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 whether a tumor is benign or malignant
- This model can be used for themselves, someone close to them, or just anyone interested in how a prediction model works

Data Overview

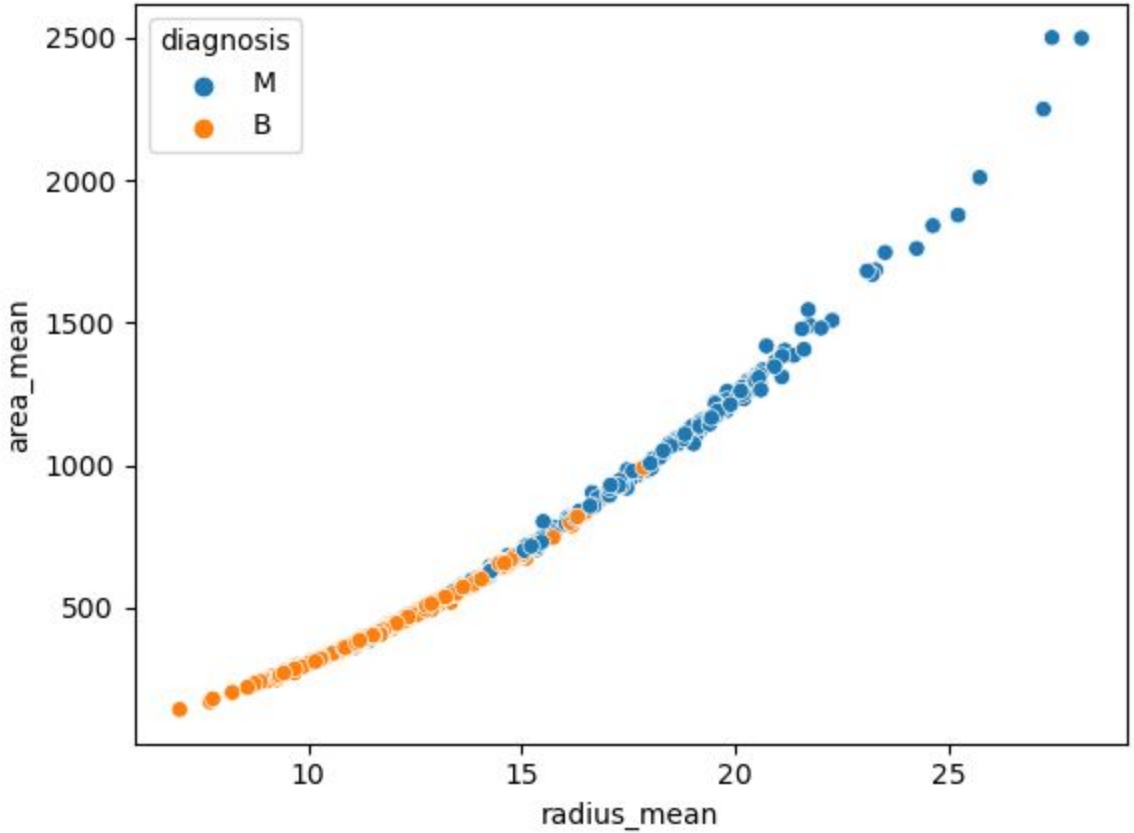
Data set

- Data set used was provided by Kaggle
- Contains data on tumor information last updated in 2022
- Rows: 569
- Columns: 32

