Paper ###-2018

Predicting a Patients Breast Cancer Diagnosis Using Data Mining Techniques

Bethany Smith, Dalton Hall, Amy C Gray, Louisiana State University

# Abstract

World wide, breast cancer is the second most common form of cancer diagnosed among women, and in 2018, it is estimated that 1,735,350 new cases of cancer will be diagnosed for both men and women (American Cancer Society , 2018). Breast cancer among women accounts for an estimated 30% of all new cancer diagnoses (American Cancer Society , 2018).

The purpose of this analysis is to determine if the less costly and less invasive Fine Needle Aspirate (FNA) procedure can be used as an alternative to surgical biopsy as a diagnostic technique. Additionally, we aim to build a classification model using FNA features to diagnose patient breast mass malignancy and, thereby, diagnose patient breast cancer. Lastly, we look to identify the most significant features of the FNA procedure that aid in classification and patient diagnostic accuracy. We utilize the Wisconsin Breast Cancer Diagnostic dataset composed of 569 patients (63% benign and 37% malignant) for assessing the nuclei cell features of the FNA procedure .

In this analysis, we conduct a performance comparison between a variety of data mining techniques and machine learning algorithms: linear classification, quadratic classification, K nearest neighbor (K-NN), logistic regression, decision tree, and neural network. We assess the efficiency and effectiveness of each technique in classifying a patient’s breast mass as malignant in terms of accuracy, sensitivity, specificity, and ROC index of the validation data. Ultimately, the HP neural network is selected as the best predictive model with a 98% classification rate, 95% sensitivity, 100% specificity and 0.995 ROC index for the validation data.

# Introduction

Breast cancer is a disease caused by the uncontrolled division of abnormal cells that grow into malignant tumors found in the breast of a patient (American Cancer Society, 2018). Early detection, diagnosis, and treatment of breast cancer is vital to patient survival. When breast cancer is detected in early stages, there is an estimated 30% chance that the patient can be treated effectively. However, the longer breast cancer goes untreated, the smaller the chance of patient survival.

Currently, there are three diagnostic techniques used when screening patients for malignant breast masses: mammography, fine needle aspirate (FNA), and surgical biopsy. The mammography procedure is the least invasive, least costly procedure of the three; however, diagnostic sensitivity is known to fluctuate between 68% and 80% (American Cancer Society, 2018). Thus, mammography is the least accurate of the three procedures and is primarily used as an initial screening tool. Surgical biopsy is a diagnostic procedure in which surgery is used to remove all or part of a breast mass believed to potentially be cancerous; the mass is analyzed after surgery to determine if it is in fact cancerous or not. Marketed with the slogan “A biopsy is the only way to know for sure if it’s cancer”, surgical biopsy is the primary diagnostic tool used in the healthcare industry. Surgical biopsy, while offering nearly 100% sensitivity, is costly, highly invasive, mentally and emotionally draining, and often accompanied by an inpatient hospital stay. Fine needle aspiration (FNA) is a diagnostic procedure in which a small gauge needle is used to remove fluid directly from the patients potentially cancerous breast mass. This fluid is then stained with a colored dye and placed on a glass slide to reveal the nuclei of those cells collected. Software is used to compute various characteristics of each nuclei; these characteristics are analyzed to produce the final diagnosis of whether the breast mass is malignant or benign. The sensitivity of the FNA diagnostic procedure fluctuates between 65% and 95%; however, when used alongside mammography, FNA is estimated to have a 5% rate of false negatives.

If researchers are able to make improvements to the FNA procedure, the procedure has the potential to offer patients and physicians many benefits including: low cost, minimally invasive, reduction in time between procedure and results/diagnosis, reduction in amount of time patients diagnosed with breast cancer go before their first treatment, etc. The purpose of this analysis is to determine if the FNA procedure can be used as an alternative to surgical biopsy as a breast cancer diagnostic technique. Additionally, we look to identify the most significant characteristics of the FNA procedure that aid in classification and patient diagnostic accuracy.

# Background & Methodology

## DAta description & preparation

The dataset used in this paper is publicly available and was obtained by Dr. William H. Wolberg, a physician at the University of Wisconsin Hospital at Madison, Wisconsin, USA. The dataset contains 569 patients along with 30 characteristics obtained through the FNA diagnostic procedure. There is a total of 32 attributes per observation; including the ID and binary target variable. The target variable diagnoses whether the patients breast mass is malignant (37%) or benign (63%).

