BREAST CANCER INDICATORS

Group 2

OUR QUESTIONS

What model is best at predicting cell malignancy?

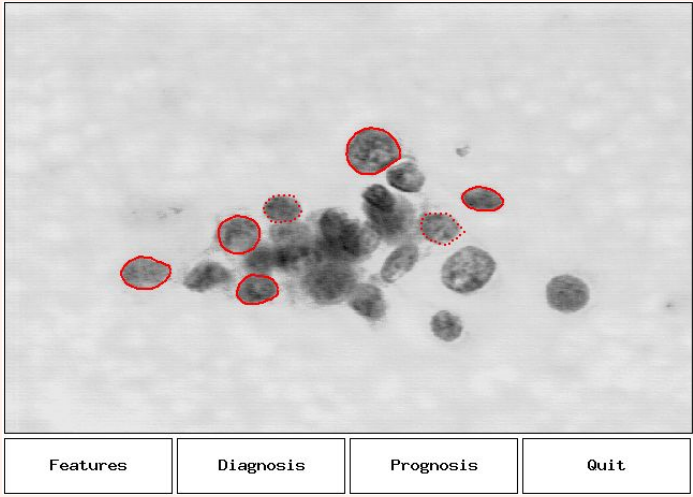
Which features are most reliable in predicting cell malignancy?

OUR DATASET

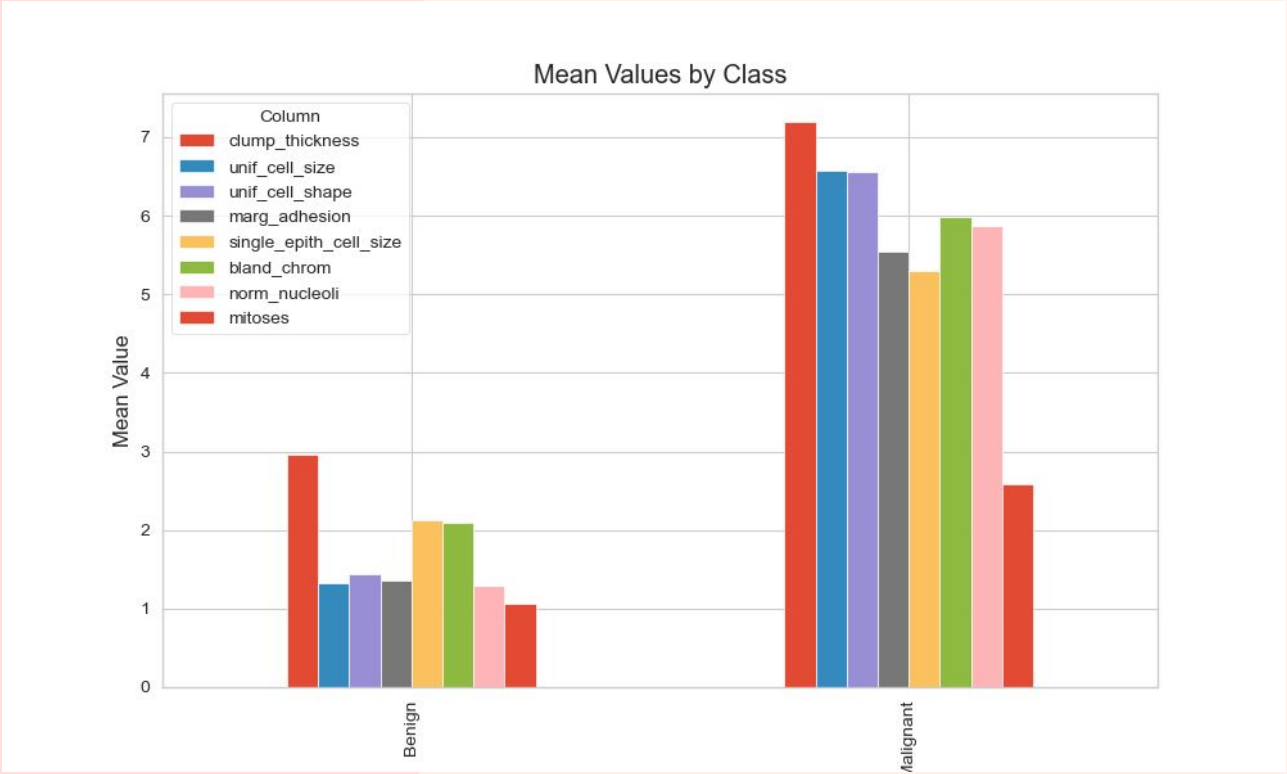
- Collected by Dr. William H. Wolberg at the University of Wisconsin Hospitals (1989 - 1991)
- Common indicators of breast cancer: clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitosis
- 699 samples, including 50 repeat patients
- 65.5% of cells designated benign

