ISyE 6740 – Summer 2020

Final Report

Team: Kyle Jones (GTID#: 903387449; GT Account: kjones374), Team ID: 37

Project Title: Predicting whether breast masses are malignant or benign

Table of Contents

Problem Statement: A physician diagnoses breast tumors as benign or malignant before beginning treatment. Malignant tumors require immediate aggressive treatment since they have cancerous cells that can spread to other tissues in the body. Benign tumors require little if any treatment because they typically remain in the tissue that they originated from (breast in this study). Due to the vastly different treatment requirements between malignant and benign tumors, it is critical that physicians properly diagnose these tumor types before beginning treatment.

Surgical biopsy is the gold standard in classifying breast tumors as benign or malignant because it has an accuracy that is close to 100%. The drawbacks of surgical biopsies however are invasiveness (requires large needle inserted into the breast), long procedure times, and costliness. A fine needle aspirate (FNA) biopsy is a less invasive and less costly procedure to diagnose tumors that involves extraction of fluid from the tumor and subsequent analysis of the nuclei within this fluid by a physician. The downside of an FNA biopsy however is poor accuracy relative to surgical biopsies, with reported accuracy metrics ranging from 65% to 98%.[3]

The dataset used for this study consisted of 569 breast tumor samples collected at University of Wisconsin hospitals.[1,2] The main goal of this analysis was to determine if machine-learning algorithms examined in ISyE 6740 could use FNA biopsy metrics to distinguish malignant vs. benign breast tumors with high accuracy. Fortunately, FNA biopsy metrics and surgical biopsy

classification were available in this dataset, where the surgical biopsy results were assumed the true classification. The second goal of this study was to identify the feature variables that were most helpful in distinguishing malignant vs. benign tumors

Feature Information: Each FNA biopsy had ten features computed for each cell nucleus within the extracted fluid. These features are listed below:

a) radius (mean of distances from the center to points on the perimeter)
b) texture (standard deviation of gray-scale values)
c) area
d) smoothness (local variation in radius lengths)
e) compactness ($perimeter^2$/area-1.0)
f) concavity (severity of concave portions of the contour)
g) concave points (number of concave portions of the contour)
h) symmetry
i) fractal dimension ("coastline approximation" – 1)

Because multiple cell nuclei existed within the fluid for an individual FNA biopsy, representative values for each of the features above were calculated with aggregate measures. Aggregate measures for each feature were as follows: mean, standard error, and worst (largest value for features a-g and smallest value for feature h).

Methodology: Three different methods were used to identify feature variables most helpful in classifying malignant vs. benign tumors. The first method used a student's T-Test and an effect size calculation to identify variables that individually were most helpful in distinguishing the two tumor types. The second method used logistic regression with a large L2 regularization penalty (Logistic Ridge Regression) to predict tumor types. Since the feature variables were scaled before fitting, the magnitude of the coefficient for a feature variable gave an indication of its relative importance to the prediction. The 3rd and final method to analyze feature importance used Principal Component Analysis (PCA) to linearly transform the thirty features to two. The magnitude of the weights of the initial Principal Components gave an indication of the amount of variance the features explained within the data, where the amount of explained variance often correlates to classification accuracy.

Four supervised machine-learning algorithms were then developed to predict benign vs. malignant tumors. The classifiers were: K Nearest Neighbor, Logistic Regression, Linear Support Vector Machine (SVM) and Random Forest. The four classifiers were trained using all features available, the top two features from the Logistic Ridge Regression analysis and the top two features using PCA. The purpose of the models that used two features for classification was to a) determine how well a simpler model could perform relative to the full feature model and b) allow for visuals that could identify patterns in the dataset and thus gain important insights.

Finally, a best classifier was determined using the full feature dataset and a best classifier was determined for the two dimensional dataset. Accuracy, negative predictive value (NPV) and positive predictive value (TPV) were determined on the test data for these classifiers to give an indication of how they are likely to generalize to future FNA biopsy data.

Results:

Importance of feature variables for classification: Tables 1-3 summarize the importance of the feature variables for classifying benign vs. malignant tumors.

