Using Supervised Machine Learning to Diagnose Breast Cancer from Fine Needle Aspiration Data

Joseph Aguilera
Brown University
https://github.com/JosephAguilera/data1030_FinalProject.git

Introduction: Breast cancer is still one of the leading causes of death amongst women; in fact, it is predicted that approximately 1 in 8 women in the United States will be diagnosed with the disease[1]. Therefore, there is a great incentive to develop novel therapies and diagnostics for breast cancer. In addition, it was reported that 65% of breast cancer cases are diagnosed at a localized stage[1], meaning that the cancer has not spread to other parts of the body, which is also known has metastasis. The approximate five-year survival rate when caught at this stage is 99%[1], underscoring the value and need for state-of-the-art diagnostics.

Joining the pursuit to develop impactful diagnostics, I am building a supervised machine learning pipeline that will diagnose breast cancer from biopsies retrieved from Fine Needle Aspirations (FNA). I will use the Wisconsin Diagnostic Breast Cancer (WDBC) dataset produced from biopsies of 569 potential breast cancer patients[2,5,6]. The WDBC dataset contains 32 features. The first two features contain the Patient ID and the *Diagnosis* target variable, and the subsequent 30 features describe the breast tissue morphology[2,5,6]. The entire dataset contains 18,208 data points. Importantly, the *Diagnosis* target variable contains two classes: the benign (non-cancerous) and malignant (cancerous) classes[2,5,6]. Therefore, the machine learning pipeline will produce and utilize a classification model.

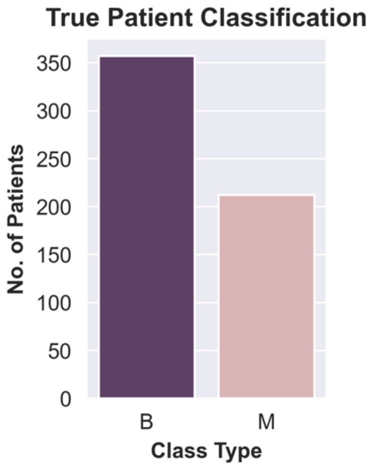

Figure 1: The total number of Benign (B) and Malignant (M) classes in the dataset is show above. Insets indicate the exact number of each class.

The FNA technique is used to collect and investigate cells residing in breast lumps or masses; a biopsy needle and ultrasound transducer are used for the extraction[2,5,6]. Once collected, immunohistochemistry staining is performed on the cells, marking the cell nuclei and other antigens of interest[2,5,6]. After staining, images are then captured and digitized. Mangasarian *et al.* probed 10 key morphological characteristics of the breast cell nuclei that are indicative of cancer; moreover, the mean, the standard error, and the mean of the top three largest values ("worst") were documented for each characteristic, producing 30 features in total[2,5,6]. The morphological nuclei characteristics include: the radius (mean of distances from center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area, smoothness (local variation in radius lengths), compactness (perimeter^2 / area - 1.0), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry, and fractal dimension ("coastline approximation" - 1) [2,5,6].

Given that the WDBC dataset has been published for approximately three decades, possesses no missing values, and contains solely continuous values in its predictive features, it has become an attractive dataset to test and optimize new machine learning algorithms for breast cancer detection. For example, at the Federal University of Rio Grande do Norte, Sizilio *et al.* previously used the WDBC dataset to develop the PDM-FNA-Fuzzy "Fuzzy" method, which has a sensitivity of 98.59% to malignant breast masses[4]. Additionally, Mohammad *et al.* performed an in-depth comparative analysis on contemporary classifiers–such J48 and K-nearest neighbor (KNN) from Naive Bayes, decision support machine (SVM), multilayer perceptron (MLP), and decision tree methods–to determine what would be the most successful one to date[3].