There are no missing values or duplicate observations present in the dataset. We use the Interactive Grouping Node, Weight of Evidence, within SAS Enterprise Miner to do an initial variable reduction. WOE is used to select features based on correlation with the dependent variable. The predictive power of each feature is assessed based on its Information Value. If the Information Value of a feature is greater than 0.1, the variable is chosen as an input variable. Figure 1 shows the twenty-six variables selected as inputs from the WOE procedure as well as the four variables rejected based on their Information Values.

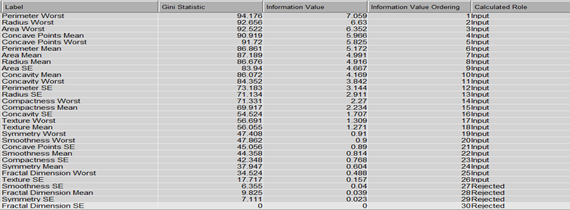


Figure 1. WOE Procedure : Calculated Role of Independent Variables

Next, the sample of 569 patients is partitioned into training and validation data sets using a 70:30 ratio. The training data is used for model fitting while validation data is used for model validation. The sample size of the training and validation samples is depicted in Table 1.

**https://lh5.googleusercontent.com/jjLNWjLkWi7EPWSTcFa2O09U6Rqtz4WMh-TFtXBKC1poq0pz6vSotW2WFY5-12w2MsfB0wKL1mMJ9snXWgoj8_AmwtnWP_tF0xRXW6jHYP4s8TbmwH_RhFPCdf_6pqM9QOdy4ZyhF_uI3zDkVg**

Table 1. The Partitioned Data for Training and Validation

## METHODology

We explore seven variable reduction techniques in SAS Enterprise Miner: variable selection, variable clustering, best variable clustering, principal components, logistic regression (forward, backward, stepwise selection), decision tree, and random forest. After variable reduction we explore nine different modeling approaches: linear classification model with variation in variable subsets and priors, quadratic classification model with variation in variable subsets and priors, K nearest neighbor with variations in K and variable subsets, logistic regression with variation in variable selection criteria (default, stepwise, backward) and variable subsets as inputs, decision tree with variation in splitting rule target criteria (default, entropy, Gini, number of branches) and variable subsets used as inputs, neural, auto neural and HP neural networks with variation in variable subsets, activation functions, hidden layers, and hidden nodes and random forest via variable selection with variation in the variable subsets used as inputs. Each variable reduction technique is used as an input for each type of model built. A total of 63 different analyses are built; this is done in an effort to achieve the highest predictive accuracy with the smallest number of predictors. We aim to find the variables that are most important in classification, and the best model, having the lowest misclassification rate, highest sensitivity, highest specificity, and highest ROC Index for the validation data.

# RESULTS

The variable selection, variable clustering, and principal component methods are the three primary variable reduction techniques utilized in the final models of this analysis. These three techniques are utilized in the final models due to the high predictive accuracy achieved with these methods reduced subsets as inputs.

First, the five variables selected from the original 26 using the variable selection, variable reduction procedure are: Concave Points Worst, Perimeter Worst, Radius SE, Smoothness Worst, and Texture mean.

Next, the variable clustering procedure reduces the 26 original WOE variables to four variable clusters. These four variable clusters are able to account for an estimated 74% of the total variation present in the data. Table 2 highlights the four variables clusters and the variables contained within each cluster. Furthermore, those variables determined to be the “best” variables within each cluster, those variables that are able to account for the most variation within each cluster, are identified in bold in Table 2 below.

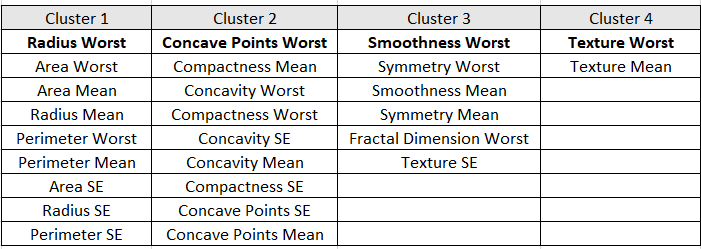


Table 2. Variable Subset Produced using Variable Clustering Procedure

Finally, the principal components, variable reduction procedure takes the 26 original WOE variables as inputs and creates orthogonal linear combinations of the original variables. Five principal components are chosen as inputs, for further analysis as they have eigenvalues greater than one. These five components are able to account for nearly 82% of the total variation. Table 3 highlights the properties of the 5 principal components.