Table 1 lists the average value of each feature for malignant and benign tumors. Additionally, statistical test results are presented to help identify statistically meaningful differences in feature metrics between the two tumor types. Feature variables with p values less than 0.05 and effect sizes greater than 0.8 indicate a "large" and significant difference between the two tumor types. Results are sorted by effect size, from largest to smallest.

Table 1. Basic Summary Statistics of Feature Variables between Malignant and Benign Tumors

Feature Variable	Average Value		Student's T-Test		Effect Size
	Malignant	Benign	T value	P Value	Cohen's D
concave points_worst	0.1822	0.0744	31.055	0.000	2.693
perimeter_worst	141.3703	87.0059	29.966	0.000	2.598
concave points_mean	0.088	0.0257	29.354	0.000	2.545
radius_worst	21.1348	13.3798	29.339	0.000	2.544
perimeter_mean	115.3654	78.0754	26.405	0.000	2.29
area_worst	1422.286	558.8994	25.722	0.000	2.23
radius_mean	17.4628	12.1465	25.436	0.000	2.205
area_mean	978.3764	462.7902	23.939	0.000	2.076
concavity_mean	0.1608	0.0461	23.104	0.000	2.003
concavity_worst	0.4506	0.1662	20.897	0.000	1.812
compactness_mean	0.1452	0.0801	17.698	0.000	1.535
compactness_worst	0.3748	0.1827	17.445	0.000	1.513
radius_se	0.6091	0.2841	16.396	0.000	1.422
perimeter_se	4.3239	2.0003	15.934	0.000	1.382
area_se	72.6724	21.1351	15.609	0.000	1.353
texture_worst	29.3182	23.5151	12.231	0.000	1.061
smoothness_worst	0.1448	0.125	11.067	0.000	0.96
symmetry_worst	0.3235	0.2702	10.902	0.000	0.945
texture_mean	21.6049	17.9148	10.867	0.000	0.942
concave points_se	0.0151	0.0099	10.642	0.000	0.923
smoothness_mean	0.1029	0.0925	9.146	0.000	0.793
symmetry_mean	0.1929	0.1742	8.338	0.000	0.723
fractal_dimension_worst	0.0915	0.0794	8.151	0.000	0.707
compactness_se	0.0323	0.0214	7.297	0.000	0.633
concavity_se	0.0418	0.026	6.246	0.000	0.542
fractal_dimension_se	0.0041	0.0036	1.862	0.063	0.161
symmetry_se	0.0205	0.0206	-0.155	0.877	-0.013
texture_se	1.2109	1.2204	-0.198	0.843	-0.017
fractal_dimension_mean	0.0627	0.0629	-0.306	0.76	-0.027
smoothness_se	0.0068	0.0072	-1.599	0.11	-0.139

Table 2 lists parameter estimates for logistic regression models built using an "l2" penalty. All features in the dataset were used in each model to predict the probability of a malignant tumor, where the difference in the four models was simply the value of the C hyperparameter. Scaling was performed before model fitting so that the magnitude of the parameter estimates could be compared. The features are sorted from largest to smallest parameter estimate at the C value of 0.01, which is the model that puts the highest weight on regularization.

Table 2. Ridge Regression Estimates

Feature Variable	**C Hyperparameter**			
	10	**1**	**0.1**	**0.01**
radius_worst	2.376	1.029	0.539	0.259
concave points_worst	0.953	0.912	0.525	0.258
perimeter_worst	1.657	0.823	0.493	0.25
texture_worst	2.681	1.315	0.598	0.236
concave points_mean	2.048	0.962	0.461	0.228
area_worst	2.78	1.011	0.485	0.228
radius_mean	0.651	0.363	0.39	0.227
perimeter_mean	0.582	0.351	0.38	0.225
area_mean	0.141	0.436	0.379	0.211
texture_mean	0.128	0.388	0.417	0.193
concavity_worst	1.601	0.873	0.419	0.186
smoothness_worst	0.271	0.671	0.43	0.18
concavity_mean	1.981	0.86	0.382	0.174
symmetry_worst	1.353	0.888	0.434	0.174
radius_se	2.656	1.291	0.502	0.174
area_se	2.747	1.013	0.39	0.151
perimeter_se	0.001	0.66	0.367	0.149
compactness_worst	0.691	0.045	0.141	0.14
fractal_dimension_mean	0.003	0.322	0.254	0.096
compactness_mean	2.379	0.563	0.018	0.095
smoothness_mean	0.525	0.162	0.153	0.092
fractal_dimension_se	2.425	0.681	0.265	0.084
fractal_dimension_worst	1.815	0.48	0.149	0.074
concave points_se	1.3	0.334	0.136	0.068
symmetry_mean	0.335	0.076	0.062	0.066
symmetry_se	0.508	0.296	0.149	0.044
compactness_se	0.085	0.736	0.273	0.031
concavity_se	1.07	0.111	0.045	0.021
smoothness_se	0.705	0.277	0.058	0.012
texture_se	0.842	0.269	0.048	0.012