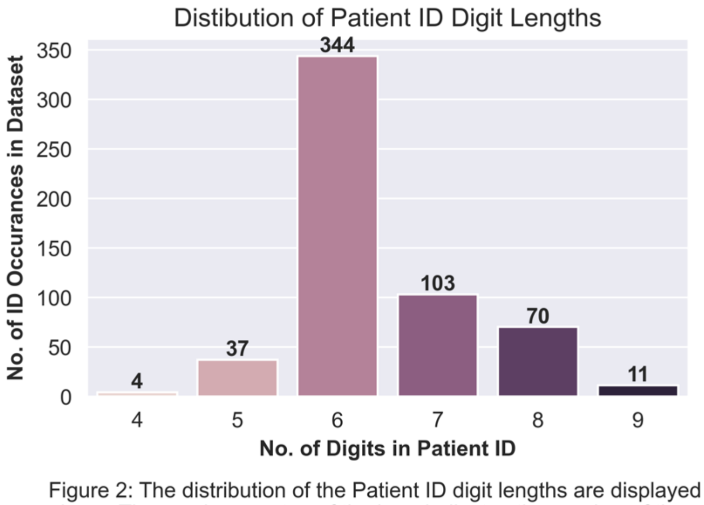

Figure 2: The distribution of the Patient ID digit lengths are displayed above. The numbers on top of the bars indicates the number of times the Patient ID digit length appeared in the dataset.

Exploratory Data Analysis: In the WDBC dataset, my first objective was to determine the characteristics of the raw dataset. The WDBC dataset is 569 by 32, indicating that there are 569 instances (patient samples) with 32 features describing them. The first two features, Patient ID (int64 Dtype) and Diagnosis (object Dtype), pertain to key facts regarding the patient. Interestingly, there is no group structure observed nor indicated in the Patient ID feature, therefore, this feature could be dropped post-EDA. The *Diagnosis* feature, which is the target variable, contains two classes: Benign (non-cancerous) and Malignant (cancerous). The distribution of the target variable classes indicates that there are more patients with the benign classification than the malignant classification (63% Benign) (Fig. 1). This gives us valuable insight to use in the subsequent steps, such

