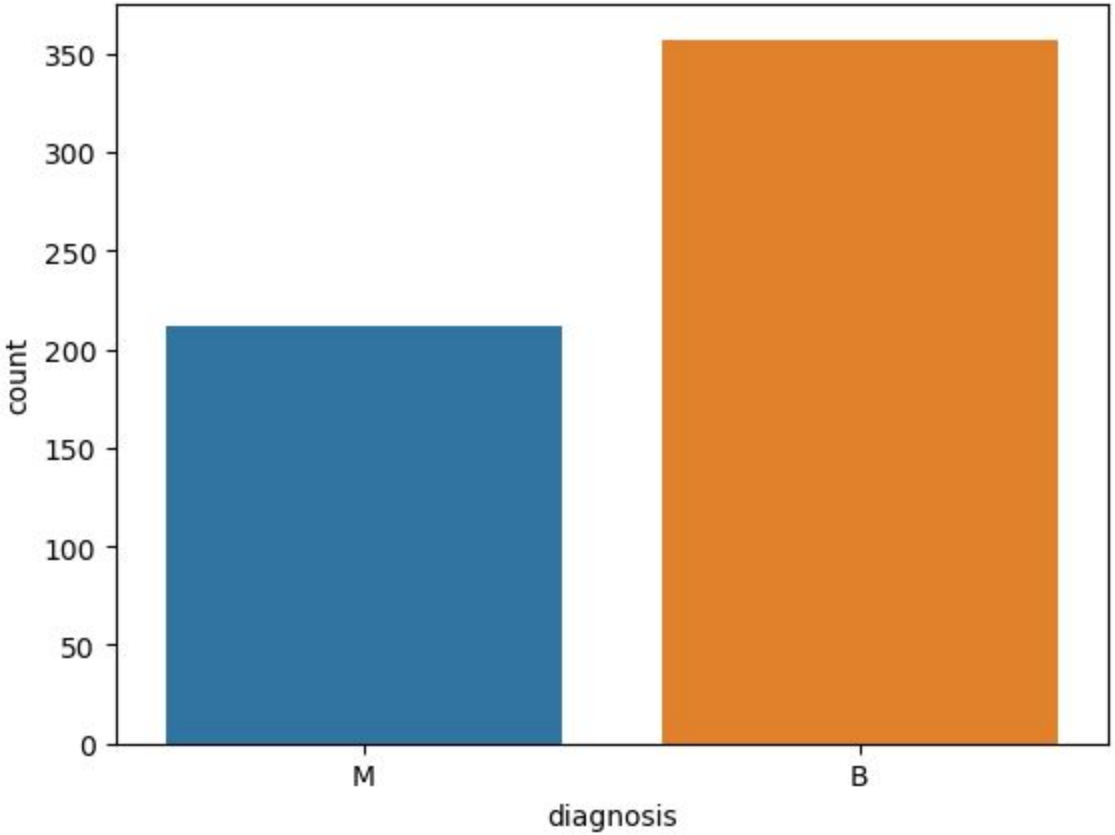
Data Overview

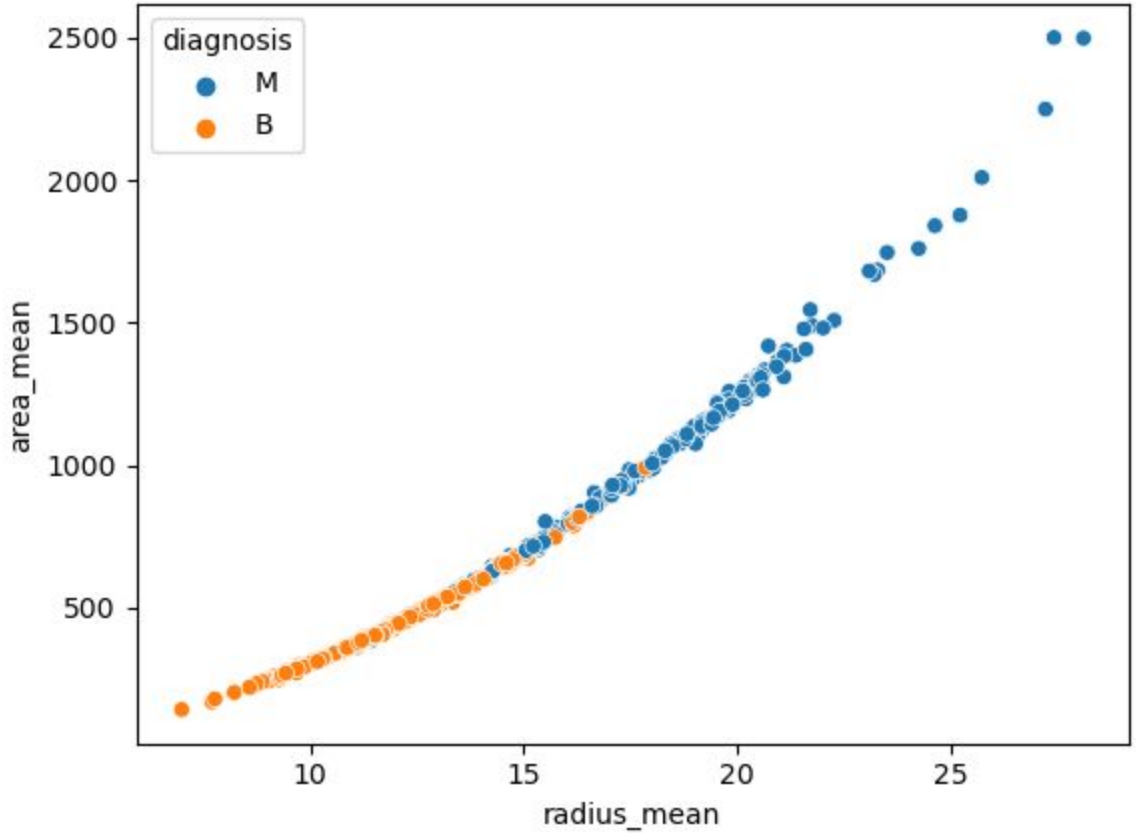
Data set

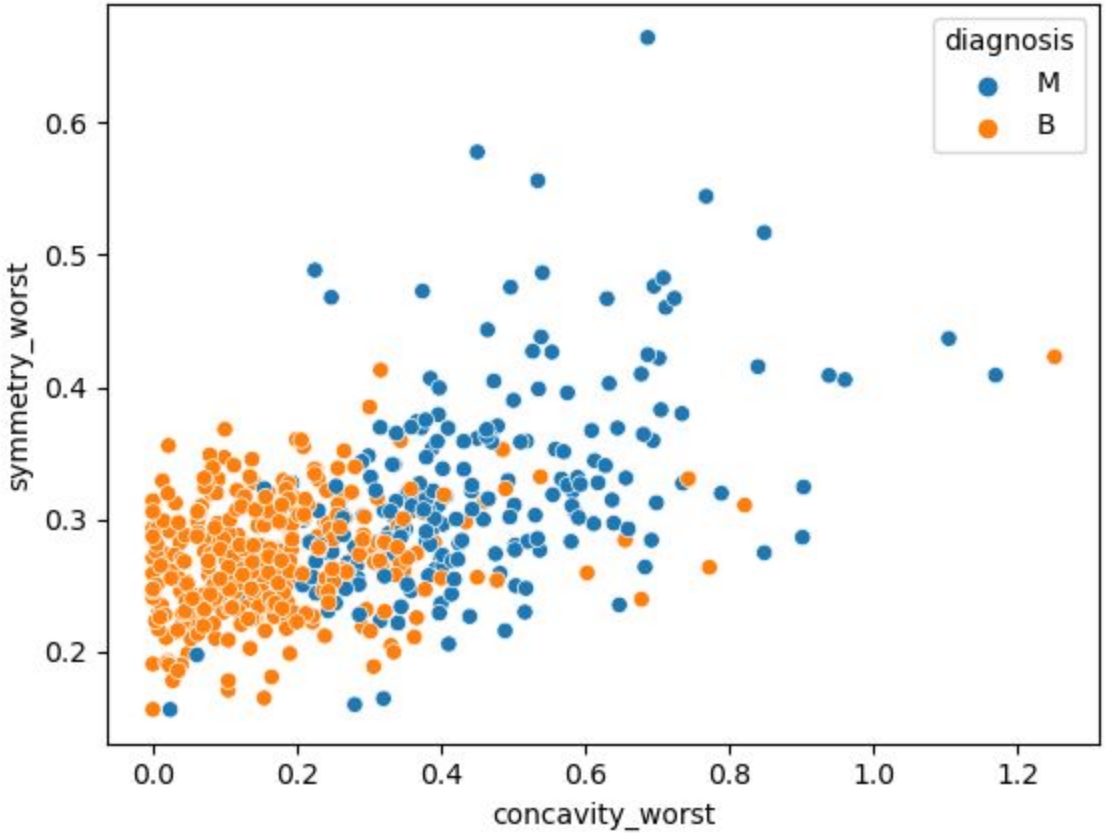
- Data set used was provided by Kaggle
- Contains data on tumor information last updated in 2022
- Rows: 16,735
- Columns: 19