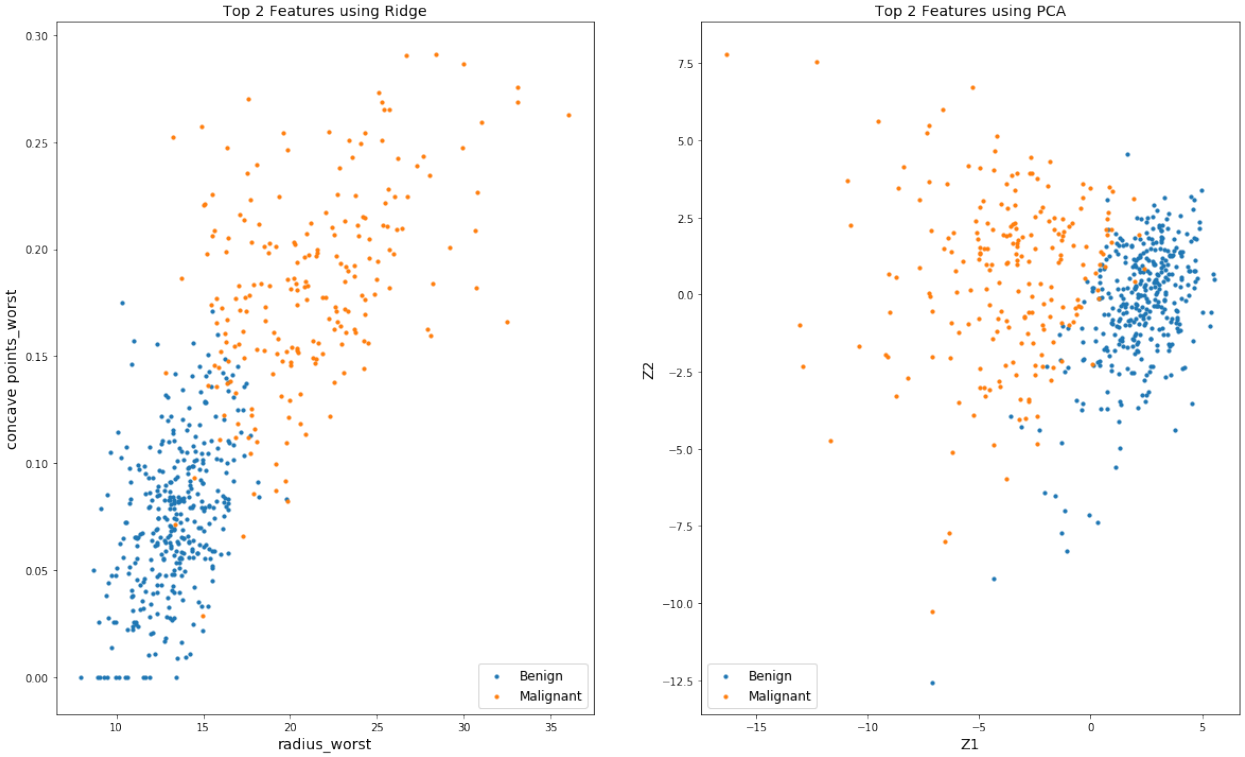
Table 3 plots the absolute value of the feature weights for the first and second Principal Components (PCs) after performing Principal Component Analysis (PCA). The magnitude of these weights gives an indication of the amount of variance the features explain within the data, which may correlate to how helpful the features are in classification.