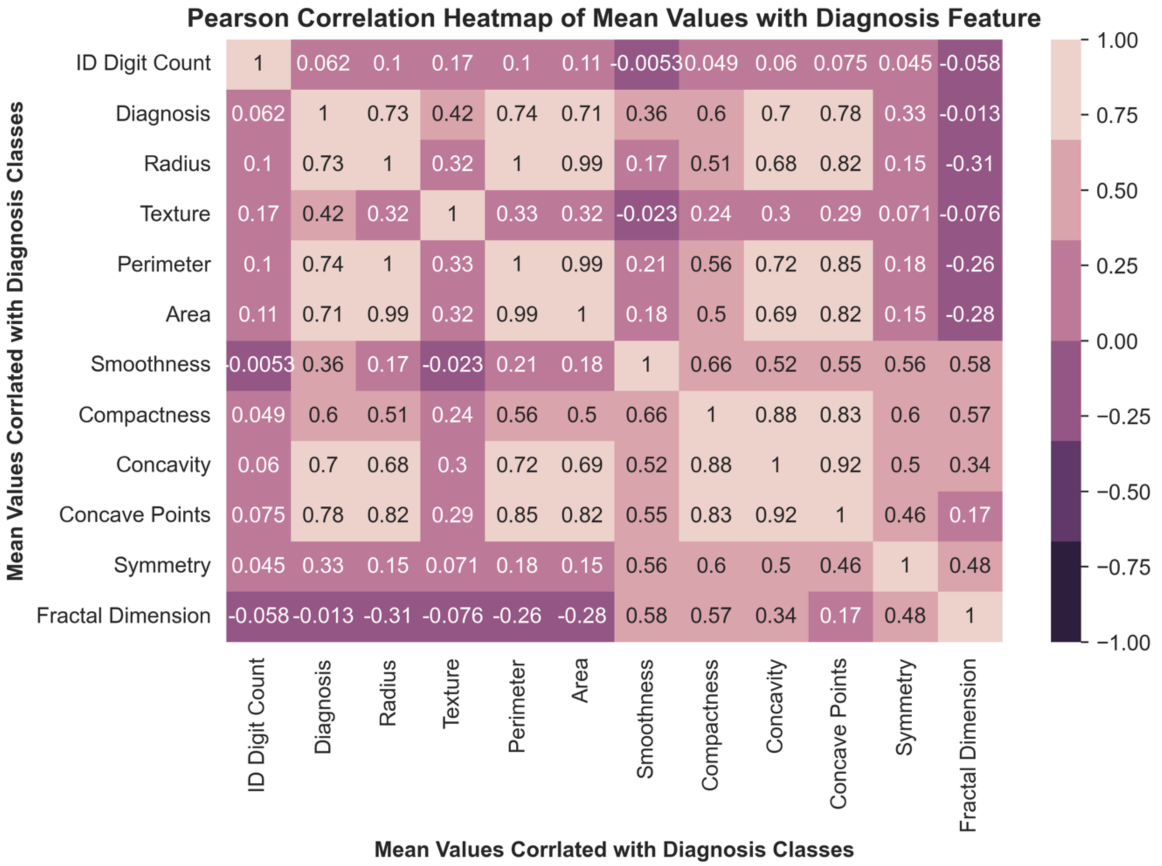

Figure 3: The Pearson Correlation was determined using the mean values in the dataset and is shown above. Specifically, the plot shows the correlation of the mean values with the Diagnosis Feature (Target Variable).

as stratifying the data upon splitting. The benign and malignant classes will indicate whether the subsequent 30 features are associated with cancerous or non-cancerous breast tissue. These 30 features (float 64Dtype) are all continues values which document the mean, standard error, and the mean of the top three largest values (referenced as “worst) for the following morphological characteristics of nuclei: Radius (µm), Texture (variance of the gray scale intensities), Perimeter (µm), Area (µm^2), Smoothness (*radius – perimeter*), Compactness (*perimeter^2/area*), Concavity (vector), Concave Points (vector), Symmetry (length difference between lines perpendicular to the major axis), and Fractal Dimension (µm).

To gain insight into how the features correlate with the target variable, I constructed a Pearson Correlation Heatmap (Fig. 3). But before constructing, I engineered and inserted an additional feature showing the Patient ID digit count because the Patient IDs differed in digit length (Fig. 2). It could be possible that patients with different digit lengths had their breast tissue biopsied in different departments or hospitals. However, there was little to no correlation shown for the ID Digit Count feature with the target variable and the morphological features. Therefore, although the Patient IDs different in digit length, the instances in the dataset should be treated as independent and identically distributed (i.i.d.).

In the Pearson Correlation Heatmap

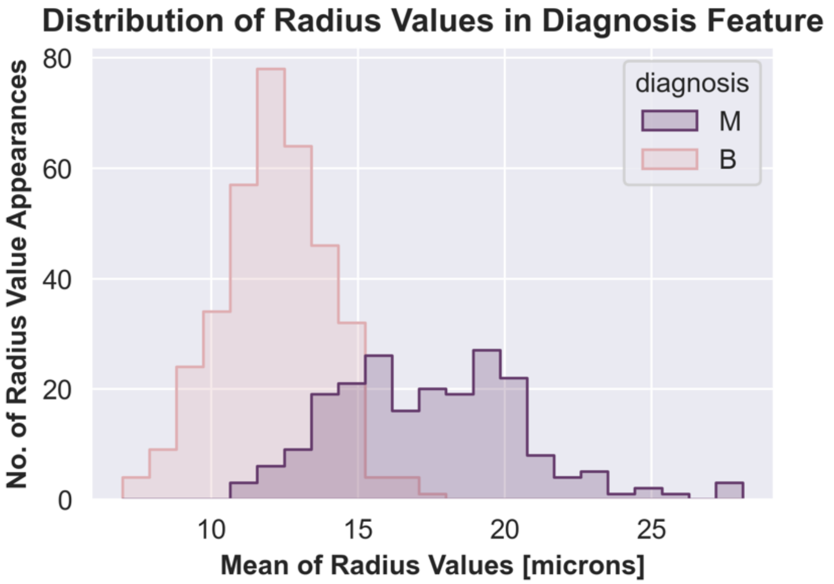

Figure 4: The distribution of the Radius Mean Values is shown above, specifically highlighting the values belonging to each diagnosis class (B: Benign, M: Malignant). The Radius Mean feature was previously shown to have a high correlation with the Diagnosis feature (target variable).

