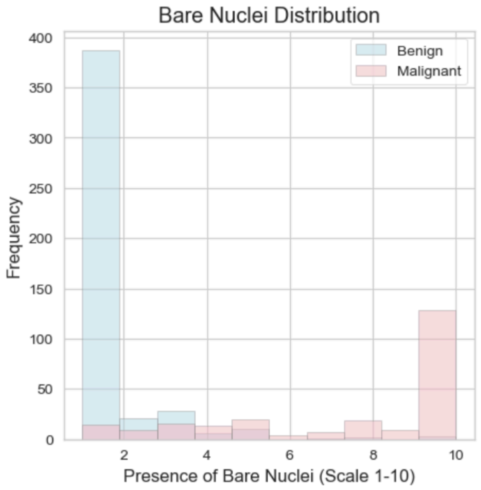
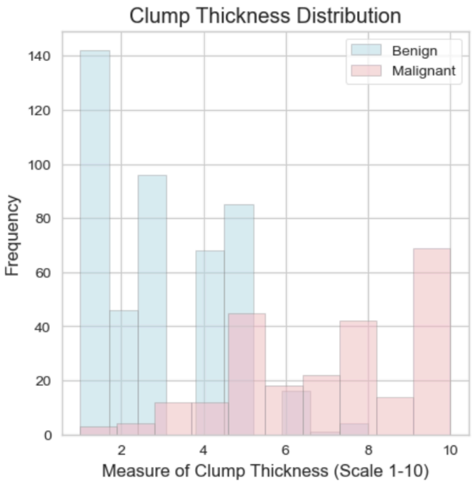
Breast Cancer Indicators

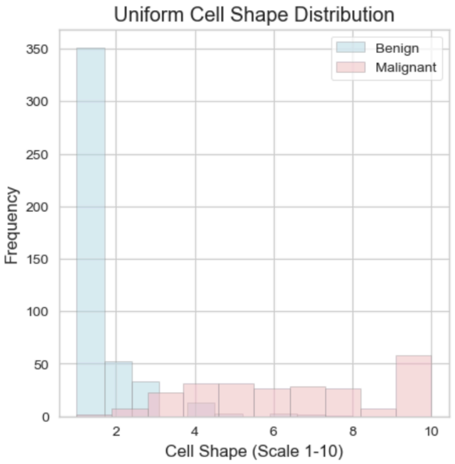
Michelle Kenton, Chelsey Severson, Alexandra White, Randy Wong

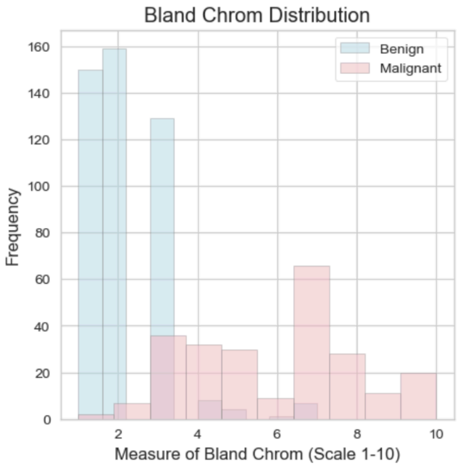
Our project uses machine learning to predict cell malignancy in breast tissue in females. Our dataset contains physical traits of cells collected from breast tissue biopsies and was created by Dr. William H. Wolberg at the University of Wisconsin Hospitals in Madison between 1989 and 1991. We sought to answer two main questions: what model is best at predicting cell malignancy and which features are most reliable in predicting cell malignancy? We used Grid Search techniques to determine that logistic regression is best at predicting cell malignancy. Next, we used permutation importance testing to determine the most reliable indicators of malignancy. Finally, using our selected features and optimized model, we achieved a malignancy prediction accuracy of .98.

The dataset comprises variables including clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, and mitosis. The class variable signifies the test result of the biopsy, with 2 indicating benign and 4 indicating malignant. All other variables are graded on a scale from 1 to 10, and together create the "nine standard technical indicators" of breast cancer. In total, there are 699 samples, with 34.5% designated as malignant cells. 2.3% of samples contained missing data, which were replaced with the mean value of the respective feature. It is important to note that some patients underwent multiple biopsies. However, as the analysis focused on cellular-level data, and the included data varied between biopsies from the same patient, all data points were retained. Below are graphs describing the distribution of the predictors categorized by their class status.