XCYT (NUCLEUS TRAINING)

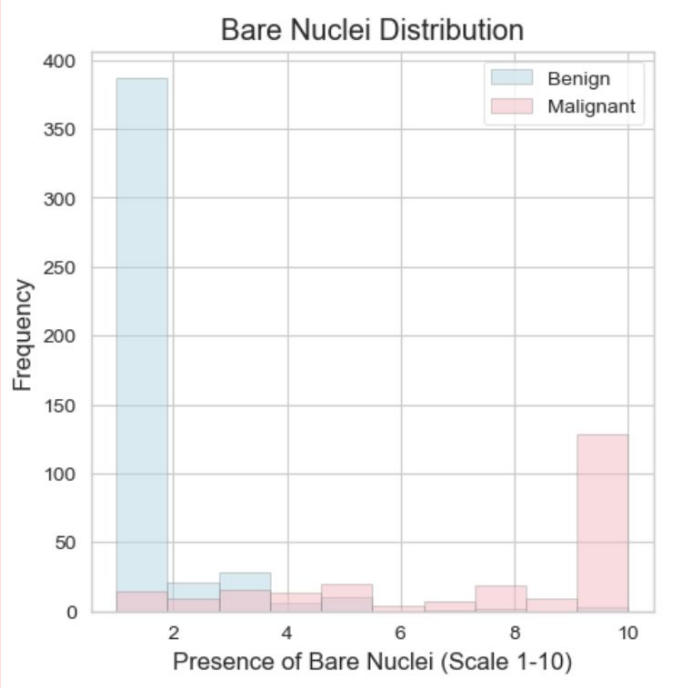
- The nucleus of each cell is hand-drawn by the user and then it is processed by Xcyt
- Once the nucleus of each cell is clearly delineated the program calculates the score of each feature.
- Scores are calculated with a value of 1-10 for 9 features