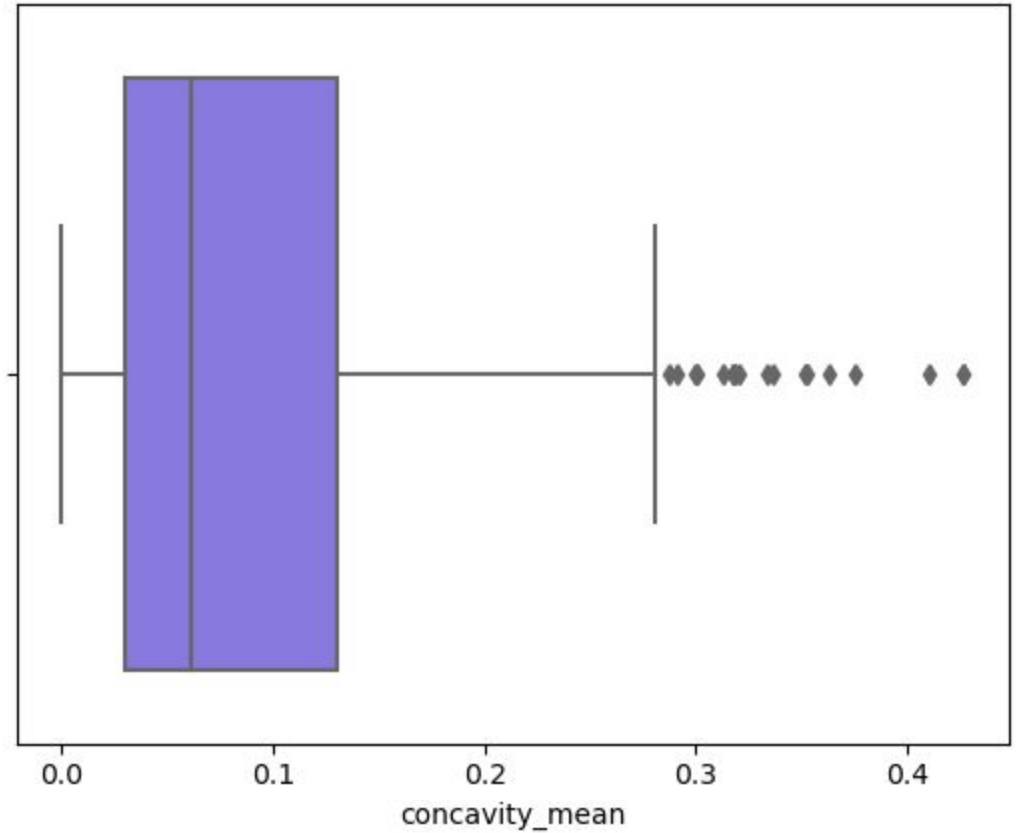
Key Features

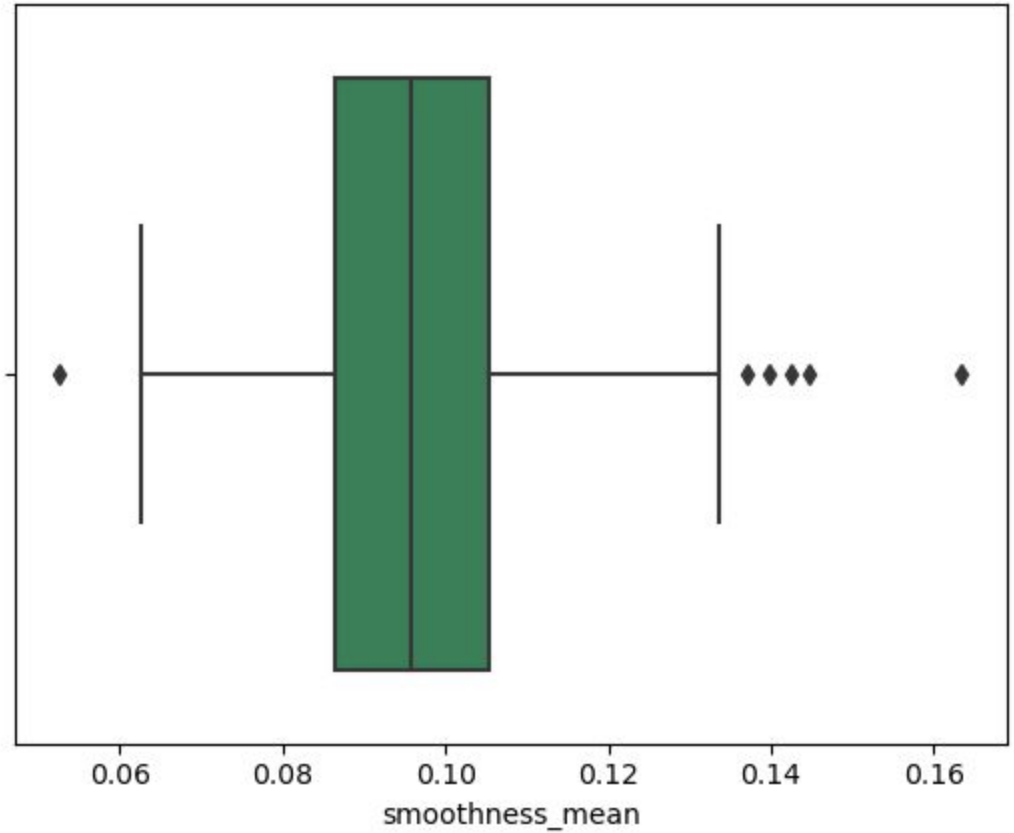
- Radius
- Texture
- Perimeter
- Area
- Smoothness
- Compactness
- Concavity
- Symmetry

