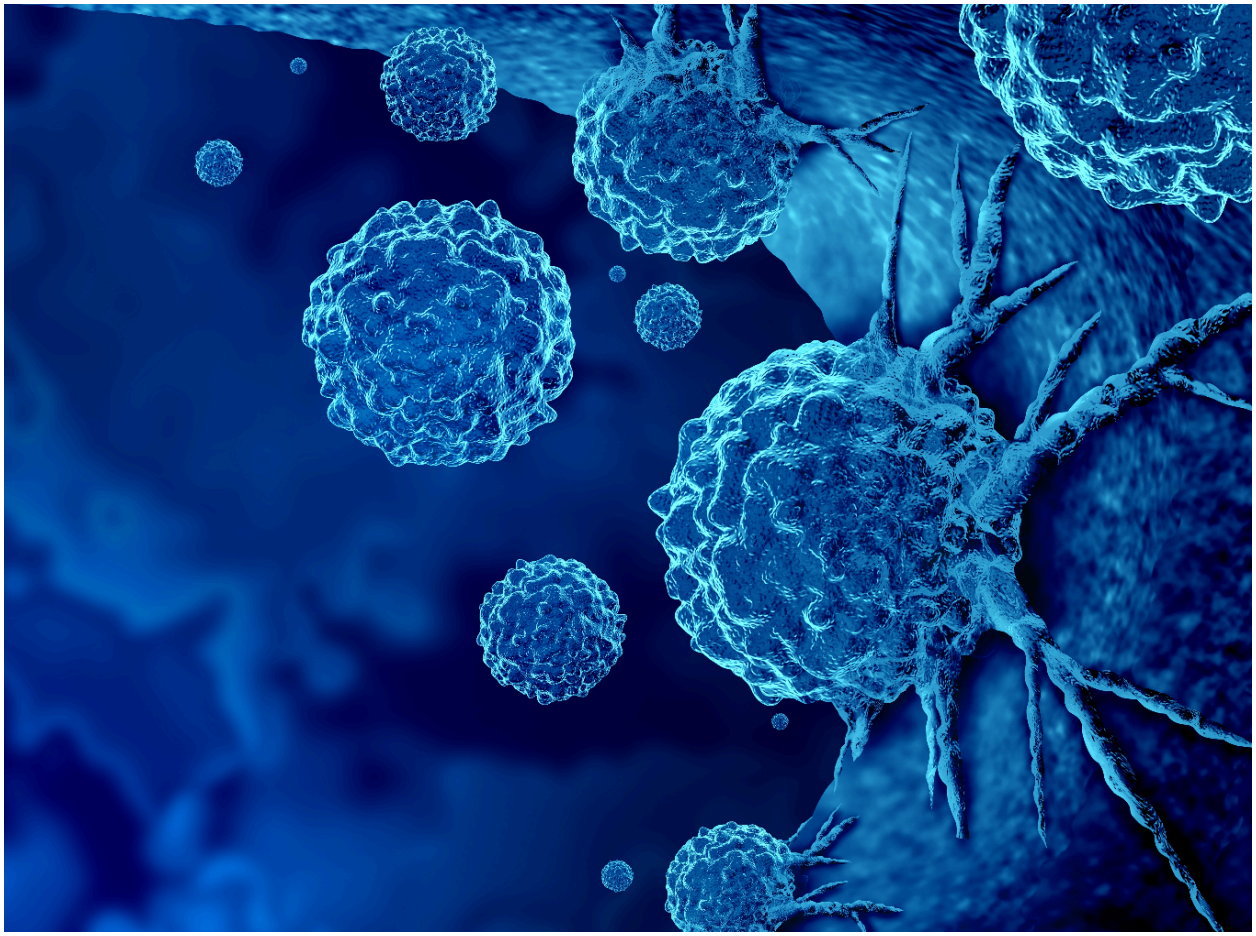


Predicting Breast Cancer as Malignant or Benign

Elijah Wikenheiser

Springboard Capstone Project

Context

Breast cancer accounts for over 12% of all new annual cancer cases worldwide. An estimated 297,790 new cases of invasive breast cancer are expected in the U.S. alone, which equates to 1 in every 8 women.[1] A tumor consisting of normal cells contained within a localized area is deemed benign and non-life threatening. A malignant tumor with abnormal cells has the potential to become invasive, leading to further complications.

