Predicting Breast Cancer

A report by Jessica Montealvo

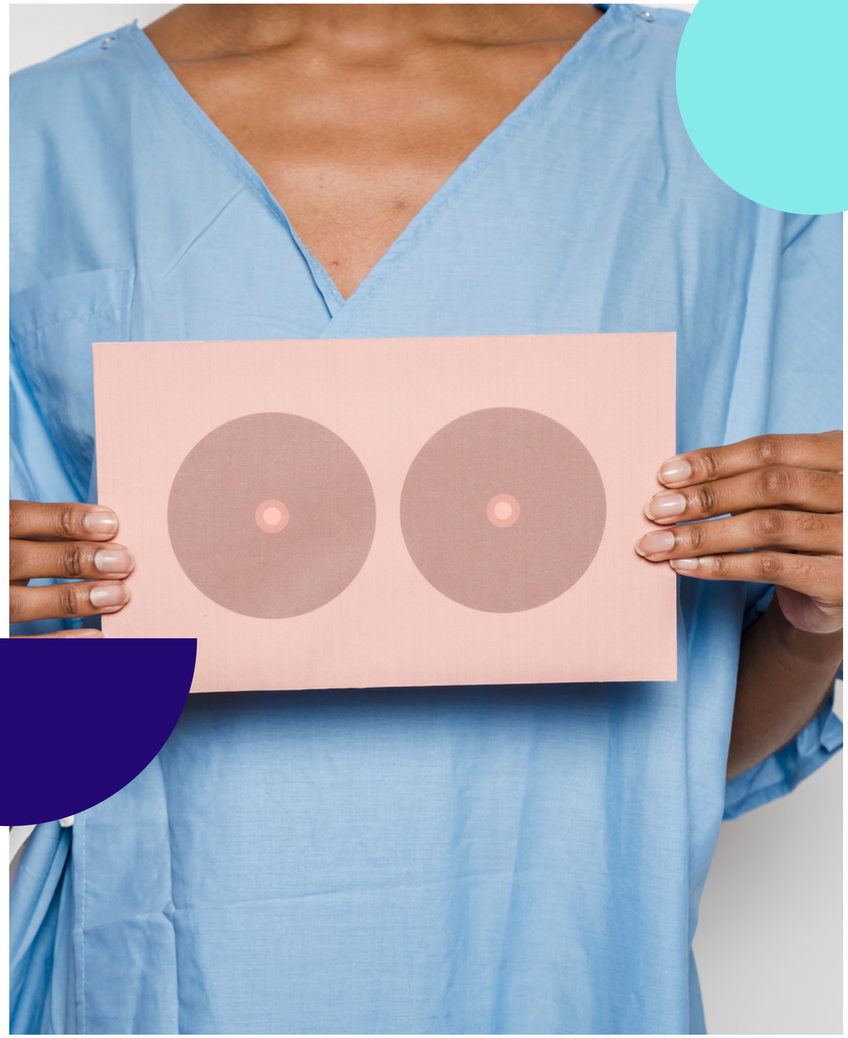
Introduction

Last year, breast cancer took the lead in the number of new cases reported, with 2.26 million diagnoses

Although prominent, breast cancer is not as deadly as other cancers, mainly because it benefits from early detection. Early detection strategies have been implemented throughout the years and rely on both clinical and self-breast exams as well as mammogram screening. Earlier-stage cancers can be treated more effectively, which readily contributes to survival. So the goal is always to identify cancer early.

Probelm Statement

In this project, I will use the Breast Cancer Wisconsin (Diagnostic) Data Set to train a model to predict whether a tumor is benign or malignant.



The Data

UCI MACHINE LEARNING REPOSITORY

The data set used for this project can be found on the UCI Machine Learning Repository. It was gathered by Dr. William H. Wolberg and his team at the Univeristy of Wisconsin in 1995

569 CASES

The data describes 569 cases of digitized images of biopsies, in particular a fine needle aspirate (FNA) of breast tissue, when breast cancer was suspected.