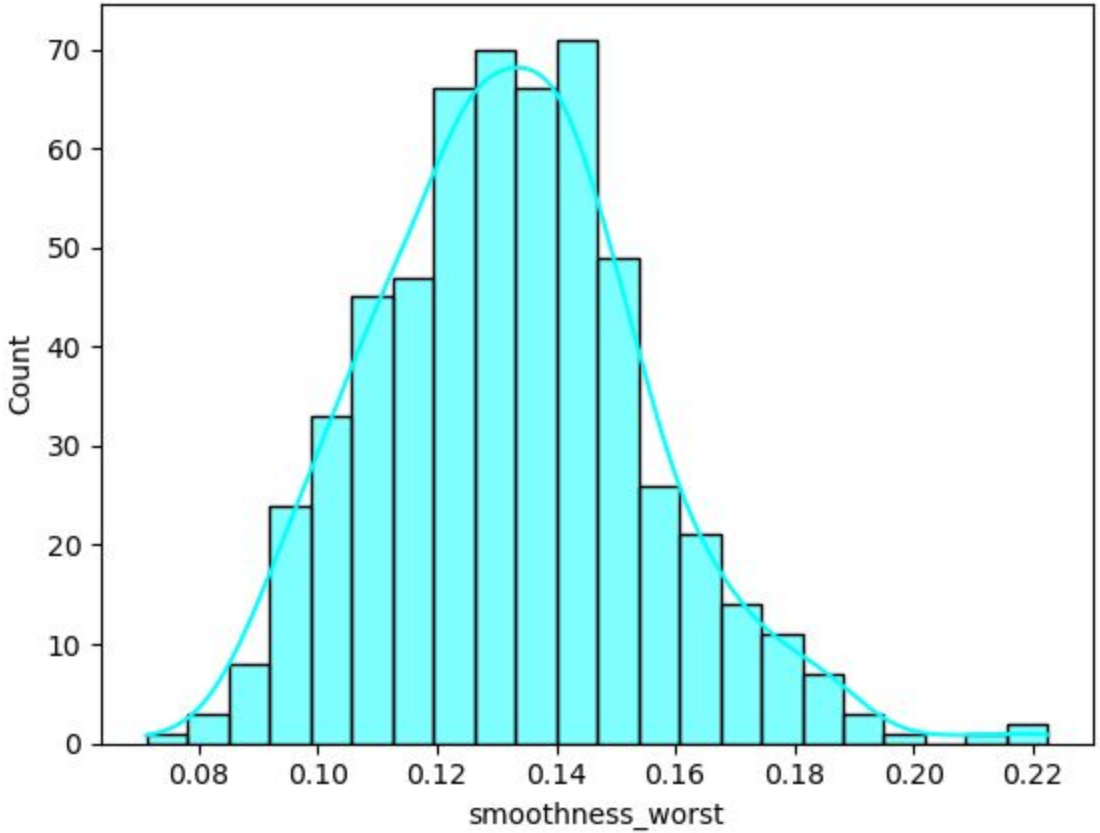

350
300
250
200
150
100
50
0
count
M
B
diagnosis

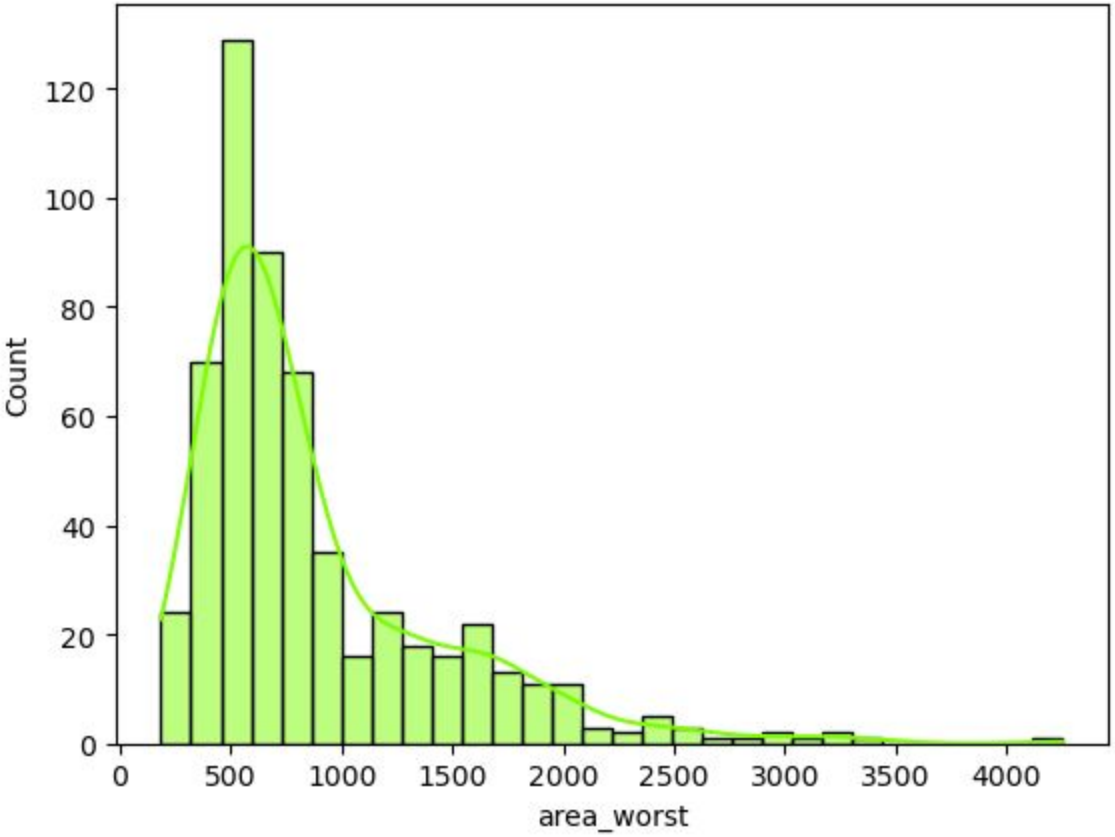

diagnosis
M
B
2500
2000
1500
1000
500
area_mean
10
15
20
25
radius_mean


diagnosis
M
B
symmetry_worst
0.2
0.3
0.4
0.5
0.6