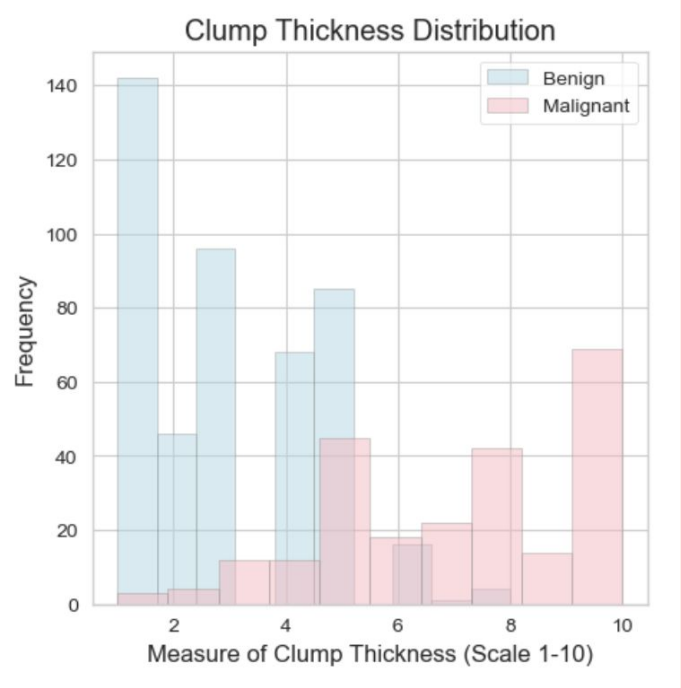


OUR DATASET IN NUMBERS

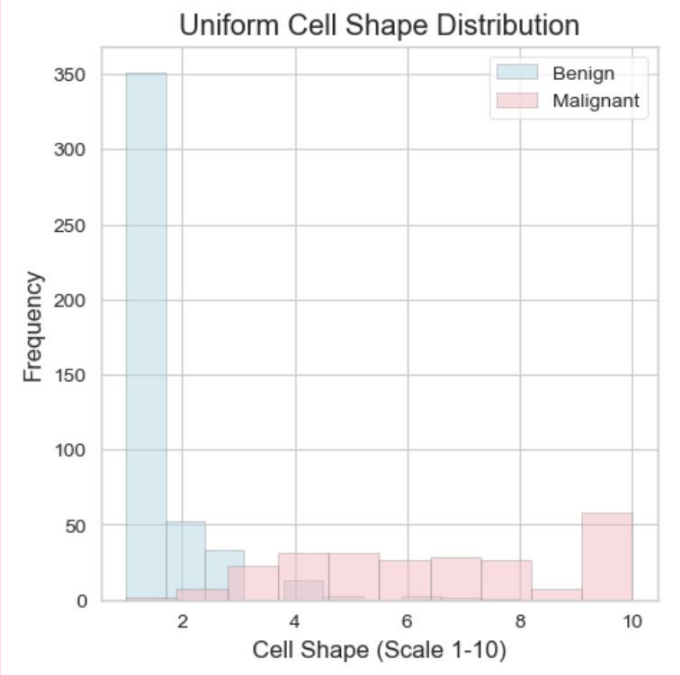


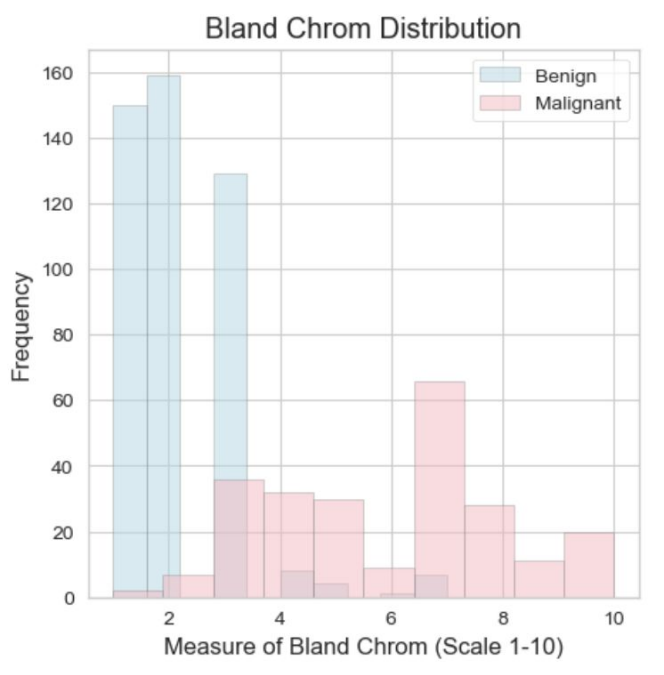
DATASET VISUALIZATION





DATASET VISUALIZATION





FEATURE ENGINEERING

DATA CLEANING

Changing feature values to explicitly indicate ‘benign’ and ‘malignant’

IMPUTATION

2.3% of samples contained missing data

REPEAT PATIENTS

50 patients were repeat patients but with different data each time

CROSS VALIDATION

80% Train

10% Validate

10% Test

CHOOSING A MODEL

01 Logistic Regression

{‘C’: [0.01, 1, 100]}

02 SVM

{'kernel': ['linear', 'rbf'], 'C': [0.01, 1, 100], 'class_weight': [{'Benign': 0.5, 'Malignant': 1}, {'Benign': 0.75, 'Malignant': 1}, {'Benign': 1, 'Malignant': 1}]}