Most breast cancers begin in the ducts or lobules. Breast cancer can spread outside the breast through blood vessels and lymph vessels. When breast cancer spreads to other parts of the body, it is said to have metastasized.[2] Ductal carcinoma in situ (DCIS) is a breast disease that may lead to invasive breast cancer, but the cancer cells are only in the lining of the ducts and have not spread to other tissues in the breast.

During initial testing, determining the type of tumor is vital to beginning treatment as early as possible. While there is a lack of available data, medical diagnostic errors range from 5% to 28%.[3]

Data

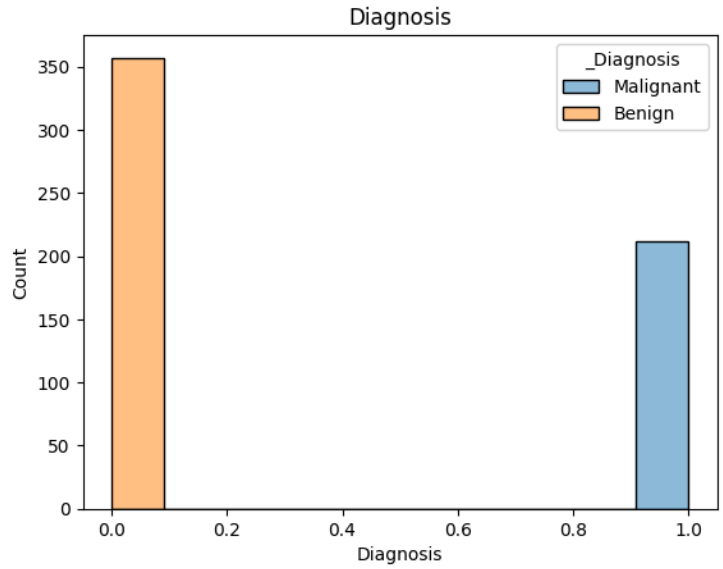
Data was obtained from the Diagnostic Wisconsin Breast Cancer Database[4]. The dataset is composed from images of tumor cell nuclei. The information includes:

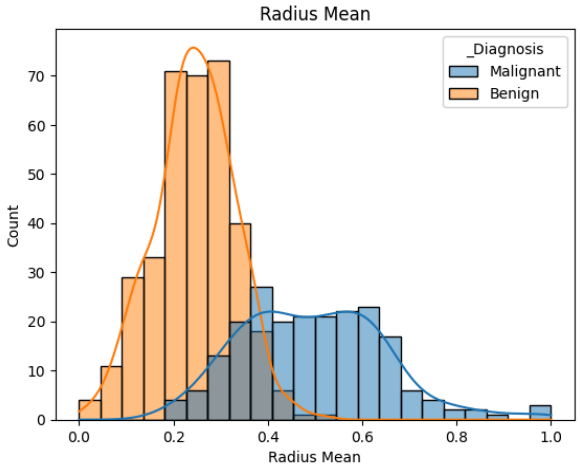
- Diagnosis
- Radius
- Texture
- Area
- Perimeter
- Smoothness
- Compactness
- Concavity
- Concave Points
- Symmetry
- Fractal Dimension

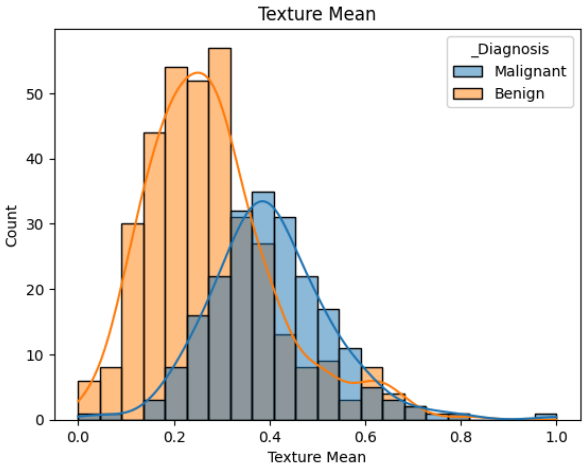
Each feature further includes the mean, standard error, and ‘worst’, which is the average of the 3 largest values.

