UNDERSTANDING BREAST MASSES THROUGH DATA MINING AND MACHINE LEARNING

GERALD GAITOS

UNDERSTANDING BREAST MASSES THROUGH DATA MINING AND MACHINE LEARNING

INTRODUCTION

The project's primary purpose is to apply different classification algorithms to the mammographic mass dataset by Schulz-Wendtland (Elter et al., 2007) to comprehend the complexities of malignant and benign breast masses and to create a model that can accurately predict the presence of malignant breast masses. In addition, the goal is to identify the critical characteristics that differentiate malignant breast masses from benign breast masses.

This project intends to provide information about the utility of mammography in differentiating malignant breast masses from benign breast masses, which will help healthcare providers, researchers, and patients understand breast mass behaviors. This project also presents the mammographic mass dataset, the method and the classification algorithms used, the result and discussion, and the references.

BACKGROUND

Breast mass is among the most common breast disorders seen in women and men. It is usually due to the abnormal growth of breast tissue cells, resulting in an abnormal lump or mass. A breast mass can either be a noncancerous or cancerous type. A noncancerous type of breast mass does not spread outside the region where it emerges, while a cancerous type spreads outside the region where it appears. The dissemination process can be through the lymphatic or vascular systems. Noncancerous or benign breast masses are usually treatable or resectable, while cancerous or malignant breast masses may require many surgical, medical, and radiological treatments (Dimagno et al., 2013).

Breast cancer, or a malignant breast mass, is the leading cancer type with the second-highest mortality rate among women (Nazari & Mukherjee, 2018). Usually, it presents with a tender, palpable breast mass, but not always; some can also initially show without a palpable mass. Hence, it is essential to perform different screening tests, such as mammography, to detect the possible presence of cancer cells (Dimagno et al., 2013).

This project focuses on understanding breast masses through the mammographic mass dataset by Schulz-Wendtland (Elter et al., 2007). Through data mining, characteristics of malignant breast masses can be identified and determined from benign breast masses, which can help in future predictions of breast cancers and the early treatment of patients likely to have breast cancer.

PROJECT OBJECTIVES

The project aims to comprehend the behavior of malignant and benign breast masses through the mammographic mass dataset by Schulz-Wendtland (Elter et al., 2007) and to create a model that can accurately predict and differentiate malignant breast masses from benign breast masses. Furthermore, this project aims to help healthcare providers decide whether to initialize early intervention against malignant breast masses that will help improve the quality of patients with malignant breast masses.

The project intends to answer the following questions:

1. Is there a difference in the characteristics of malignant and benign breast masses regarding their respective mammographic data?
2. Which attributes significantly contribute to the behavior of malignant breast masses? And the behavior of benign breast masses?
3. Can the mammographic mass dataset provide an accurate model predicting malignant breast masses in patients with breast mass?
4. Among the classification algorithms, which one will perform the best in providing the best accuracy and prediction?

METHODS

TASK-RELEVANT DATA

This project utilized the mammographic mass dataset by Dr. Rüdiger Schulz-Wendtland (Elter et al., 2007), which he obtained from Fraunhofer Institute for Integrated Circuits (IIS) Image Processing and Medical Engineering Department (BMT) in 2007. This dataset comprised 961 instances with six attributes in CSV format. The attributes and their descriptions can be seen in the Appendix.

TOOLS

This project utilized the R language to perform statistical analysis, create prediction models, and determine each model's accuracy based on the classification algorithms: classification tree, artificial neural network, support vector machine, and K nearest neighbors.

DATA PREPROCESSING

The data quality assessment of the mammographic mass dataset showed 162 missing values across attributes. The attributes were also converted from discrete to continuous data types to be able to account for statistical analysis. All 162 missing values were filled with the mean values of their respective attributes. The dataset was partitioned randomly to 70% training and 30% test sets, then subjected to four machine learning algorithms: classification tree, artificial neural networks, support vector machine, and K nearest neighbors.

DATA ANALYTICS METHODS

This project used data cleaning, mining, modeling, and visualization to extract critical information and create compelling and accurate predicting models. Statistical tools (correlation matrix, multiple linear regression, and ANOVA) were applied to determine the difference and significance in the characteristics of malignant and benign breast masses and how the attributes contributed to such behaviors. Moreover, to identify which of the predictive models, i.e., classification tree, artificial neural networks, support vector machine, and K nearest neighbors, performed the best, the respective predictive accuracy scores were tabulated and compared. The codes utilized in this project can be seen in the appendix.

RESULTS

The mammographic dataset was subjected to the correlation matrix, multiple linear regression, and ANOVA to determine the association of the variables, i.e., age, BI-RADS, shape, margin, and

density, to the severity of the breast masses. The results of these statistical tools are shown in Figures 1, 2, and 3.