(Fig. 3), the features showing high correlation and anti-correlation were specifically documented, such as the Mean Radius, Mean Perimeter, and Mean Area features. When visualizing the features showing high correlation or anti-correlation, a clear distinction between the *Diagnosis* classes was observed (Fig. 4). Contrastingly, when visualizing the features with little to no correlation to the *Diagnosis* feature, no distinct between the *Diagnosis* classes was observed (Fig. 5). Therefore, the morphological features with a high correlation or anti-correlation with the *Diagnosis* feature will have the highest predictive power. If there are difficulties regarding computational speed and features must be dropped, the features with the most neutral Pearson Correlation Coefficients will be considered first.

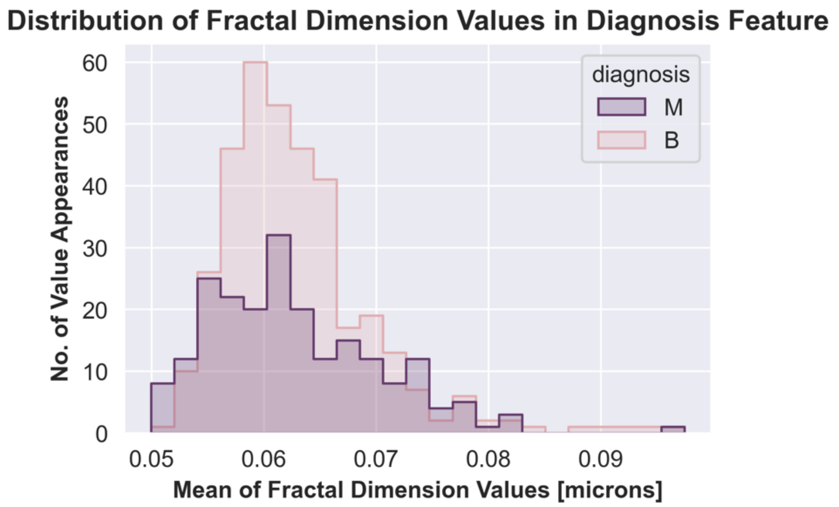

Figure 5: The distribution of the Fractal Dimensional Values is shown above, specifically highlighting the values belonging to each diagnosis class (B: Benign, M: Malignant). The Fractal dimension mean values were shown to have little corrlation with the Diagnosis feature (target variable).

Methods: To construct the best *Diagnosis* classifier, I used four machine learning algorithms, which are Logistic Regression (L1, L2, and Elastic Net), K-nearest Neighbors, Random Forest, and Support Vector Machine. The hyperparameters tuned for each machine learning algorithm are represented in Table 1.

To begin, I built a pipeline that takes in an initialized machine learning algorithm, then performs proper splitting, scaling, cross-validation, and uncertainty measurements. The pipeline's output are the appropriate test scores (detailed below), and the best models over 10 random states. First, the pipeline will the split the WDBC dataset 80:20, stratifying the target variable. The 80% split will be used in the upcoming cross-validation step, and the 20% will be stored for testing at the end of the pipeline. Then, using Sci-kit Learn's *make_pipeline* function, I initialized the *StandardScaler* preprocessor and the inputted machine learning algorithm. I solely utilized the *StandardScaler* because all 30 morphological nuclei features contained continuous values. Next, to cross-validate, I utilized the *GridSearchCV* function, which uses the initialized *make_pipeline* object (*StandardScaler*, Machine Learning Algorithm), the grid of algorithm-specific hyperparameters, an initialized *StratifiedKFold* object, and the recall scorer. The *StratifiedKFold* splitting technique will be used in the cross-validation step due to the dataset being mildly imbalanced. In addition, the *StratifiedKFold* object has 4 splits, which will split the 80% portion of the data into 20% pieces; therefore, for each model trained, 60% of the data will be used for training and 20% for validating. To train the best model, I specifically used the recall metric due to its real-world implications. Recall measures the rate of true-positives over true-positives and false-negatives. By maximizing this value, I am ensuring that those who have cancer will not be missed by the classifier, thus minimizing the false-negatives.