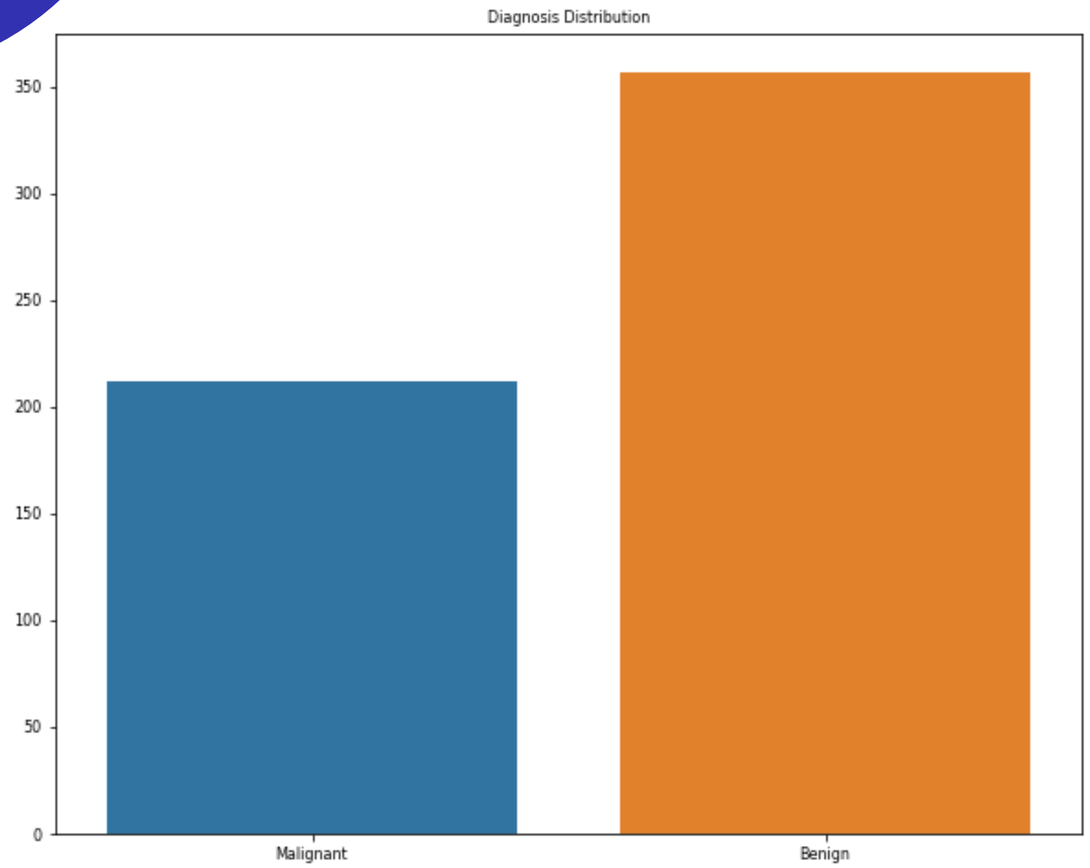
30 FEATURES

The features in this data set are measurements derived from the images and describe the cell nuclei in the tissue.

Data Wrangling

- There were no missing values
- Drop ID column (present in some sources)
- Diagnosis column translated to
  - malignant = 1
  - benign = 0

Exploration Data Analysis



Diagnosis Distribution

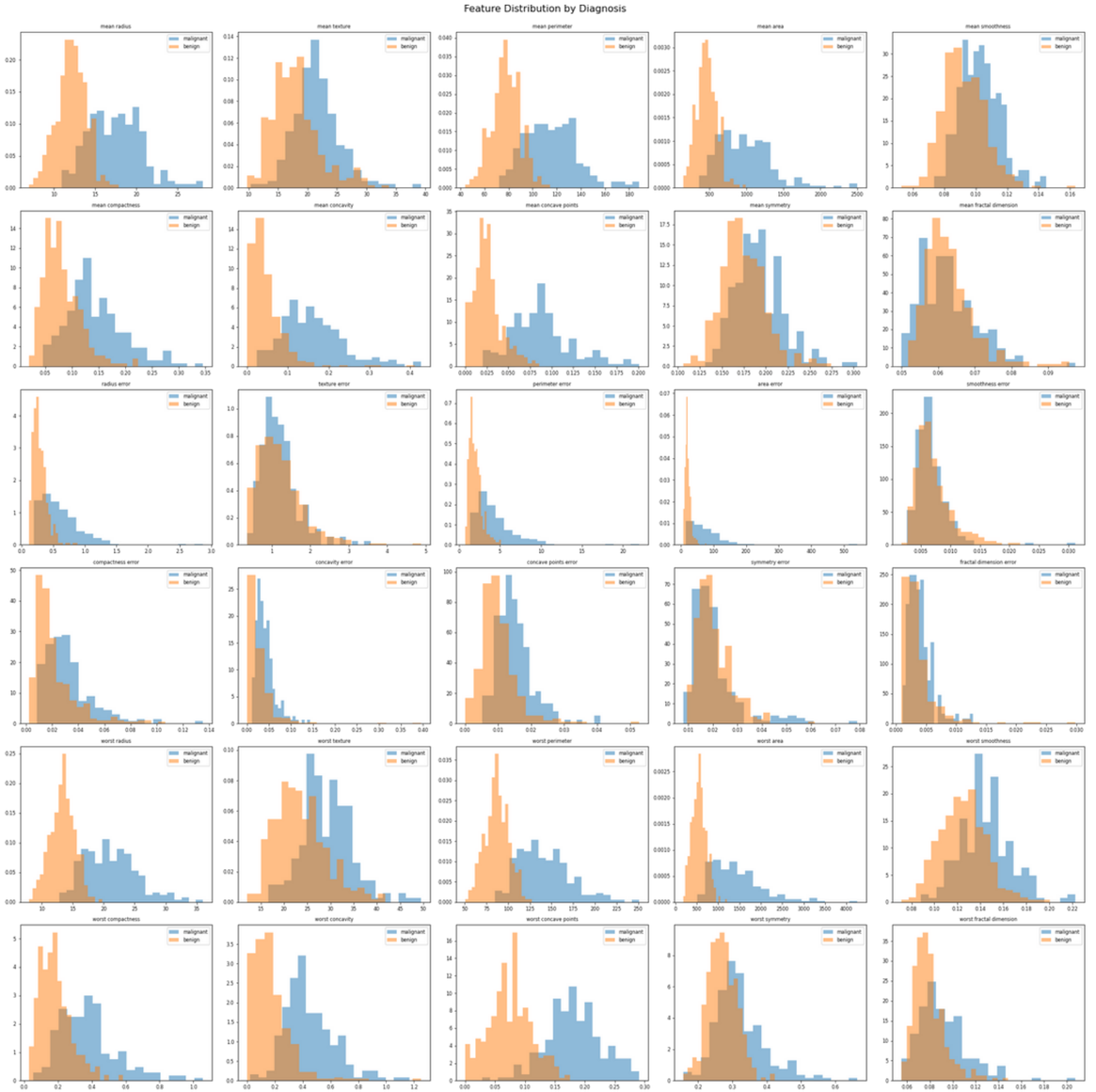
Not Cancer (benign): 62.74%
Cancer (malignant): 37.26%