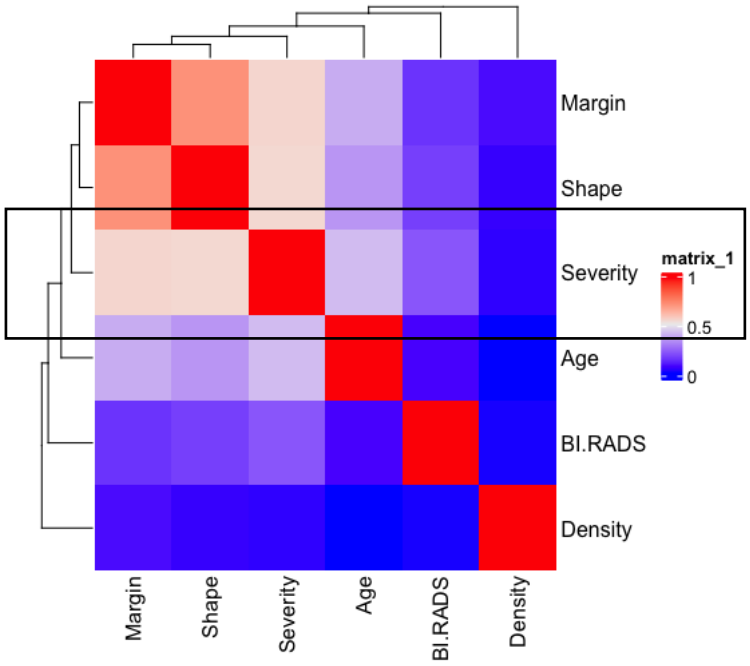
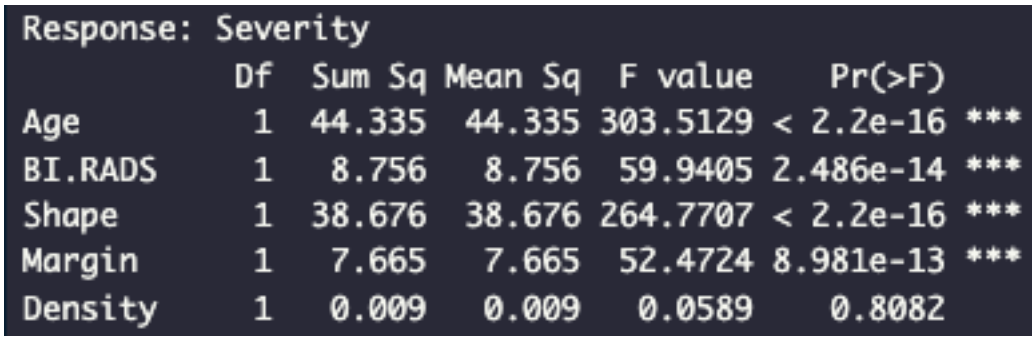


Figure 1. Heat Map representing the correlation matrix of the six attributes from the mammographic mass dataset. The row enclosed in the black box is the focus of the study. Among the five attributes, the Margin and Shape show the highest correlation coefficient (in the shade of red squares), with values of 0.55434900 and 0.5612300, respectively, correlating to the attribute Severity. The results suggest that if the breast mass observed through mammography has an irregular shape and ill-defined margin, the breast mass will most likely fall into the malignant class.

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.6580053  0.1117363  -5.889 5.38e-09 ***
Age          0.0075362  0.0009392   8.024 2.98e-15 ***
BI.RADS      0.0333176  0.0070567   4.721 2.69e-06 ***
Shape        0.1076578  0.0146727   7.337 4.65e-13 ***
Margin       0.0861963  0.0119599   7.207 1.16e-12 ***
Density      0.0082473  0.0339719   0.243    0.808
Residual standard error: 0.3822 on 955 degrees of freedom
Multiple R-squared:  0.4162,    Adjusted R-squared:  0.4131
F-statistic: 136.2 on 5 and 955 DF,  p-value: < 2.2e-16
```

Figure 2. Multiple linear regression results. This figure shows that four out of five attributes, namely, Age, BI-RADS, Shape, and Margin, are significantly associated with attribute Severity. The attribute Age has the lowest p-value, 2.98e-15, which entails that the attribute Age has a significant correlation to the attribute Severity. This finding suggests that Age is a significant

factor in associating whether the breast mass observed through mammography is malignant or benign. The figure also demonstrates the R-squared and the adjusted R-squared values, which are 0.4162 and 0.4131. Surprisingly, the values are low, which questions the actual association of the five attributes to the attribute Severity.