Table 3. Feature Weights at 1st and 2nd PC

Feature Variable	Weight at 1st PC	Weight at 2nd PC
concave points_mean	0.261	0.035
concavity_mean	0.258	0.06
concave points_worst	0.251	0.008
compactness_mean	0.239	0.152
perimeter_worst	0.237	0.2
concavity_worst	0.229	0.098
perimeter_mean	0.228	0.215
radius_worst	0.228	0.22
area_worst	0.225	0.219
area_mean	0.221	0.231
radius_mean	0.219	0.234
perimeter_se	0.211	0.089
compactness_worst	0.21	0.144
radius_se	0.206	0.106
area_se	0.203	0.152
concave points_se	0.183	0.13
compactness_se	0.17	0.233
concavity_se	0.154	0.197
smoothness_mean	0.143	0.186
symmetry_mean	0.138	0.19
fractal_dimension_worst	0.132	0.275
smoothness_worst	0.128	0.172
symmetry_worst	0.123	0.142
texture_mean	0.104	0.06
texture_worst	0.104	0.045
fractal_dimension_se	0.103	0.28
fractal_dimension_mean	0.064	0.367
symmetry_se	0.042	0.184
texture_se	0.017	0.09
smoothness_se	0.015	0.204

Dimensionality reduction: Reducing the thirty features in the full dataset to two features makes it possible to visualize patterns in the dataset, which can help gain important insights. Sometimes, reducing the number of dimensions can even improve the model by filtering out noise and unnecessary details. Figure 1 displays a scatter plot of the two features from Table 2 with the largest coefficients and a scatter plot of the top two features using PCA from Table 3. Each data point is color coded by its true classification.

Figure 1. Scatter plots of two-dimensional data using Ridge and PCA, disaggregated by tumor type



Machine Learning Results: The last section of the results summarizes the performance of various supervised machine-learning classifiers built to predict tumor type. These models were trained with the full and reduced datasets

Table 4 lists the accuracy scores on training and test data of different classification models. No cross validation was performed during this stage, as the purpose of this step was to simply identify the best performing classification models. Default Scikit-Learn hyperparameters were used during this stage.

Table 4. Initial classification results before fine-tuning models

Classifier	Features used to train data					
	All Features Used		Best 2 Ridge Features		Best 2 PCA Features	
	Train Accuracy	Test Accuracy	Train Accuracy	Test Accuracy	Train Accuracy	Test Accuracy
K-Nearest Neighbor	97.58%	93.86%	95.82%	92.11%	95.82%	82.46%
Logistic Regression	99.34%	97.37%	94.95%	93.86%	96.48%	83.33%
Linear SVM	99.12%	96.49%	95.38%	93.86%	96.26%	81.58%
Random Forest	99.78%	93.86%	97.36%	92.98%	97.80%	85.09%

Optimal hyperparameters for the best performing classifiers in Table 4 using 5-fold cross validation are summarized in Table 5. Logistic Regression, Linear SVM, and Random Forest classifiers were selected as the best classifiers when using all features for training due to similarity in accuracy on the train and test data (Table 4). Meanwhile, all classifiers were used from Table 4 for the dataset using the top two ridge features because each classifier had similar accuracy scores on the train and test data. Models were not built for the dataset using the top two Principal Components because of the poor performance on the test data in Table 4.

Table 5a. Optimal hyperparameters for best classifiers using all features

Classifier	Hyperparameter	Optimal Value
Logistic Regression	C	1
Linear SVM	C	0.1
Random Forest	Maximum Tree Depth	30

Table 5b. Optimal hyperparameters for best classifiers using best two ridge features

Classifier	Hyperparameter	Optimal Value
Logistic Regression	C	10
Linear SVM	C	1
Random Forest	Maximum Tree Depth	10
K-Nearest Neighbors	K	9

Two ensemble-voting classifiers were generated using the models summarized in Tables 5a and 5b to determine if the aggregated predictions could result in a better classifier than any individual classifier. The ensemble-voting method was "hard voting" whereby the tumor type that got the most votes among the individual classifier was chosen as the predicted tumor type. The results of the two ensemble-voting classifiers are summarized in Table 6.