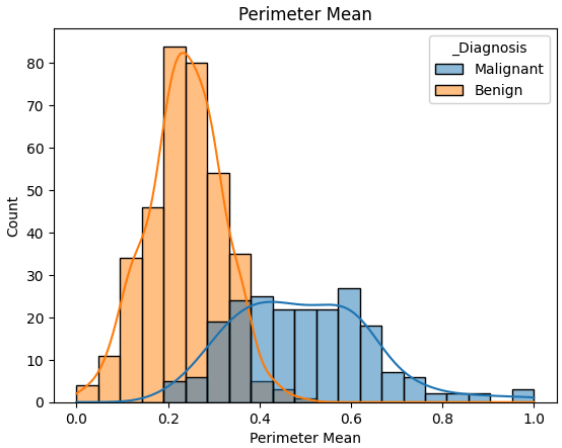
EDA

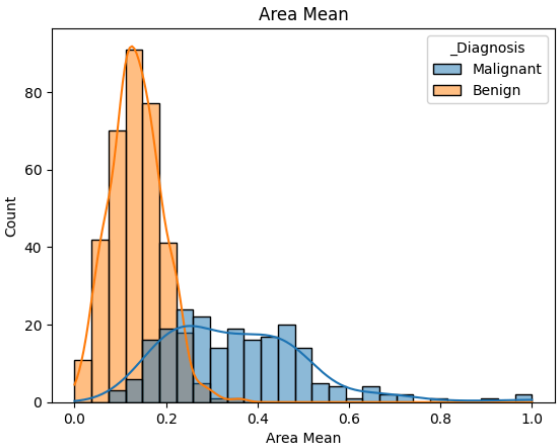
To begin, the diagnosis value was changed from a ‘M’ or ‘B’ to a ‘1’ or ‘0’. The data was then scaled for initial data exploration.