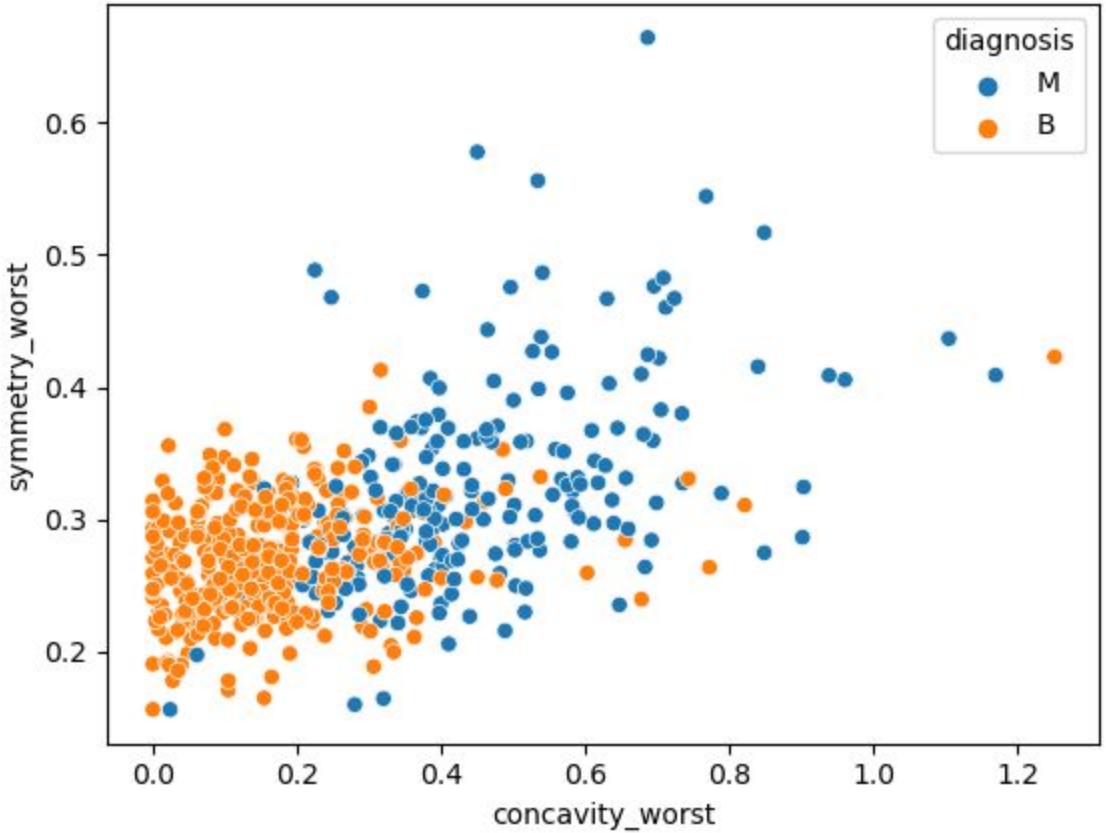
Key Features

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

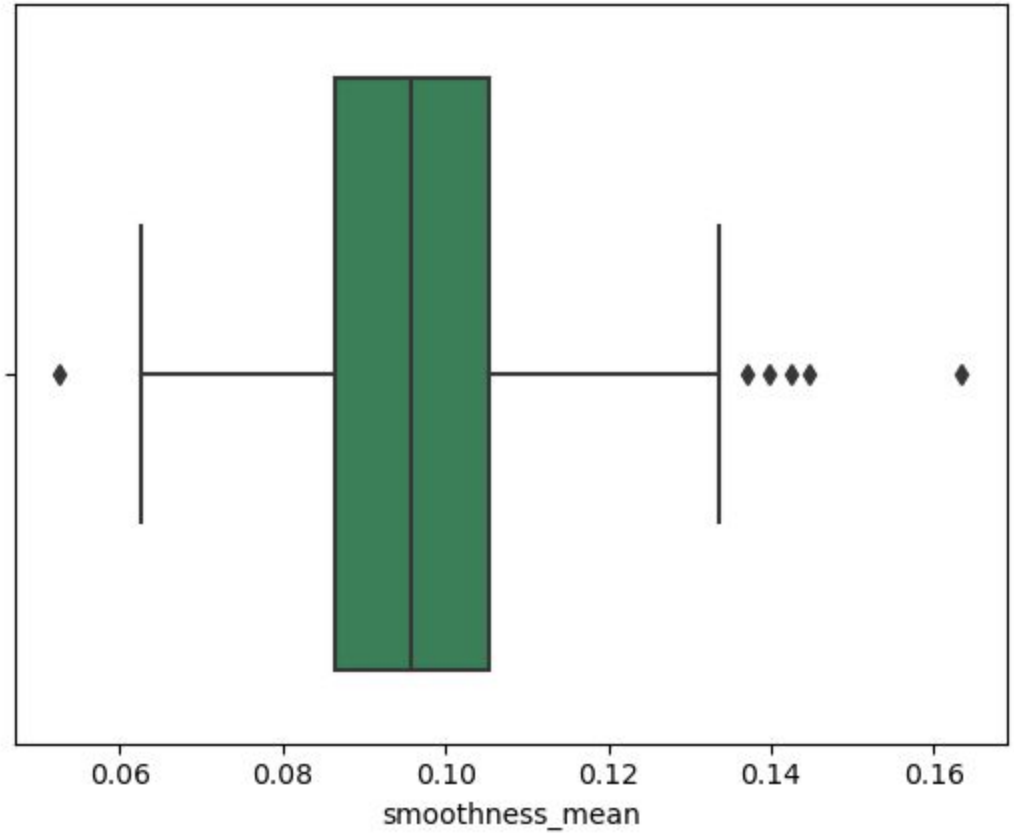