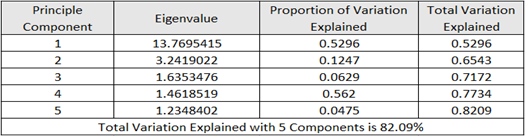


Table 3. Variation Explained from selected Principal Components

One of the main objectives of this analysis is to determine which features of the cell nuclei obtained in the fine needle aspirate (FNA) diagnostic procedure are of critical importance for correctly diagnosing patient breast masses as malignant or benign. As previously mentioned, we explore seven variable reduction techniques in SAS Enterprise Miner: variable selection, variable clustering, best variable clustering, principal components, logistic regression (forward, backward, stepwise selection), decision tree, and random forest. While the variable selection, variable clustering, and principal component variable reduction techniques are utilized in the final models of this analysis, Table 4 below displays those variables that are selected in each of the seven variable reduction procedure explored in SAS Enterprise Miner.

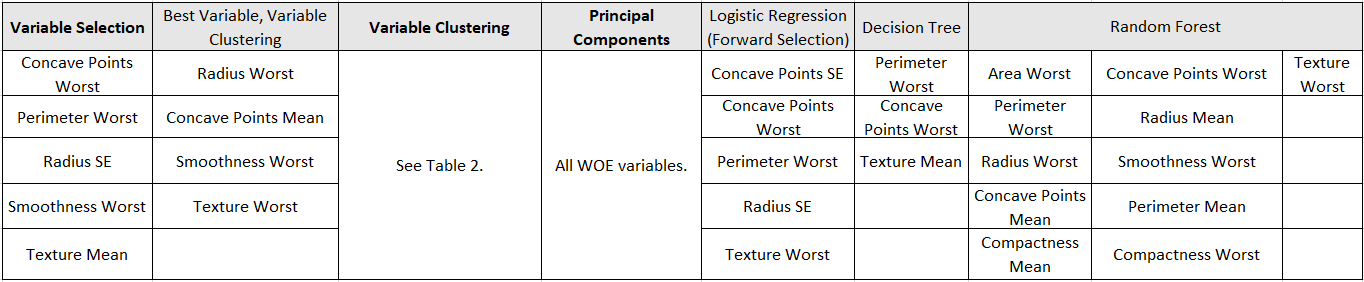


Table 4. Summary of WOE Variables selected in each Variable Reduction Procedure

Across six of the seven variable selection methods explored, Concave Points Worst and Perimeter Worst are selected as significant features for analysis. Furthermore, across five of the seven variable selection methods explored, Smoothness Worst and Texture Worst are selected as significant features to be used in analysis. Thus, to further improve outcome prediction and patient diagnosis, we recommend medical professionals consider refining those techniques used to measure and record the perimeter, concavity, texture, and smoothness of a cell’s nucleus within a potentially malignant breast mass.  
  
Once we obtain the reduced set of variable inputs, we apply the nine modeling approaches: linear classification model with variation in variable subsets and priors, quadratic classification model with variation in variable subsets and priors, K nearest neighbor with variations in K and variable subsets, logistic regression with variation in variable selection criteria (default, stepwise, backward) and variable subsets as inputs, decision tree with variation in splitting rule target criteria (default, entropy, Gini, number of branches) and variable subsets used as inputs, neural, auto neural and HP neural networks with variation in variable subsets, activation functions, hidden layers, and hidden nodes and random forest via variable selection with variation in the variable subsets used as inputs. Table 5 summarizes the Classification Rate, Sensitivity, Specificity, and ROC Index obtained for the validation data for the “best” model built for each type of model explored in this analysis along with the variable reduction technique that provides the highest predictive accuracy for that model type.