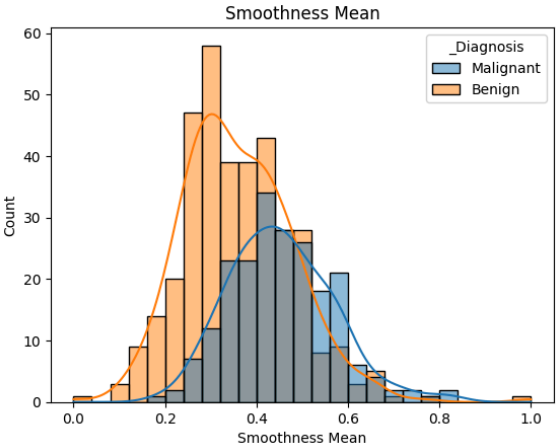


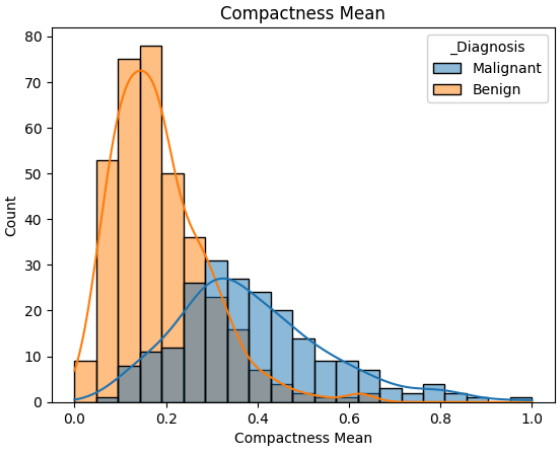


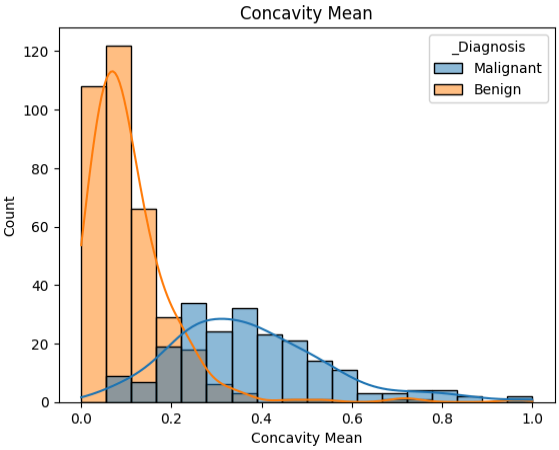


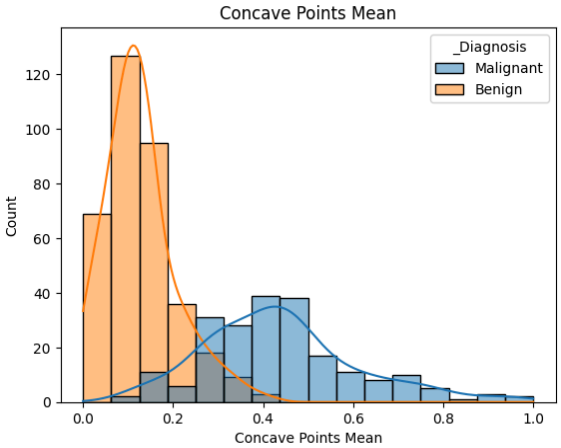

Perimeter Mean
_Diagnosis
Malignant
Benign
Count
Perimeter Mean

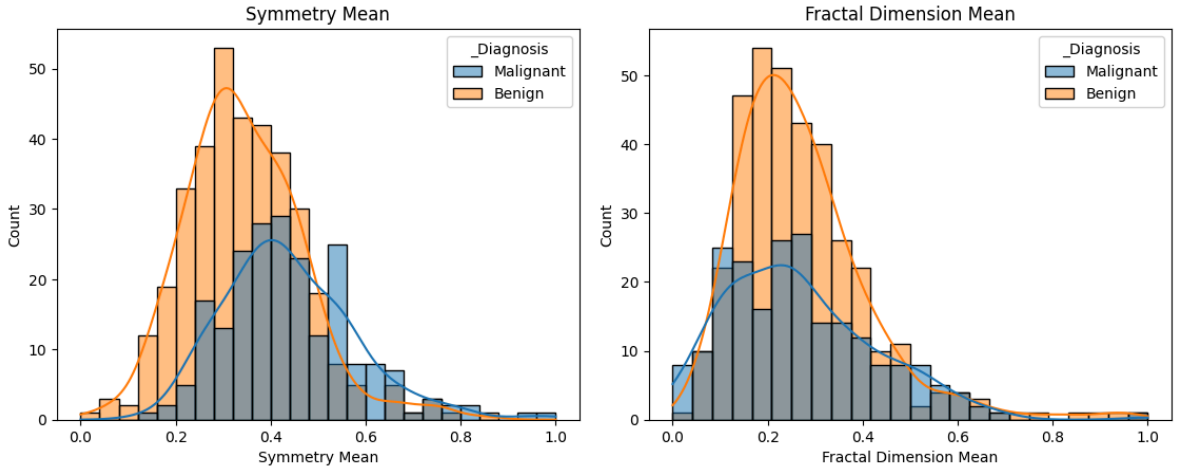

Area Mean
_Diagnosis
Malignant
Benign
Count
Area Mean

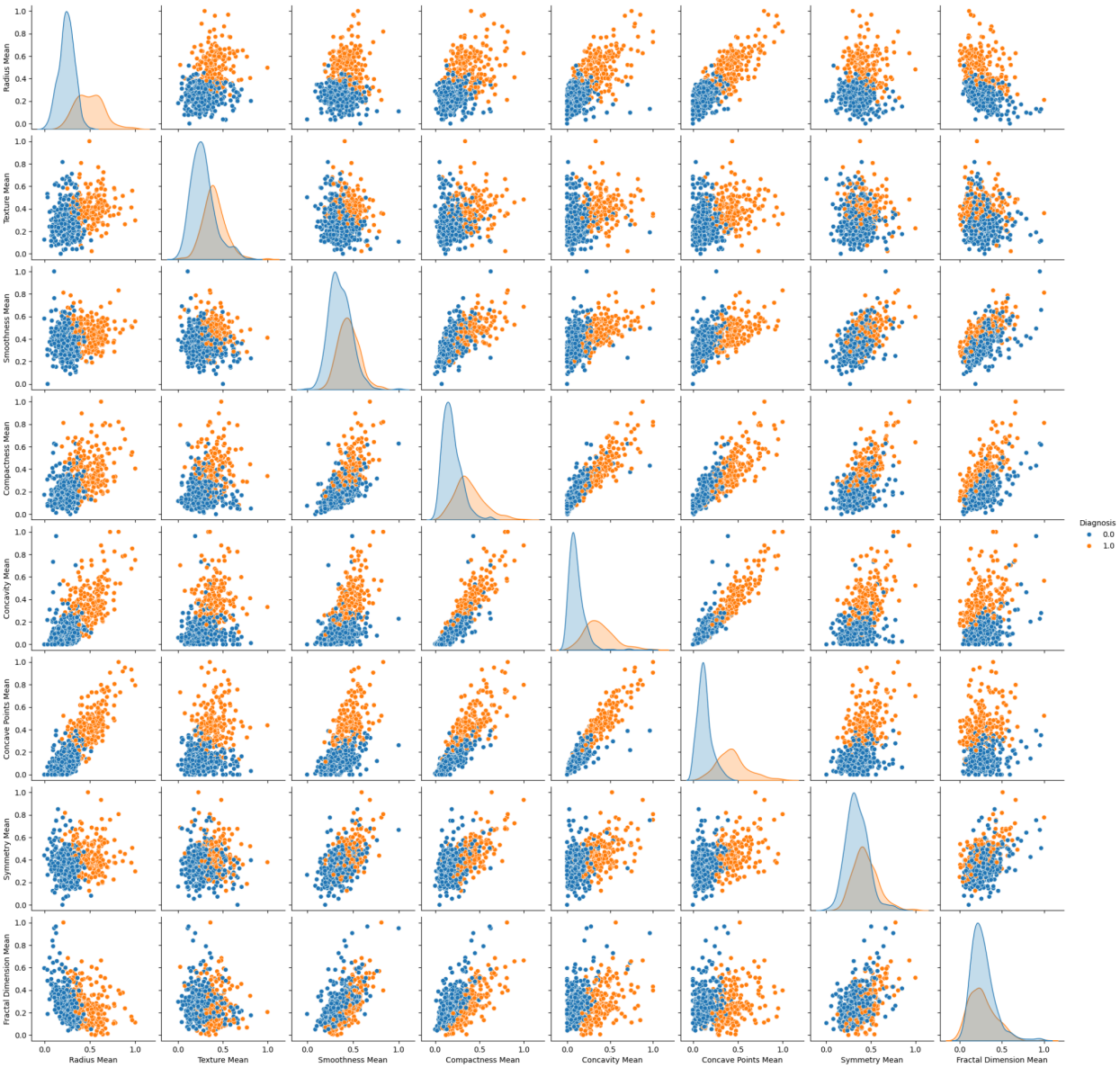

Smoothness Mean
_Diagnosis
Malignant
Benign
Count
Smoothness Mean


Compactness Mean
_Diagnosis
Malignant
Benign
Count
Compactness Mean


Concavity Mean
_Diagnosis
Malignant
Benign
Count
Concavity Mean


Concave Points Mean
_Diagnosis
Malignant
Benign
Count
Concave Points Mean


Symmetry Mean
_Diagnosis
Malignant
Benign
Count
Symmetry Mean
Fractal Dimension Mean
_Diagnosis
Malignant
Benign
Count
Fractal Dimension Mean