Radius & Area


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

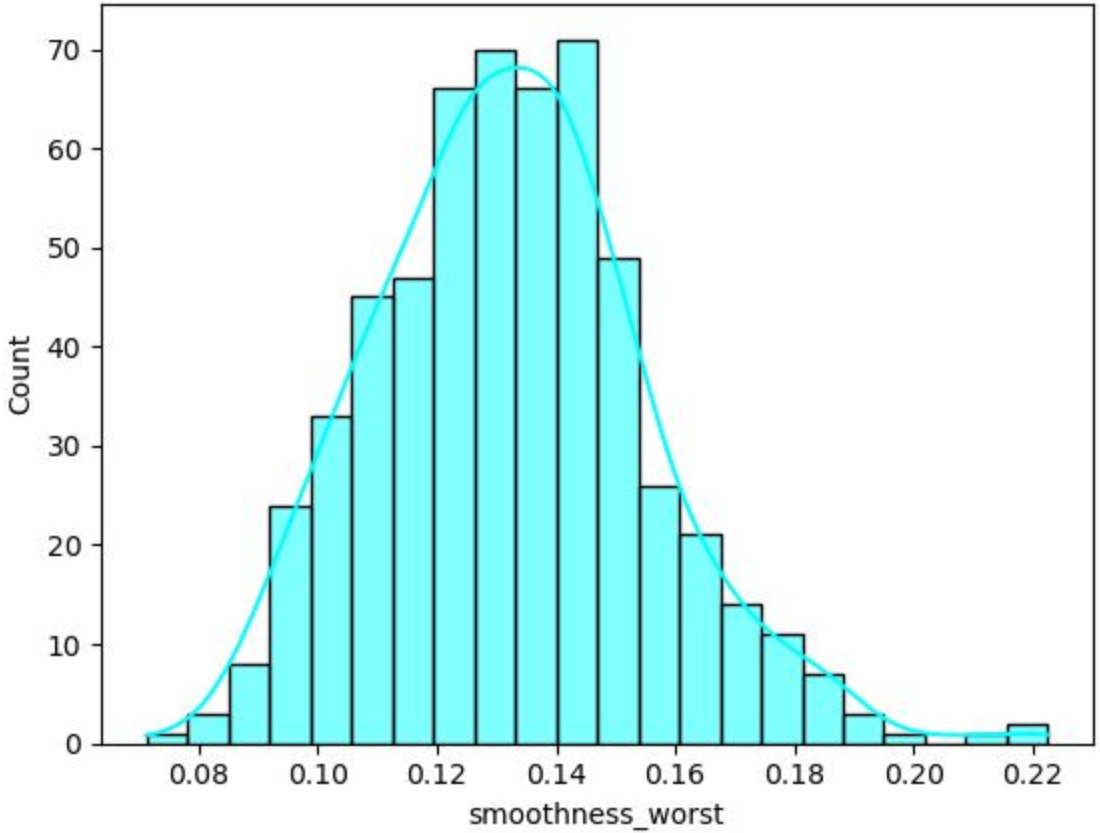
Concavity & Symmetry



Boxplot of Smoothness