concavity_worst
0.0
0.2
0.4
0.6
0.8
1.0
1.2


0.0
0.1
0.2
0.3
0.4
concavity_mean


0.06
0.08
0.10
0.12
0.14
0.16
smoothness_mean

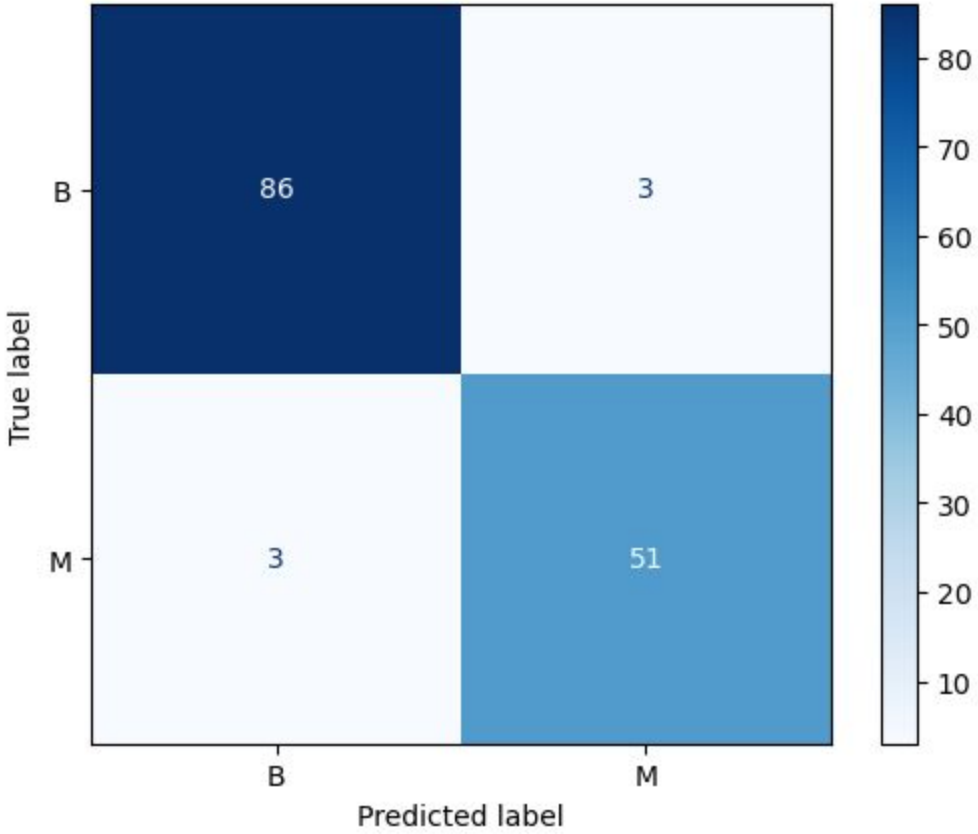

70
60
50
40
30
20
10
0
Count
0.08
0.10
0.12
0.14
0.16
0.18
0.20
0.22
smoothness_worst