The following graph shows that our data set has two classes (cancer/ not cancer). The classes are not balanced. For this reason, I will upsample in my training data set to avoid my models predicting correctly solely based on the fact that there are more Not Cancer instances.

Feature Distribution

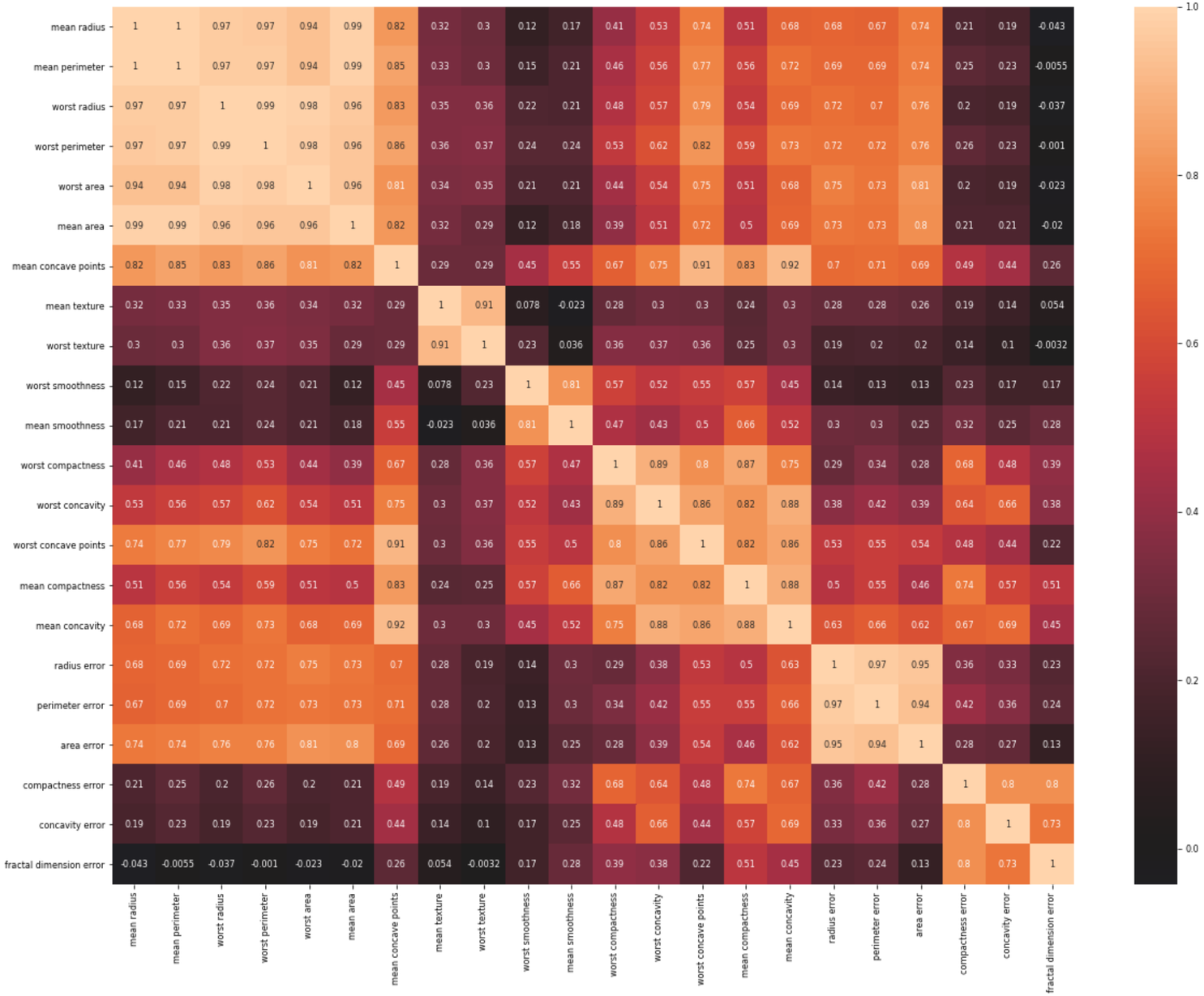
Cancer cells in blue

Cancer cells have less compact nuclei; accordingly, they are distributed on the