```
Response: Severity
           Df  Sum Sq Mean Sq  F value    Pr(>F)
Age         1  44.335  44.335 303.5129 < 2.2e-16 ***
BI.RADS     1   8.756   8.756  59.9405 2.486e-14 ***
Shape       1  38.676  38.676 264.7707 < 2.2e-16 ***
Margin      1   7.665   7.665  52.4724 8.981e-13 ***
Density     1   0.009   0.009   0.0589    0.8082
```

Figure 3. ANOVA results. This figure resembles the result from multiple linear regression results. Four out of five attributes, namely, Age, BI-RADS, Shape, and Margin, show a significant correlation with the attribute Severity. Among the four attributes, Age and Shape, with a p-value of <2.2e-16, show the highest association with the attribute Severity. This result implies that the attributes Age and Shape are crucial factors in determining the severity of the breast mass.

The mammographic mass dataset was used to create models that could be used to associate the masses with malignancy. The dataset was divided into 70% training and 30 % test sets. Different machine learning algorithms were utilized: classification tree, artificial neural networks, support vector machine, and K nearest neighbors. Table 1 shows the tabulated accuracy score results of the machine learning algorithms with modifications to optimize the predictive ability of the models.

Table 1. Accuracy Score Results

Machine Learning Algorithm	Modifications	Accuracy Score
Classification Tree		0.8373702
Artificial Neural Networks	Node = 6	0.8235294
	Node = 3	0.8269896
	Node = 9	0.8062284
Support Vector Machine	Kernel = Radial	0.7958478
	Kernel = Sigmoid	0.7508651
	Kernel = Polynomial	0.7923875
K Nearest Neighbors	K = 3	0.7958478
	K = 2	0.7543253
	K = 4	0.7923875

Table 1 shows the accuracy score results among the classification tree, artificial neural networks, support vector machine, and k nearest neighbors algorithms. The accuracy score results among all the algorithms are around 0.75 to 0.84. Moreover, Table 1 also displays different modifications applied to the three machine learning algorithms for optimization; having a node/size of three in artificial neural networks, a radial kernel in support vector machines, and a k of three in K nearest neighbors resulted in better accuracy scores. Among the machine learning algorithms, the classification tree model provided the best accuracy score among the four. Hence, the classification tree model can be used to predict correctly the attribute Severity in a given dataset.

DISCUSSION

After subjecting the mammographic mass dataset to different statistical tools, it could be interpreted that the attributes Age, BI-RADS, Shape, and Margin (having p-values of <0.05) showed significant association with the attribute Severity, while the attribute Density (having a p-value of >0.05) had a lack of evidence of association to the attribute Severity. This suggested that old-aged individuals, with high BI-RADS scores, and mammographic masses with irregular shapes and ill-defined margins were mostly at risk of getting breast masses of malignant origins.

Other studies suggested that attribute Density also played a critical role in developing breast abnormalities that were non-obligate precursors of breast malignancies. According to Boyd et al. (2011), extensive percent mammographic density (PMD) showed a significant association with invasive breast cancer. This statement is supported by Bertrand et al. (2013). Bertrand et al. found that mammographic density strongly contributed to all breast cancer types in all ages but with a great predilection to women ages <55 years. The associations observed in this paper contradicted the ones observed in the mentioned articles. One of the probable reasons why the attribute Density did not show a significant association in this study was the presence of more than 70 missing values that have resulted in favoring other attributes.

Moreover, the accuracy score of all the algorithms was around 0.75 to 0.84. Around 75 to 84 instances could be correctly predicted as having benign or malignant breast masses. Among the machine learning algorithms, the classification tree model performed the best in providing the highest accuracy score of 0.84.

CONCLUSION

In conclusion, the individual's age, BI-RADS score, and breast mass characteristics, i.e., shape and margin, would play an important role in identifying the risk of an individual having benign or malignant breast mass. Along with this, it could also be inferred that by utilizing the mammographic mass dataset, the classification tree model could be used to predict the presence of malignant and benign breast masses with an accuracy score of 0.84.

From this study, the benefits of the application of data analytics and data science in understanding huge amounts of data information could be appreciated. For future work, it would be ideal to have a larger dataset with fewer missing values for better analysis and training, and testing of machine learning models.

REFERENCES

Bertrand, K. A. et al. (2013). Mammographic density and risk of breast cancer by age and tumor characteristics. *Breast Cancer Research* 15, R104. https://doi.org/10.1186/bcr3570

Boyd, N. F., Martin, L. J., Yaffe, M. J. & Minkin, S. (2011). Mammographic density and breast cancer risk: current understanding and future prospects. *Breast Cancer Research* **13**, 223. https://doi.org/10.1186/bcr2942

Elter, M., Schulz-Wendtland, R. & Wittenberg, T. (2007). The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. *Medical Physics, 34* (11), pp. 4164-4172. Retrieved April 27, 2023, from http://archive.ics.uci.edu/ml/datasets/mammographic+mass

Dimagno, M. M. et al. (2013, June). Common Breast Problems. *UMHS Breast Problems Guidelines*. Retrieved March 14, 2023, from https://www.med.umich.edu/1info/FHP/practiceguides/breast/breast.pdf

Nazari, S. S. & Mukherjee, P. (2018). An overview of mammographic density and its association with breast cancer. *Breast Cancer* 25, 3, 259-267. https://doi.org/10.1007/s12282-018-0857-5

APPENDICES

DATA DICTIONARY

Attribute	Data Type	Unit Measurement
BI-RADS assessment	Ordinal	1 to 5
Age	Integer	years
Shape	Nominal	1 = round 2 = oval 3 = lobular 4 = irregular
Margin	Nominal	1 = circumscribed 2 = microlobulated 3 = obscured 4 = ill-defined
Density	Ordinal	1 = high 2 = iso 3 = low 4 = fat-containing
Severity	Nominal	0 = benign 1 = malignant