Count
120
100
80
60
40
20
0
0
500
1000
1500
2000
2500
3000
3500
4000
area_worst

Standard Scaler

Train:
98.12%

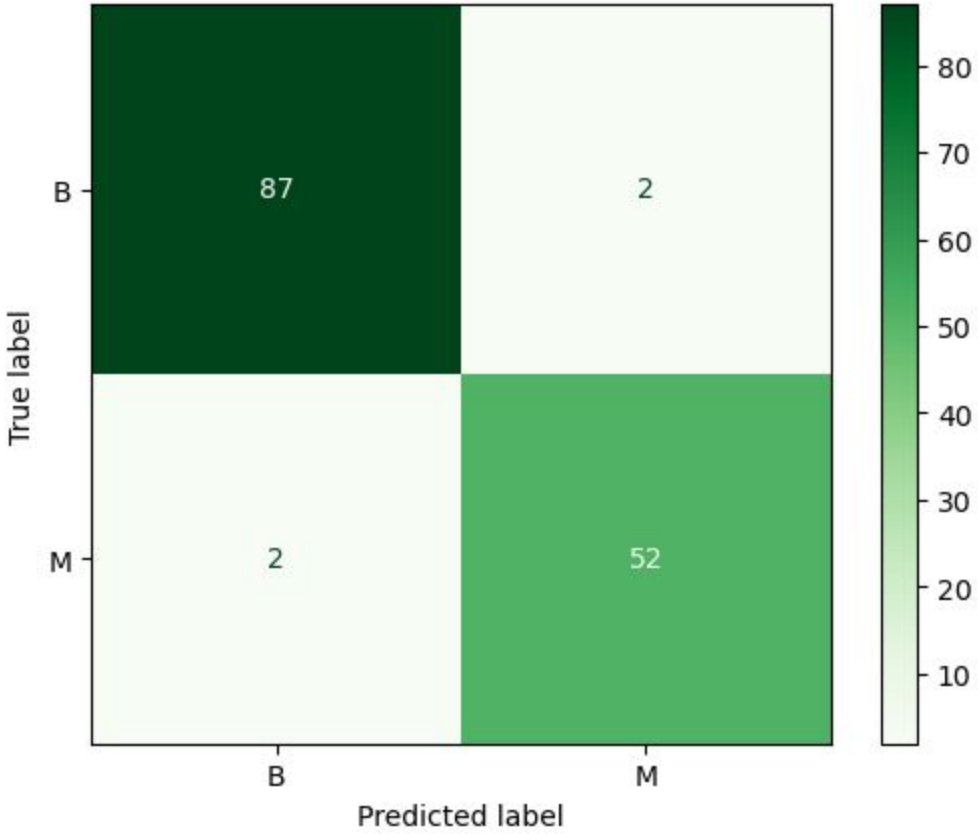
Test:
95.80%



KNeighbors

Train:
98.12%

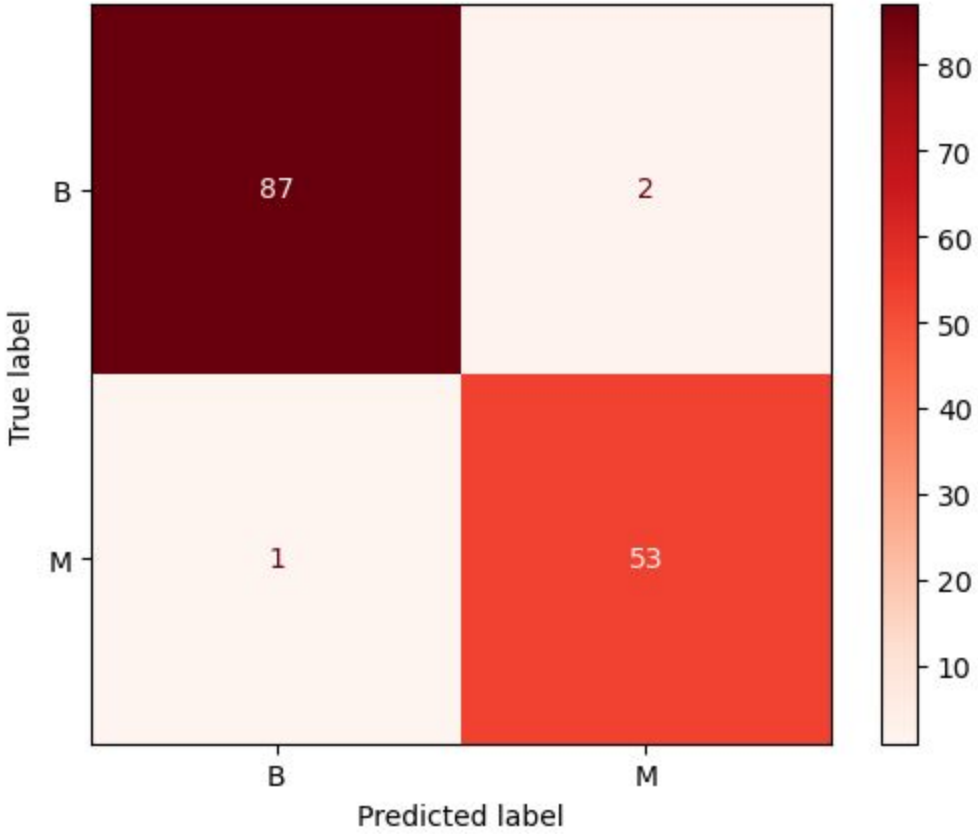
Test:
97.20%



Logistic Regression

Train:
98.83%

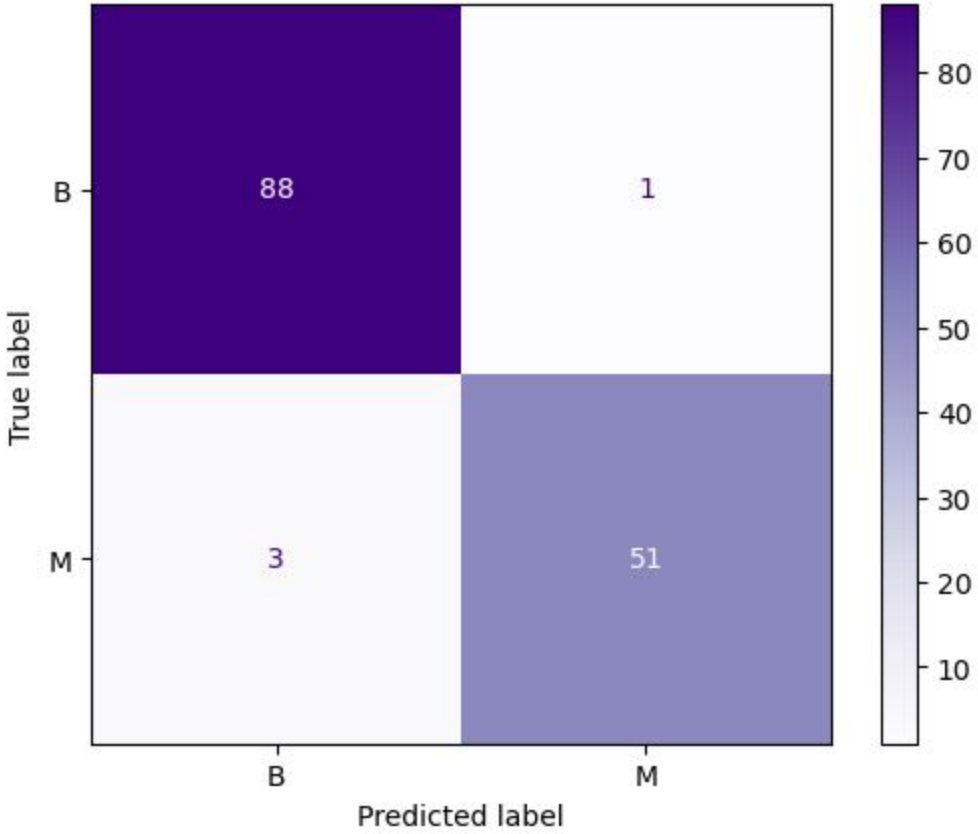
Test:
97.90%



Random Forest

Train:
100%

Test:
97.20%



Conclusion

- Overall the greater the size of the tumor the higher chance of the tumor being malignant
- There are many different variables that will affect one’s diagnosis
- Self care, tests, doctor visits will help in finding and preventing anything. The earlier the better

Thank you!

Any Questions?