Machine Learning Algorithm	Hyperparameter
Logistic Regression Classifier (L1)	**C:** Inverse of regularization strength
Logistic Regression Classifier (L2)	**C:** Inverse of regularization strength
Logistic Regression Classifier (Elastic Net)	**C:** Inverse of regularization strength **L1 Ratio:** The Elastic-Net mixing parameter, Ratio of L1 Regularization **Solver:** Saga
K-nearest Neighbors Classifier	**n_neighbors:** Number of neighbors to use **weights:** Uniform or distance
Random Forest Classifier	**n_estimators:** The number of trees in the forest **max_depth:** The maximum depth of the tree
Support Vector Classifier	**C:** Regularization parameter **Gamma:** Kernel coefficient for 'rbf', 'poly' and 'sigmoid'

Table 1. The table represents all the hyperparameters that will be tuned in the machine learning pipeline.

After fitting the *GridSearchCV* object with the 80% split of the data, the best machine learning models and scalers were all saved and prepared for testing. Lastly, I used the recall metric for testing as the goal of this study is to catch every cancerous breast mass. However, I also determined the accuracy and precision of the best models to gain a better scope of the model's performance. Importantly, the pipeline includes looping through ten unique random states, which would capture the variance between different data splits. Using this method, the mean and standard deviation of the model performances were calculated and

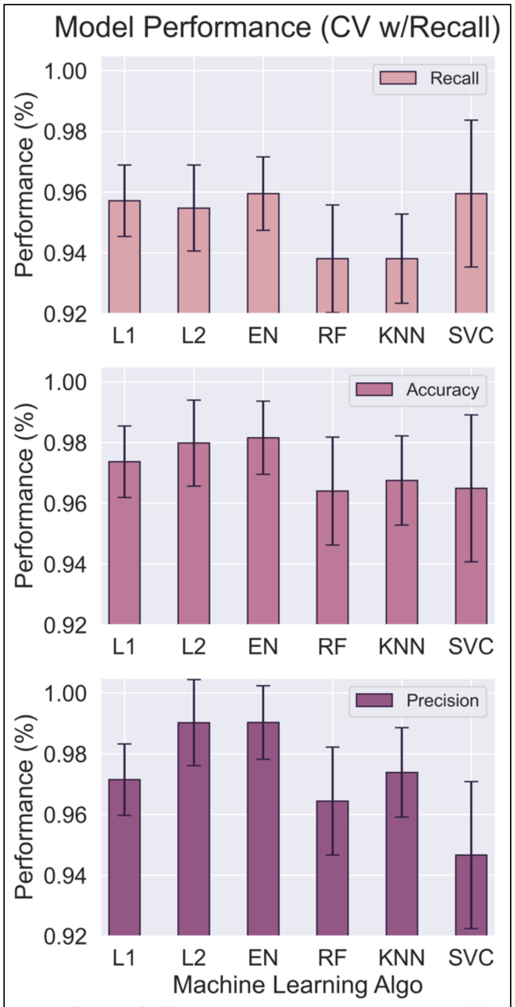

Figure 6: The plots above show how the ML models performed using Recall, Accuracy, and Precision. All the models were cross-validated using recall as the scorer.

compared.

Results: The classifier selection rationale is to select the model which has the highest recall (Fig. 6). However, when calculating the baseline performance, I was not able to use recall, since the baseline recall is zero. Therefore, the baseline metric of choice is accuracy, and the baseline performance is 63.2%.