Table 6a. Ensemble-voting classifier results using all features

	Train Data	**Test Data**
Accuracy	99.34%	97.37%

Table 6b. Ensemble-voting classifier results using best two ridge features

	Train Data	**Test Data**
Accuracy	96.26%	93.86%

The logistic regression models summarized in tables 5a and 5b were chosen as the final classification models because the accuracy metrics were virtually identical to that of the ensemble methods. Confusion matrix results are summarized for these final classifiers at two different decision thresholds: 1) the probability threshold that resulted in the highest accuracy on the train data and 2) the probability threshold that resulted in at least a 99% negative predictive value (NPV) on the train data. The second decision threshold was added because clinicians may decide that misclassifying a malignant tumor as benign is far worse than misclassifying a benign tumor as malignant. To ensure the first scenario occurs rarely, a high NPV is needed. Note that for the second probability threshold, the threshold that resulted in the highest accuracy was chosen among the thresholds that had at least a 99% NPV.

Table 7a. Confusion matrix results using logistic regression model trained on all features

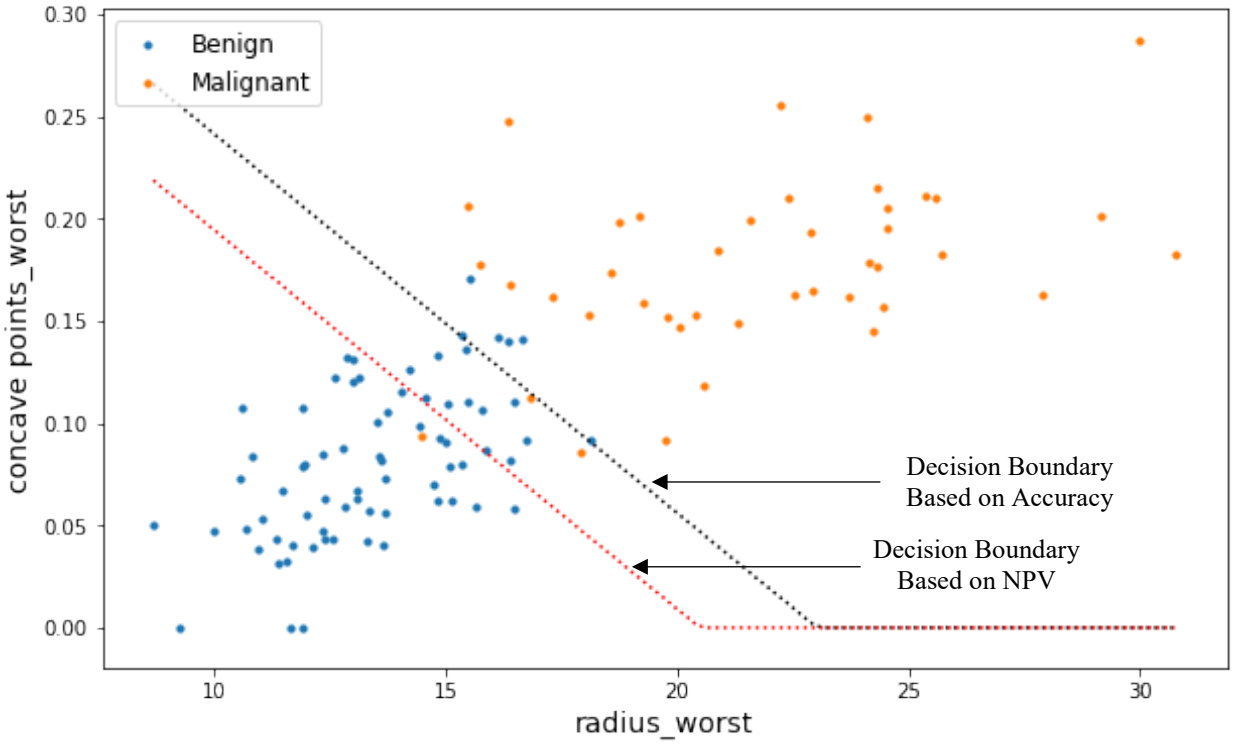
Confusion Matrix Results	**Highest Accuracy**		**Highest NPV**	
	Train Data	**Test Data**	**Train Data**	**Test Data**
Accuracy	99.34%	97.37%	97.14%	92.35%
Negative Predictive Value	98.96%	97.26%	99.28%	98.92%
True Predictive Value	100.0%	97.56%	93.85%	81.63%