03 kNN

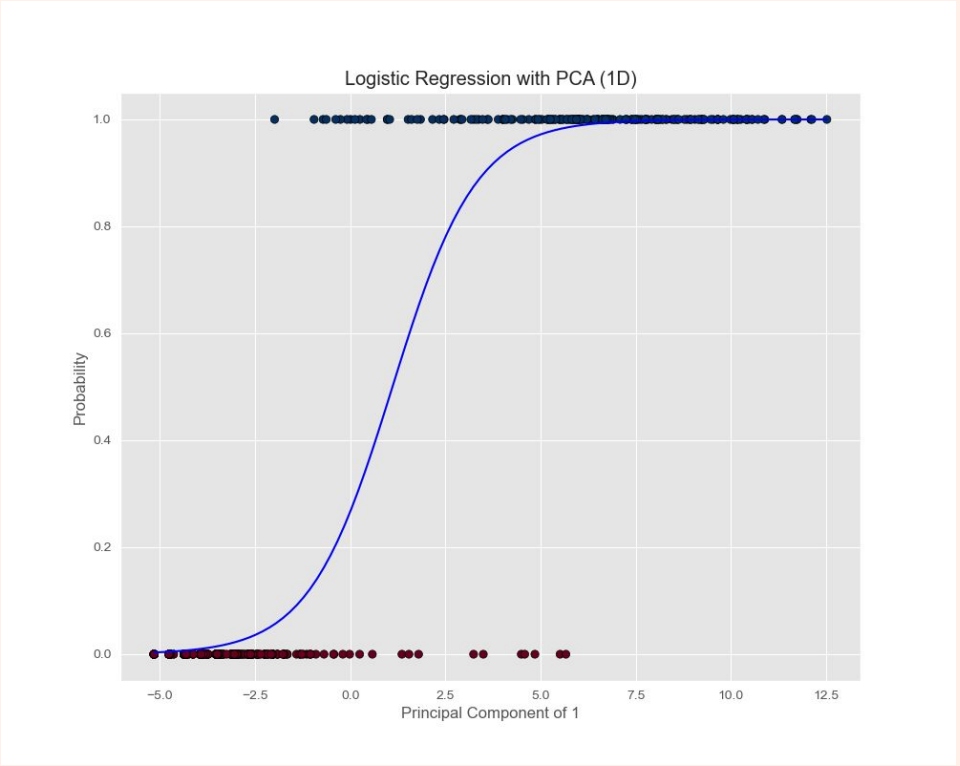
{'n_neighbors': [1, 2, 3, 4]}

04 Decision Tree

{'max_depth': [1, 3, 5, 7]}

FEATURE SELECTION

Logistic Regression (PCA1D)



CONFUSION MATRIX

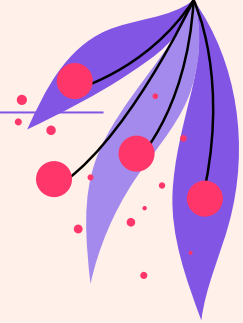
Confusion Matrix: Logistic Regression on Test Data

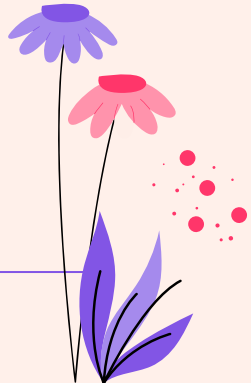
	Predicted Malignant	Predicted Benign
Actually Malignant	True Positive: 23	False Negative: 1
Actually Benign	False Positive: 0	True Negative: 46

	Metric	Performance
0	Precision	1.000000
1	Recall	0.958333
2	Accuracy	0.985714
3	AUC	0.979167

98.5% Accuracy

on testing data





CONCLUSION

- Logistic Regression is best model
- Clump thickness, bare nuclei, bland chromatin, and uniform cell shape together are strongest indicators of cell malignancy

Sources

O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

Wolberg, WIlliam. (1992). Breast Cancer Wisconsin (Original). UCI Machine Learning Repository. https://doi.org/10.24432/C5HP4Z.