The top two performing models in recall are the Logistic Regression (Elastic Net; EN) and the Support Vector Classifiers (SVC), respectively (Fig. 6). Both models had a recall performance of 96%, which is substantially greater than the other models used in the analysis. However, the Logistic Regression (Elastic Net; EN) model greatly outperformed the Support Vector Classification Model in accuracy and precision (Fig. 6); the model accuracy performances are 98.2% and 96.5% respectively, and the model precision performances are 99% and 94.7% respectively. Therefore, the Logistic Regression (Elastic Net; EN) model is the most predictive and will be used in the downstream analyses; additionally, this model is 28.89 standard deviations above the baseline model's accuracy, which indicates significant improvement in predictive capability. Although maximizing the recall for the Logistic Regression (Elastic Net; EN) model, there is still chance for a false-negative as indicated in the confusion matrix (Fig. 7).

Delving into the global feature importances–using the SHAP summary (Fig. 8), Perturbation Importances (Fig. 9), and the Logistic Regression Coefficients (Fig. 10)–most of the features that emerged were presented in all three methods. Additionally, the top feature "texture worst" was determined as the most important feature is all three methods. This grants us further confidence that the features present in all three methods are in fact the most important, including features such as "radius worst", "concave points worst", and "symmetry worst". Conversely, some features ("smoothness mean", "fractal dimension mean") that were the least important also had a low correlation with the target variable, as determined from the Pearson Correlation plot (Fig. 3). Then, from observing the SHAP force plots (Fig. 11) where single points in the dataset could be evaluated, the same features detected in global feature importance methods reemerge, driving the patient *Diagnosis* probabilities both up and down depending on the patient.

The most interesting finding was that the "worst" category for the nuclei characteristics was the most important. This indicates that the most exaggerated of the nuclei characteristics are most likely to be associated with cancer.

Outlook: Moving forward, I will attempt to increase the model performance by engineering features, combining features that are the most important with those that only slightly contribute. Also, the recall of the best model could potentially be improved by tuning the critical probability, shifting performance from precision to recall. The real-world input values for the 30 features can be obtained with ease, which will make constructing a larger dataset feasible. In addition to the nuclei characteristics in the feature matrix, the immunohistochemistry staining procedure reagents and concentrations should be documented, which could potentially increase model performance.

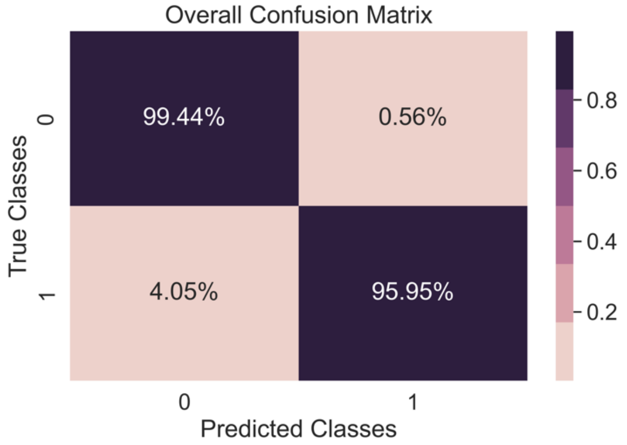

Figure 7: The plot above examines the confusion matrix for all the concatenated test set true and predicted classes. The percentages inside the plot represent the class combinations contributing to the overall confusion matrix.

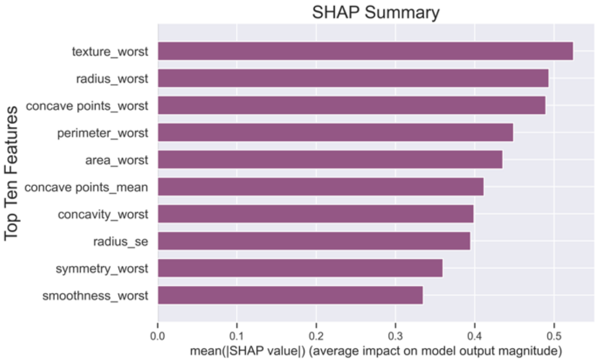

Figure 8: The SHAP summary plot above shows the top 10 features which have the most predictive power. The top feature is "texture worst".