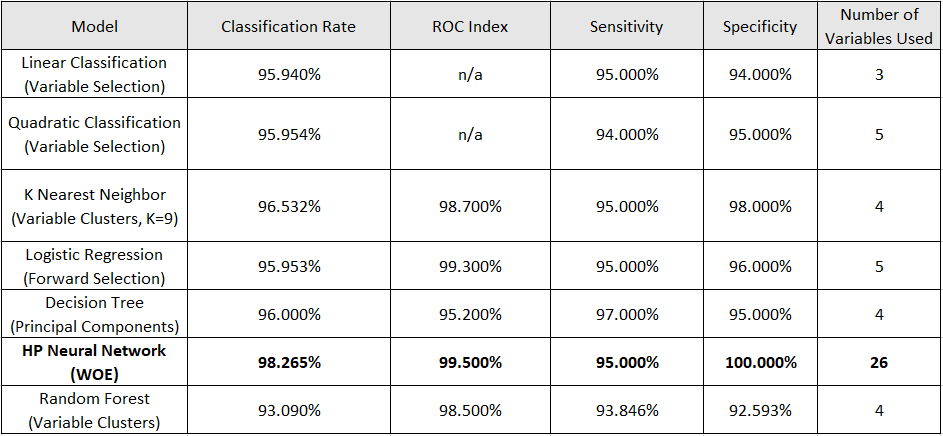


Table 5. Model Assessment Fit Statistics

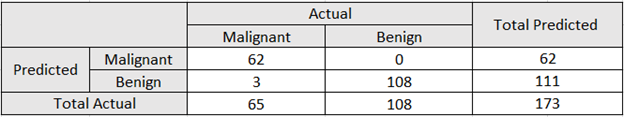
After comparing all models, the HP neural network with the 26 WOE variables as inputs is deemed to be the most diagnostically accurate model, possessing the following fit statistics: 98.265 classification rate, 99.5% ROC Index, 95% sensitivity (true positive), and 100% specificity (true negative) for the validation data. The HP neural network model surpasses all other models across all cut offs of interest under the validation ROC index criteria. Furthermore, the HP neural network outperforms all other models in terms of misclassification with a validation misclassification error rate of .01 or 99% accuracy. Table 6 displays the classification matrix of the validation data for the HP neural network.  
  


Table 6. HP Neural Network Classification Matrix

# Conclusion

The best model for predictive purposes is the HP neural network. When compared to other models the neural network has the lowest misclassification rate (.017), the highest ROC index (.995), and the highest specificity of any model (100%). However, if health care professionals were to be more interested in interpretation of results rather than prediction, we would favor the logistic regression forward selection model. That is, with the logistic regression model we are able to accurately make statements such as the following: for every 1-unit increase in the radius standard error of cell nuclei the estimated odds that a patient's breast mass is malignant increase by 22%. The HP neural network provides physicians and patients increased levels of diagnostic accuracy with the use of the less invasive, less costly FNA procedure. Additionally, with the use of the HP neural network patients and physicians will see a reduction in time between accurate diagnosis of patients and medical intervention as well as a reduction in the number of patients who receive both false positives and false negatives. The neural network model provides patients with a diagnostically accurate alternative to the surgical biopsy procedure. In an effort to further improve outcome prediction, moving forward, we recommend medical professionals consider refining those FNA techniques used to measure and record the perimeter, concavity, texture, and smoothness of a cell’s nucleus within a potentially malignant breast mass.

# References

American Cancer Society . (2018, January 5). *Facts & Figures 2018 : Rate of Deaths From Cancer Continues Decline*. Retrieved from American Cancer Society : https://www.cancer.org/latest-news/facts-and-figures-2018-rate-of-deaths-from-cancer-continues-decline.html.

American Cancer Society. (2017, September 12). *Fine Needle Aspiration Biopsy of the Breast*. Retrieved from American Cancer Society: https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/breast-biopsy/fine-needle-aspiration-biopsy-of-the-breast.html.

American Cancer Society. (2017, September 12). *Surgical Breast Biopsy*. Retrieved from American Cancer Society: https://www.cancer.org/cancer/breast-cancer/screening-tests-and-early-detection/breast-biopsy/surgical-breast-biopsy.html.

National Breast Cancer Foundation Inc. . (2017, October 11). *Biopsy*. Retrieved from National Breast Cancer Foundation Inc. : http://www.nationalbreastcancer.org/breast-cancer-biopsy.

National Breast Cancer Foundation Inc. . (2017, October 8). *Mammogram*. Retrieved from National Breast Cancer Foundation Inc. : http://www.nationalbreastcancer.org/diagnostic-mammogram.

Wolberg, D. W. (2017, August 24). *Breast Cancer Wisconsin (diagnostic) Data Set*. Retrieved from UCI Machine Learning Repository: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic).

# Acknowledgments

We would like to thank Dr. Joni Shreve, of Louisiana State University for her guidance throughout this analysis and for her unwavering support of our team and the LSU, Master of Science in Analytics Program.

# Contact Information

Your comments and questions are valued and encouraged. Contact the authors at:

Bethany Smith

Louisiana State University

337.308.3716

[Bsmi168@lsu.edu](mailto:Bsmi168@lsu.edu)

Dalton Hall

Louisiana State University

Dhall21@lsu.edu

Amy C. Gray

Louisiana State University

Agray44@lsu.edu

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration.

Other brand and product names are trademarks of their respective companies.