larger portion of the spectrum for radius, mean texture, mean perimeter, area, concavity, and concave points. Correlations are evident.



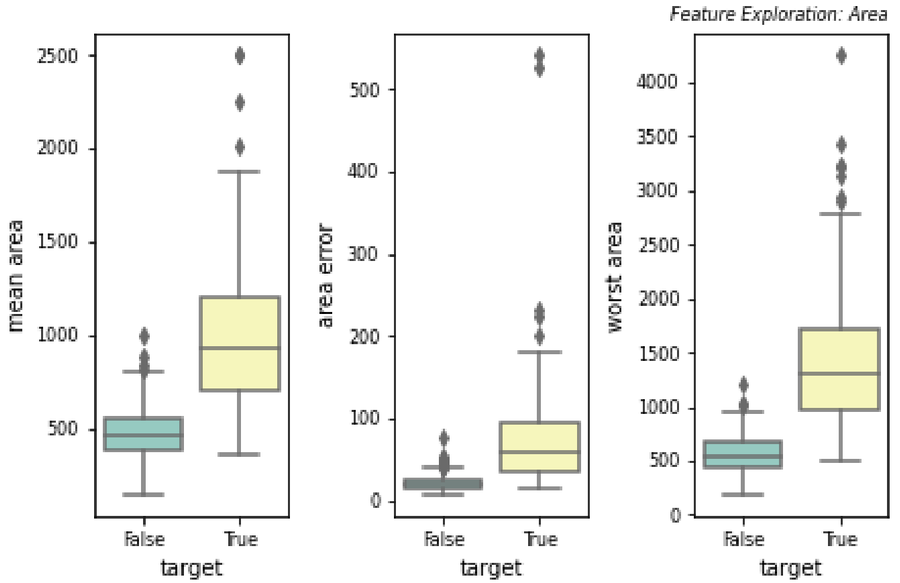
Correlation Matrix Snapshot

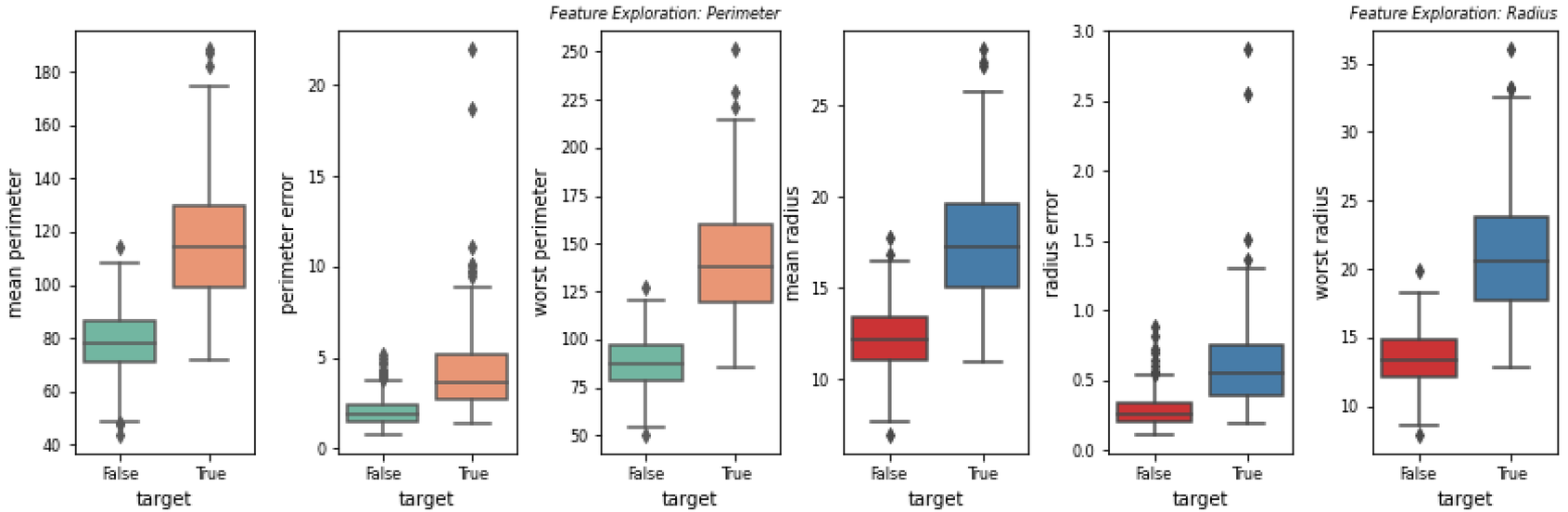
As evidenced by the correlation matrix, namely by the peach-colored squares in the matrix, in many instances, the mean, worst, and error measurements for each category correlate to each other.