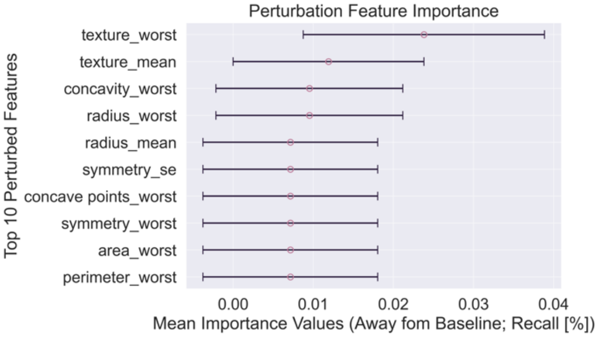

Figure 9: Using the perturbation importance method, a plot was constructed to show the features which have the most predictive power. The top ten features are displayed above. The top feature is "texture worst."

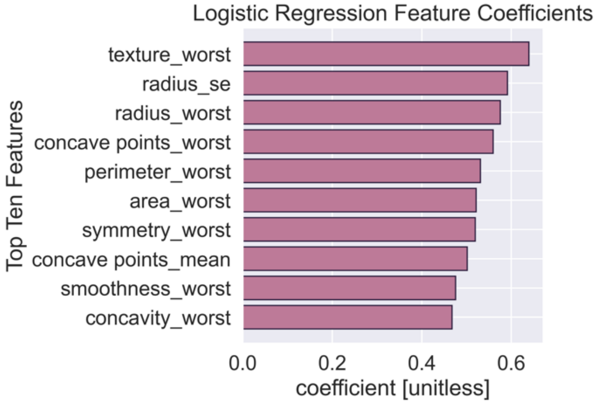

Figure 10: Using the Logistic Regression Coefficients, the top ten most important features are shown above. The top feature is "texture worst".

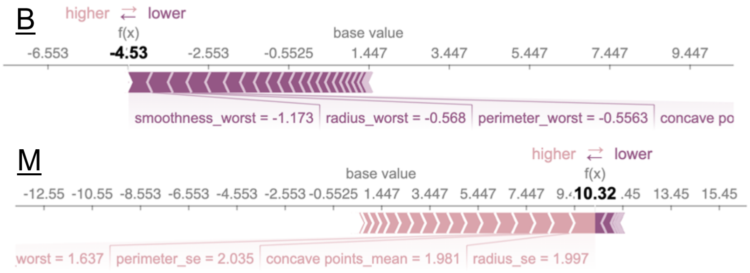

Figure 11: Two SHAP force plots are represented above. “B” represents Benign, and “M” represented malignant.

References:

[1] *Breast Cancer Facts & Statistics for 2022*. (n.d.). National Breast Cancer Foundation. Retrieved October 21, 2022, from https://www.nationalbreastcancer.org/breast-cancer-facts

[2] Mangasarian, O. L., Street, W. N., & Wolberg, W. H. (1995). Breast Cancer Diagnosis and Prognosis Via Linear Programming. *Operations Research*, *43*(4), 570–577.

[3] Mohammad, W. T., Teete, R., Al-Aaraj, H., Rubbai, Y. S. Y., & Arabyat, M. M. (2022). Diagnosis of Breast Cancer Pathology on the Wisconsin Dataset with the Help of Data Mining Classification and Clustering Techniques. *Applied Bionics and Biomechanics*, *2022*, 1–9.

[4] Sizilio, G. R., Leite, C. R., Guerreiro, A. M., & Neto, A. D. D. (2012). Fuzzy method for pre-diagnosis of breast cancer from the Fine Needle Aspirate analysis. *BioMedical Engineering OnLine*, *11*(1), 83.

[5] Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993). *Nuclear feature extraction for breast tumor diagnosis* (R. S. Acharya & D. B. Goldgof, Eds.; pp. 861–870).

[6] Wolberg, W. H., Street, W. N., & Mangasarian, O. L. (1994). Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. *Cancer Letters*, *77*(2–3), 163–171.

GitHub Repository:

https://github.com/JosephAguilera/data1030_FinalProject.git