Radius Mean
Texture Mean
Smoothness Mean
Compactness Mean
Concavity Mean
Concave Points Mean
Symmetry Mean
Fractal Dimension Mean
Diagnosis
0.0
1.0

A pair plot

Models to test:

- Decision Tree Classifier
- Decision Tree Regressor
- Gradient Boosting Regression
- K Nearest Neighbors
- K Nearest Neighbors Grid Search
- Logistic Regression
- Neat Neural Network
- Random Forest Classifier
- Random Forest Grid Search
- Random Forest Regressor

Columns chosen:

- Radius Mean
- Texture Mean
- Smoothness Mean
- Compactness Mean
- Concavity Mean
- Concave Points Mean
- Symmetry Mean
- Fractal Dimension Mean

Results

Model	Accuracy	F1 - Score
Gradient Boosting Classifier	98.246%	0.972
Random Forest Regressor	96.491%	0.946
Support Vector Machine	96.491%	0.944
Decision Tree Classifier	95.614%	0.933
Logistic Regression	95.614%	0.933
NEAT Neural Network	95.614%	0.933
KNN Grid Search	95.165%	0.944
Random Forest Grid Search	94.505%	0.897
Random Forest Classifier	93.860%	0.904
Decision Tree Regressor	92.982%	0.897
K Nearest Neighbors	92.982%	0.944

Secondary Criteria: ‘Priyanki841’ SVM accuracy of 96.49%
Columns chosen:

- Radius Mean
- Texture Mean
- Smoothness Mean
- Compactness Mean
- Symmetry Mean
- Fractal Dimension Mean
- Radius Standard Error
- Texture Standard Error
- Smoothness Standard Error
- Compactness Standard Error
- Symmetry Standard Error
- Fractal Dimension Standard Error

Results - Priyanka841 Columns

Model	Accuracy	F1 - Score
Gradient Boosting Classifier	98.246%	0.952
Random Forest Classifier	97.368%	0.923
Support Vector Machine	96.491%	0.923
Decision Tree Classifier	96.491%	0.911
KNN Grid Search	96.491%	0.872
K Nearest Neighbors	95.614%	0.909
Logistic Regression	95.614%	0.900
Random Forest Regressor	94.725%	0.897
Random Forest Grid Search	93.860%	0.925
Decision Tree Regressor	93.407%	0.871
NEAT Neural Network	92.982%	0.857

Winning Model: Gradient Boosting Classifier

Classification Report:

	precision	recall	f1-score	support
0	0.97	0.97	0.97	78
1	0.94	0.94	0.94	36
accuracy			0.96	114
macro avg	0.96	0.96	0.96	114
weighted avg	0.96	0.96	0.96	114

Confusion Matrix:		True Class	
		Benign	Malignant
Predicted Class	Benign	74	4
	Malignant	1	35

For a pathologist, high sensitivity is preferred over high specificity. Incorrectly diagnosed as positive is easier to catch and accommodate for during or after treatment.[3]

Conclusions

- Out of 10 classification models, Gradient Boosting Classifier provided the best results at 98.246%.
- With further data, model could be further trained with a preference to specificity.
- In conjunction with other cytogenetic tests, classification models could provide another tool for catching breast cancer earlier and with higher accuracy.

Citations

- [1] Breast Cancer Facts and Statistics. Breast cancer facts and statistics 2023. (n.d.). https://www.breastcancer.org/facts-statistics
- [2] Support. (2021, February 8). How common are breast cancer misdiagnoses?. Hale & Monico. https://www.halemonico.com/2021/01/13/how-common-are-breast-cancer-misdiagnoses/
- [3] William Sukov, M.D.