Collinear Features

Features that describe the same phenomenon, which may confuse or mislead a model. A robust feature selection will need to tease these apart.

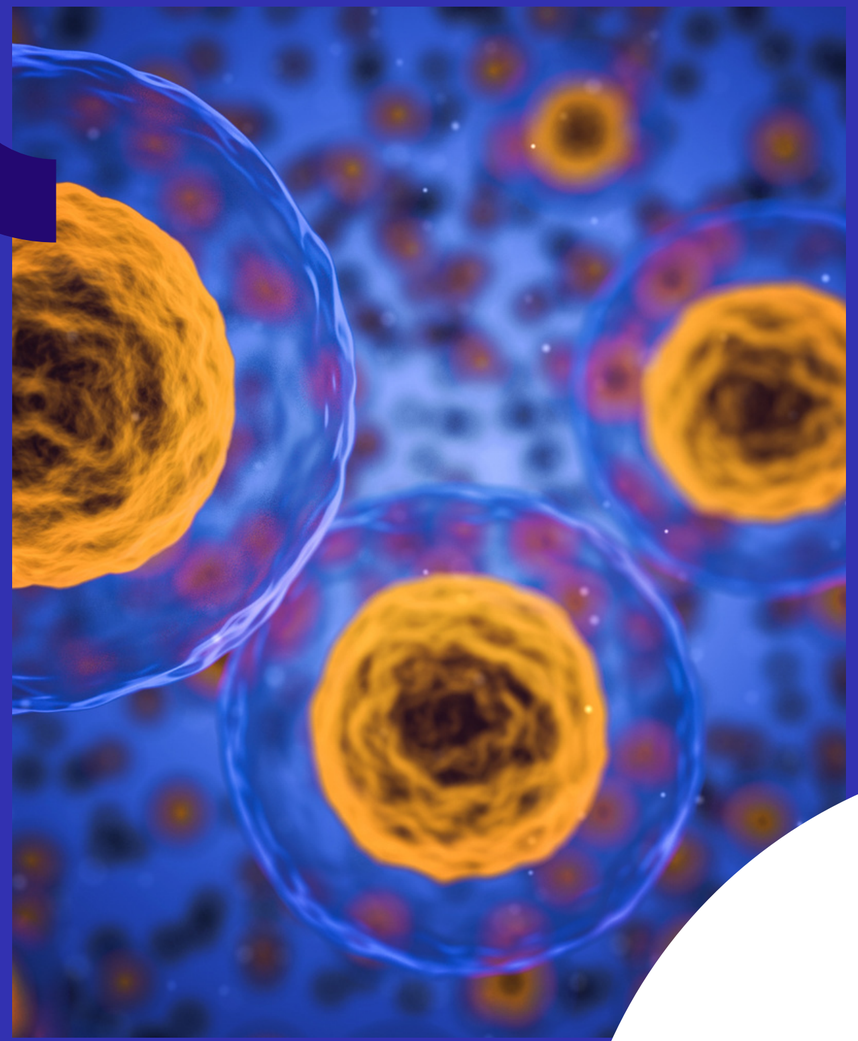




Feature Selection

Six feature selection methods used

- Select K Best
- Recursive Feature Elimination
- LassoCV
- Random Forest Feature Importance + Hierarchical clustering
- Only Means
- <80% correlated features



Feature Selection Results

Feature Set	Selected based on	Number of Features	LogisticRegression ROC AUC Score on Test Data
Recursive Feature Elimination with Cross-Validation	RFECV	16	0.9977
Random Forest Feature Importance	Hierarchy based on Correlation and node distance	14	0.9974
Lasso CV	LassoCV/LogReg	22	0.9972
Original	Kept all features	30	0.9971
SelectKBest	SelectKBest	15	0.9915
Mean Features	Removed error and worst feature sets	10	0.9903
Least Correlated Features	Removed correlated features (threshold=0.8)	16	0.9900