Table 7b. Confusion matrix results using logistic regression model trained on Ridge features

Confusion Matrix Results	**Highest Accuracy**		**Highest NPV**	
	Train Data	**Test Data**	**Train Data**	**Test Data**
Accuracy	95.38%	92.11%	89.67%	84.21%
Negative Predictive Value	97.48%	95.65%	99.17%	98.21%
True Predictive Value	92.09%	86.67%	78.87%	70.69%

Figure 3 is a plot of the test data using the best two ridge features. The decision boundaries for the two methods of classification are shown with dashed lines. Points that fall on or above the decision boundary were classified as malignant tumors.

Figure 3. Classification results on test data using two different decision boundaries



Summary: The three methods used to identify feature variables most important in classifying benign vs malignant tumors identified “concave points_worst” as probably the most helpful. This variable was ranked first in Table 1, second in Table 2 and third in Table 3. A simple threshold method using this variable alone and the full dataset showed that classifying tumors with values of 0.144 or greater as malignant resulted in an overall accuracy of 91.74% (results not shown in report). A second important variable for classification was “radius_worst”, which was highly ranked in Tables 1-3. Because it was ranked second in the logistic ridge regression analysis for variable selection, “radius_worst” was chosen with “concave points_worst” to be part of the simpler two-dimensional dataset. A more general conclusion for Tables 1-3 was that the aggregate measure of “worst” was the most helpful in classifying tumor types whereas the aggregate measure of “se” was the least helpful.

Machine learning results for the full feature dataset and the two dimensional dataset using “concave points_worst” and “radius_worst” as predicting variables revealed that the Linear SVM and Logistic Regression models had the best testing accuracy. This is probably because the data appears mostly linearly separable even at two dimensions, as seen in Figures 1 and 3. Interestingly, an ensemble method using the results of all classifiers to make a prediction based on the classification that received the most votes revealed no improvement in accuracy relative to the individual logistic regression model (Table 6). This was true for the models trained using the full dataset and the models trained using “concave points_worst” and “radius_worst” as predicting variables. Thus, the final classifier used for the full and two-dimensional feature datasets was a logistic regression model with an optimized regularization hyperparameter.

The results of these final classifiers are shown in Table 7. In addition to accuracy, NPV and TPV metrics are shown for two decision boundaries created, whereby one decision boundary was determined at the highest training accuracy possible and one decision boundary was determined at the highest training accuracy while ensuring a TPV of 99%. This second decision boundary was included since clinicians may prefer a model with a very low chance of misclassifying a malignant tumor as benign (false negative) compared to a model with very high accuracy but a fair amount of false negatives. Results on test data showed that the logistic regression model had 97.37% accuracy (or 92.35% with high TPV decision boundary) when using all feature variables and 92.11% accuracy (or 84.21% with high TPV decision boundary) when using only “concave points_worst” and “radius_worst” as feature variables. Although the use of all features did result in better classification results, “concave points_worst” and “radius_worst” still produced decent results. Therefore, these two metrics should be measured carefully during an FNA biopsy, as they appear to be very helpful in classifying tumor types.

Future analysis could explore neural networks as a classifier since it has been shown to produce state of the art classification results in other studies. Future analysis could also explore the number of additional features that need to be included with “concave points_worst” and “radius_worst” to achieve similar accuracy as the full dimensional dataset. It is possible only a handful of additional features need to be added to achieve sufficient accuracy, which would simplify the FNA biopsy procedure, as less features would need to be measured.

References:

1) https://www.kaggle.com/uciml/breast-cancer-wisconsin-data

2) Wolberg, W. H., and O. L. Mangasarian. “Multisurface Method of Pattern Separation for Medical Diagnosis Applied to Breast Cytology.” *Proceedings of the National Academy of Sciences*, vol. 87, no. 23, 1990, pp. 9193–9196., doi:10.1073/pnas.87.23.9193.

3) R.W.M. Girard and J. Hermans. The value of aspiration cytologic examination of the breast. A statistical review of the medical literature. Cancer, 69:2104-2110, 1992.