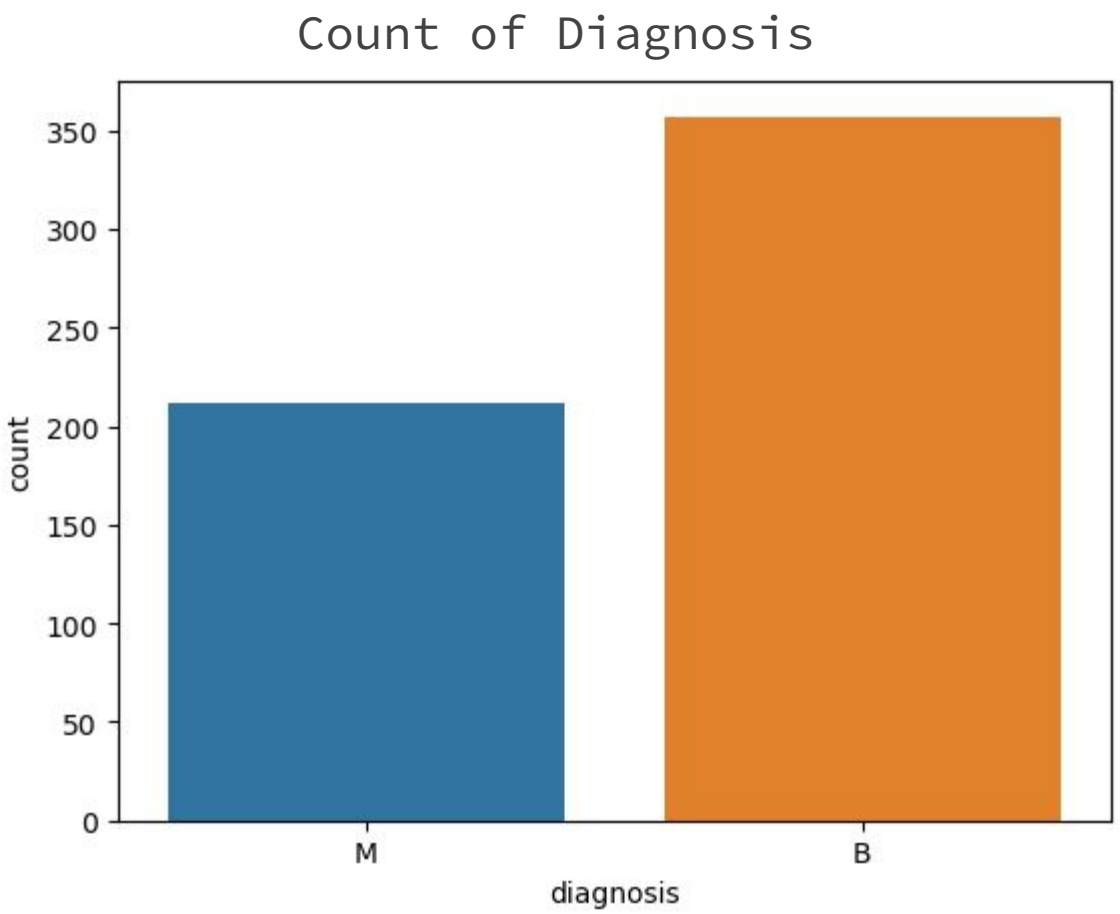


Histogram of Smoothness



Benign = 62.7%

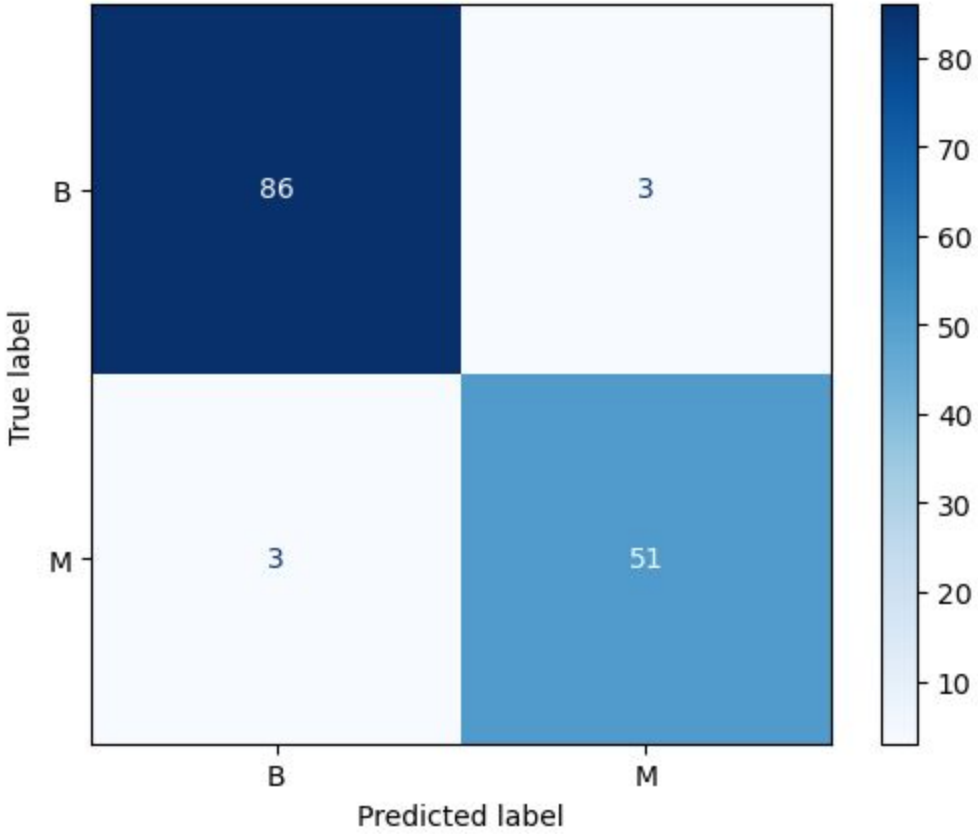
Malignant = 37.3%



Standard Scaler