Machine Learning

Modeling

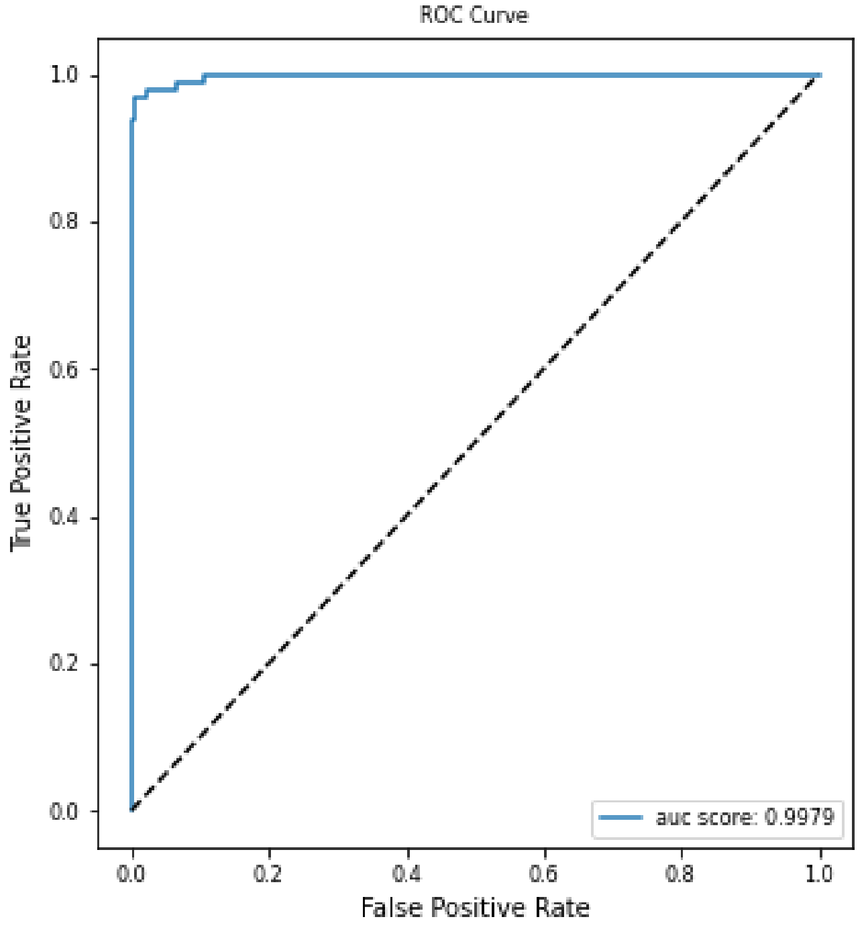
Data set selected from the Recursive Feature Elimination with Cross-Validation.

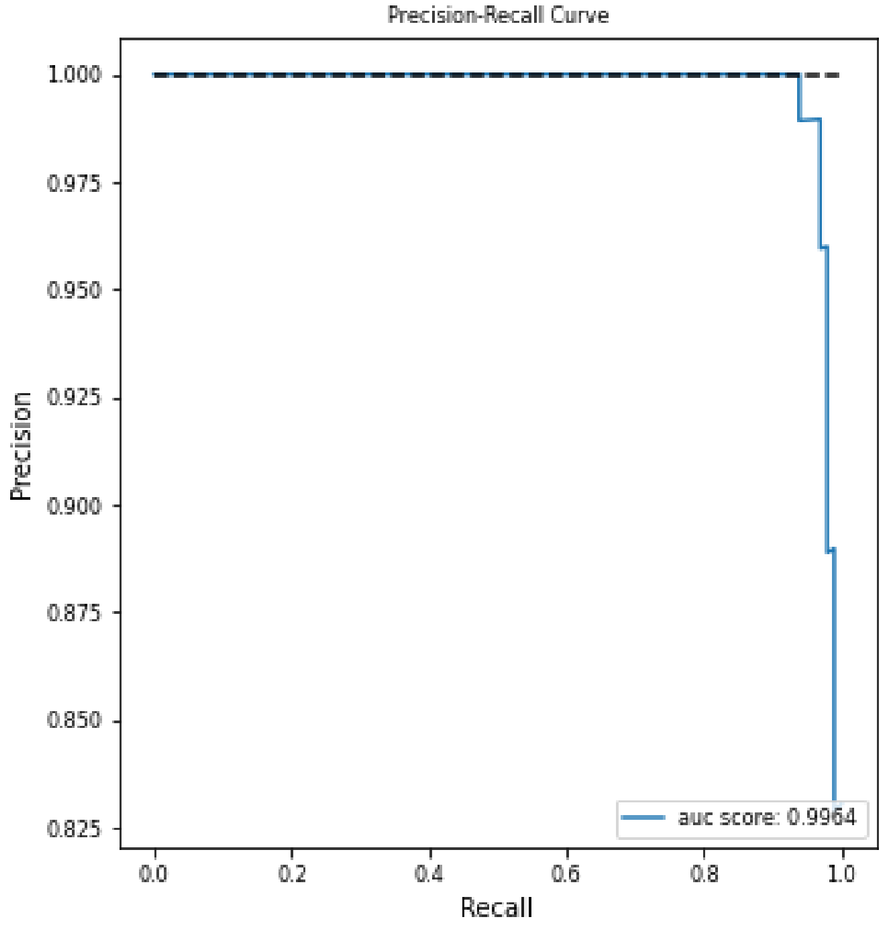
Three different models were assessed via Grid Search CV.