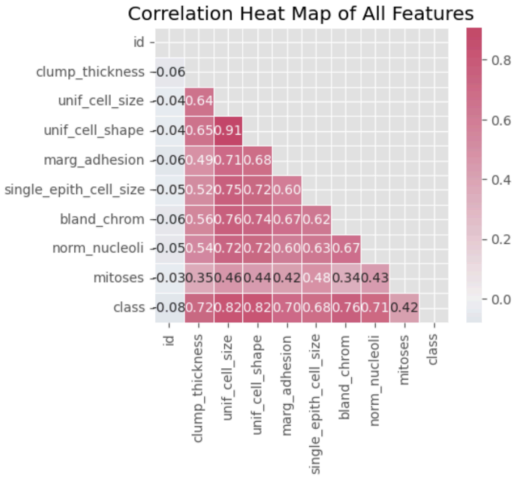


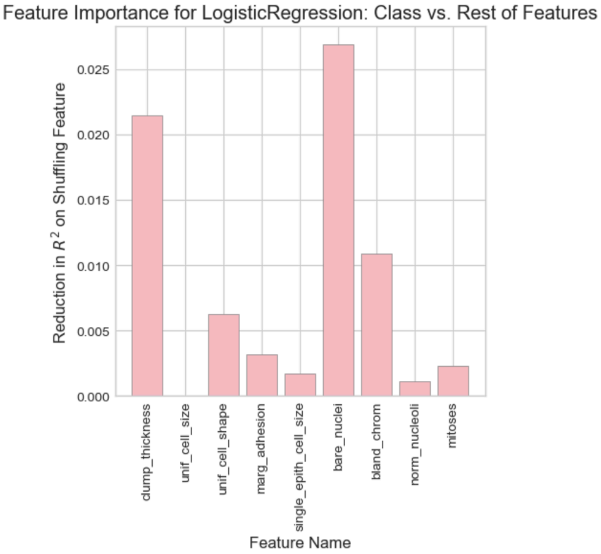




Prior to training models, we split our data into training, validation, and testing data, with 80% of samples becoming training data, 10% becoming validation data, and the last 10% was reserved for testing. We then performed a Grid Search to determine the best model for predicting malignancy. The logistic regression model with a small C-value (C=.01) outperformed the RBF SVM model, linear SVM models, decision tree models, and the K-NN classification models. It had an accuracy score of .98 on the validation data. The low c-value of our model accounts for the issue of over-fitting.

We next moved to feature selection. Below is a correlation heatmap of all our features. This graph shows a high correlation between uniformity of cell size and uniformity of cell shape, indicating that including both of these features in our model may not be necessary. In addition, we used a permutation importance test on our validation data to show us the reduction in R-squared values across models when features were excluded. This graph showed us the four most influential features were bare nuclei, clump thickness, bland chromatin, and uniform cell shape. We decided to only use the top four most influential features because the other features showed very small reductions in R-squared values when they were shuffled out of the model. In addition, by limiting the number of features used in our final model, we avoided overfitting our model.





Finally, we scored our fully optimized model on the testing data. This resulted in an accuracy score of .98. While this is a promising test result, it is important to also consider the limitations of our model. Our dataset was relatively small and dated. This could impact the implications of using our model to make claims about what factors influence breast cell malignancy today due to changes in biopsy procedures and how the studied features are measured. In addition, our model only predicts *cell* malignancy, not patient prognosis. There are many other factors and systems within the body that influence a breast cancer diagnosis (such as lymph node metastasis).

We began this project with two questions: what machine learning model is best at predicting cell malignancy and which of the nine standard technical indicators of breast cancer are most reliable in predicting cell malignancy? We found that a logistic regression with a small C-value using the features bare nuclei, clump thickness, bland chromatin, and uniform cell shape produced the highest accuracy rating of 98%. Although there are limitations to our model, we believe that it has potential to assist in the process of predicting and diagnosing breast cancer in females.

Member	Proposal	Coding	Presentation	Report
Michelle Kenton	1	1	1	1
Chelsey Severson	1	1	1	1
Alexandra White	1	1	1	1
Randy Wong	1	1	1	1

Notes:

- In the chart above, 1 = full contribution, 0.1-0.9 = partial contribution, 0 = no contribution.

References

Akinnuwesi, B. A., Macaulay, B. O., & Aribisala, B. S. (2020). Breast cancer risk assessment and early diagnosis using Principal Component Analysis and support vector machine techniques. *Informatics in Medicine Unlocked*, *21*, 100459. https://doi.org/10.1016/j.imu.2020.100459

Lopez, R. (2023, August 31). *Diagnose breast cancer using machine learning*. https://www.neuraldesigner.com/learning/examples/breast-cancer-diagnosis/

Wolberg, W. (1992). *Breast Cancer Wisconsin (Original)* [dataset]. [UCI Machine Learning Repository]. https://doi.org/10.24432/C5HP4Z