Train:
98.12%

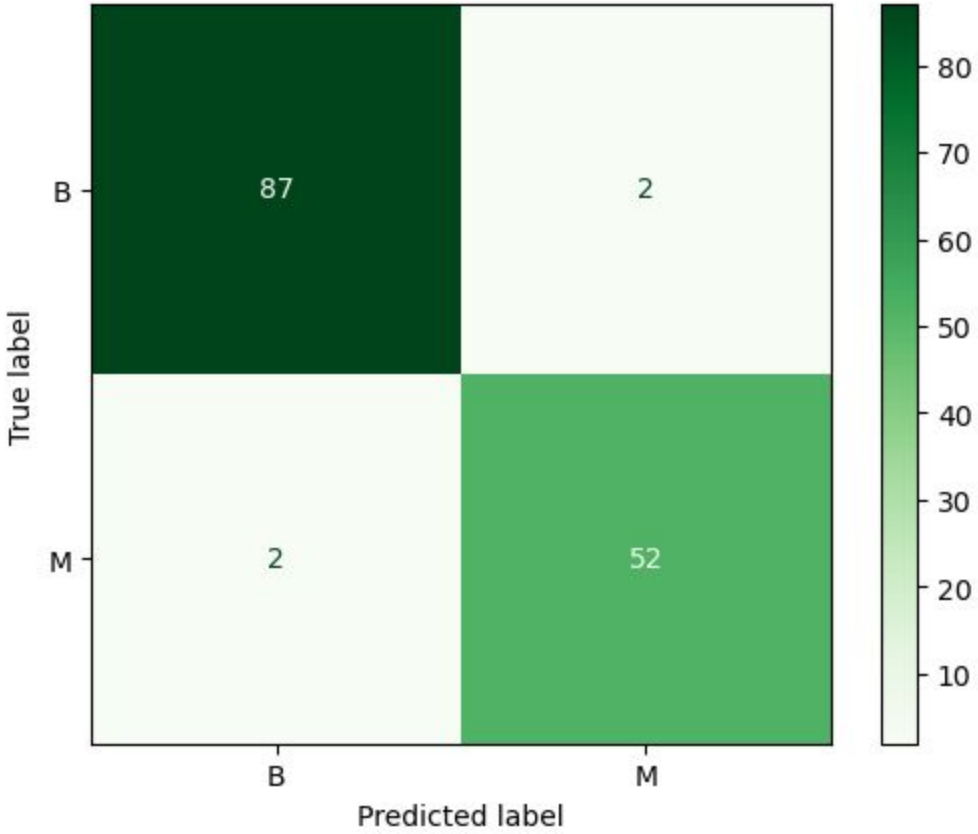
Test:
95.80%



KNNeighbors

Train:
98.12%

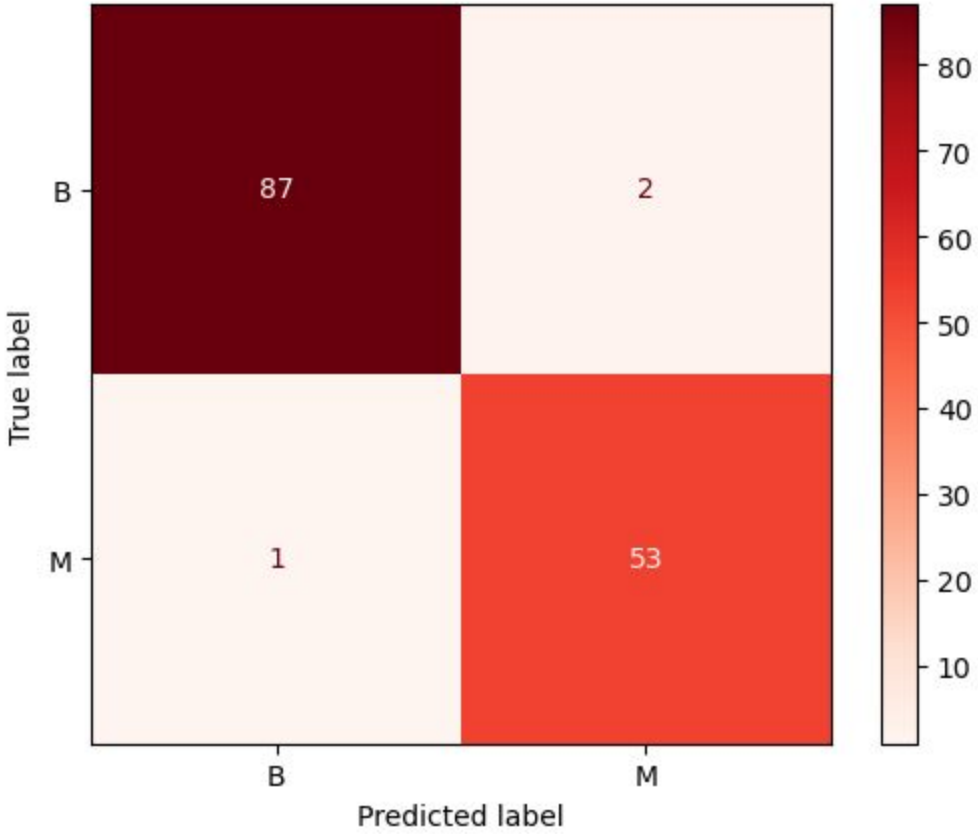
Test:
97.20%



Logistic Regression

Train:
98.83%

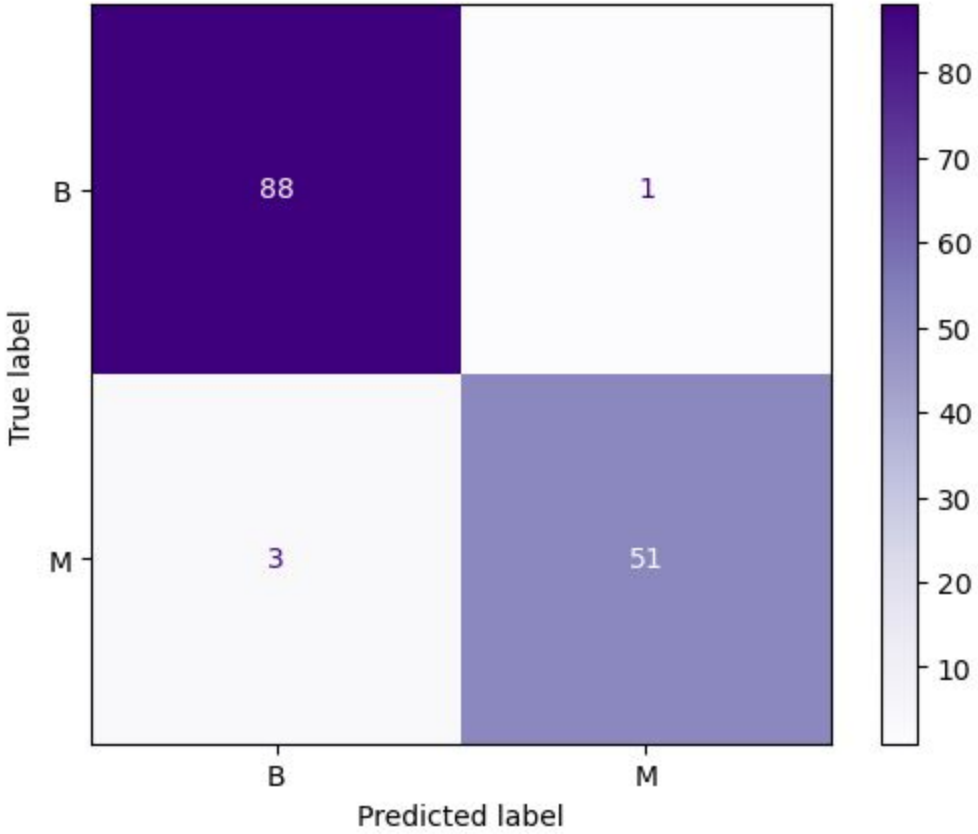
Test:
97.90%



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?