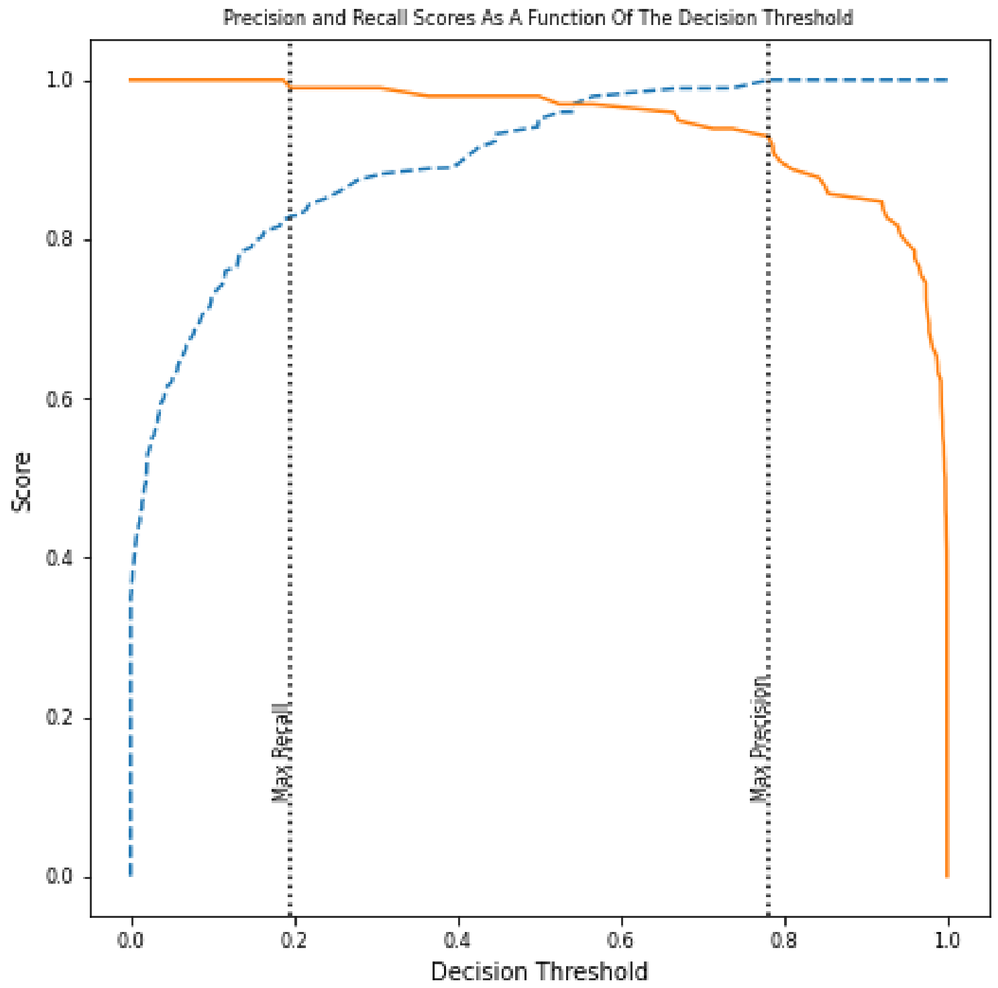
Model	Best Score	Best Parameters	Cross-Validated ROC AUC Score (Test Set)
LogisticRegression	0.9823	{'C': 0.4, 'max_iter': 100, 'penalty': 'l2', 'solver': 'liblinear'}	0.9979
Gradient Boosting	0.9764	{'learning_rate': 0.5, 'max_depth': 2, 'n_estimators': 500}	0.9973
Random Forest	0.9705	{'criterion': 'entropy', 'max_depth': 5, 'max_features': 'log2', 'min_samples_leaf': 1, 'min_samples_split': 5, 'n_estimators': 15}	0.9946

Best: Logistic Regression

ROC AUC = 0.9979 Precision/Recall = 0.9964







Precision and Recall Decision Threshold

PRIORITIZING RECALL

Minimize false-negative errors. These would be the instances where a person walks away with a “not cancer” diagnosis, all the while harvesting a malignant tumor.

PRIORITIZING PRECISION

Minimize false-positive errors, which in this project are less harmful since diagnosing a human with cancer while not having cancer will not be detrimental.

Max Recall

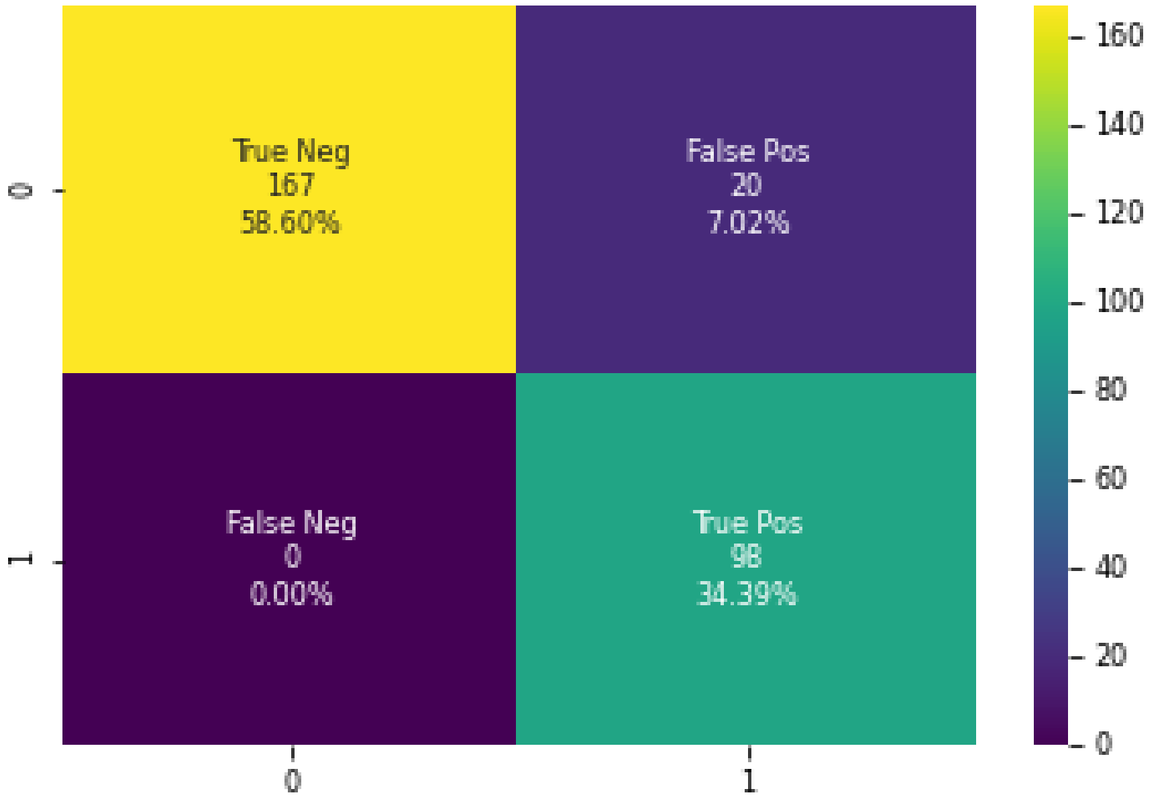
Threshold: 0.1961

Classification Report – Max Recall

	precision	recall	f1-score	support
False	1.00	0.89	0.94	187
True	0.83	1.00	0.91	98
accuracy			0.93	285
macro avg	0.92	0.95	0.93	285
weighted avg	0.94	0.93	0.93	285

Setting a threshold of 0.1961 will prevent false negative predictions, thereby giving the patient a chance at early detection and treatment.

Success!



Although our threshold allows for more people to get a false cancer diagnosis, no patient will walk away thinking they are healthy when they really have cancer.

Conclusions

FEATURE SELECTION

A thorough feature selection process only kept the most essential and informative features, effectively avoiding collinear features.

MODELING AND THRESHOLDING

Logistic regression model:
ROC AUC: 0.9979
Precision/Recall: 0.9964
Decision Threshold: 0.19 -> Recall= 1.00

NEXT STEPS

Analyze the images directly and derive the nuclei measurements via an image processing model

References

WORLD HEALTH ORGANIZATION
https://www.who.int/news-room/fact-sheets/detail/cancer

SCIKIT-LEARN DOCUMENTATION
https://scikit-learn.org/stable/index.html