Breast Cancer Detection using Metabolic Parameters

Group 1

Hannah Abraham, Nada Rabbat

George Mason University

Professor Isuru Dassanayake

STAT 515 – Applied Statistics & Visualization for Analytics

May 17, 2022

Abstract

The purpose of this paper is to assess different machine learning algorithms to achieve breast cancer diagnostics prior to the disease, based on data and parameters which can be gathered during routine blood analysis. Supervised learning models are utilized to predict if a patient has breast cancer by utilizing k-nearest neighbor, support vector machines, classification trees, and logistic regression. The results of the models will be analyzed by comparing the accuracy in prediction, misclassification error, and false-negative rates using the training and testing sets.

Introduction

Breast cancer is one of the most common cancers among women.[1] Approximately 255,000 cases of breast cancer are diagnosed in women and about 2,300 in men each year, with a mortality rate of around 16.47% in women.[2] It is a disease that involves the rapid division and growth of cells in the breast. There have been claims that suggest a link between adiposity and cancer onset. However, imbalances in certain metabolic parameters (glucose, insulin, resistin...etc.) that often lead to obesity, have been suggested to be the underlying cause of cancer in the breast tissue.[3] There is also seldom stress, even societally, for men or women of a younger age to screen for breast cancer, leading many patients to go undiagnosed for years, which in turn reduces the chances of successful and permanent cancer treatment results. A tumor is dealt with best at an early stage of its onset. According to the American Cancer Society database, breast cancer survival rates decrease from 99% at early detection, to 28% when it is detected at a later stage. This statistic deems it necessary to become aware of an individual's possible breast cancer predisposition and to diagnose it as early as possible. Unfortunately, the tumors are internal and not visible, and therefore not as easily self-diagnosed. A patient will not know there is a tumor growing unless it has already become large enough to feel during a self-test, or at a routine yearly checkup, which patients usually avoid unless the unusual bodily activity is sensed, due to the high cost involved, especially for non-insured patients. However, what is more frequently conducted, and at a lower cost than a doctor's appointment, are routine blood tests. Previous works suggest there is a link between certain metabolic parameters that are retrieved during routine blood tests and cancer. A dataset was constructed such that results from the blood tests of each patient were accompanied by a patient's age data, and the results of a mammography screening diagnosing a tumor as malignant.[4] This paper will use this data to construct predictive models to classify a patient at risk of having a

Commented [NM1]: low mortality rate, might cause reader to think its not that important of a cause.

Commented [HA2R1]: do you not think that's still considered a lot, especially within a year?

Commented [NM3R1]: not really but if you think its significant we'll keep it idm

[1] https://www.kucancercenter.org/news-room/blog/2020/08/most-common-cancers-women-how-to-detect-them-early

[2] https://www.cdc.gov/cancer/breast/basic_info/index.htm#:~:text=Each%20year%20in%20the%20United,breast%20cancer%20than%20White%20women.

[3] J. Crisóstomo *et al.*, "Hyperresistinemia and metabolic dysregulation: a risky crosstalk in obese breast cancer," *Endocrine*, vol. 53, no. 2, pp. 433–442, Aug. 2016, doi: 10.1007/s12020-016-0893-x.

[4] M. Patrício *et al.*, "Using Resistin, glucose, age and BMI to predict the presence of breast cancer," *BMC Cancer*, vol. 18, no. 1, p. 29, Dec. 2018, doi: 10.1186/s12885-017-3877-1.

breast tumor. Other models performed their analysis on the intended topic by using algorithms such as Random Forest, logistic regression, and support vector machines. This paper will attempt to recreate the previously outlined models as well as a classification tree model and compare their accuracy, misclassification error scores, and false-negative rates. Additionally, age will be surveyed as an important factor in the malignancy of breast tumors, along with assessing other parameters that are statistically significant in the classification problems.

Methods

1. Dataset

The datasets for the metabolic parameters were acquired from the study conducted by Patricio et al.[5] in which they predicted breast cancer biomarkers. The data was made up of 116 female patients, 64 of which were diagnosed with breast cancer using an imaging technique, such as mammography, and 52 healthy individuals.

Commented [HA4]: can you include the source for the dataset as well? thanks!

a. ***Structure:*** Each patient's data was represented by their own row, with the features spanning the columns of the dataset. The column labels were age, BMI, glucose, insulin, HOMA, leptin, adiponectin, resistin, MCP-1, and lastly, class, which represented the healthy vs tumorous nature of the patient, with 1 representing a healthy individual (benign), and 2 corresponding to a patient with diagnosed breast cancer.

b. ***Preprocessing:*** The source data had already been preprocessed slightly prior to uploading it into the repository. Following data retrieval, it was evident that all data was ordered such that all breast cancer positive patients appeared first, and the healthy patients appeared after. To improve classification results and avoid overfitting the model to only one class, the data rows were randomly shuffled. Following the shuffling, all missing values from the set were dropped. The data was also assigned column names after loading it into R studio, and the response variables were transformed into factors for model readability. 80% of the data was assigned to training, and 20% to testing.

c. ***Summary Statistics:*** Before conducting our analysis, the dataset is explored to identify any patterns and potential issues that could impact our predictive models and skew our results. To visualize all our variables in a scatterplot without overcrowding our graph, principal component analysis (PCA) is used to plot our variables using the ggfortify library (R CRAN). Although our dataset isn't necessarily 'wide', PCA is used to find any trends present within the data, which helps us identify any clusters and groups of samples that are similar. Autoplot was used to plot PC1 and PC2 from our principal components. The output shows us that there is no relationship between the variables, especially between patients with breast cancer and healthy ones (Appendix A). There were no clusters identified in the scatterplot, which further proves that there is little to no similarity between the samples. Some outliers were present, with the majority being from the malignant samples. However,

[5] M. Patrício *et al.*, "Using Resistin, glucose, age and BMI to predict the presence of breast cancer," *BMC Cancer*, vol. 18, no. 1, p. 29, Dec. 2018, doi: 10.1186/s12885-017-3877-1.

it is not significant enough for us to remove these outliers prior to conducting the analysis. To further analyze the relationships between the variables, a correlation matrix was created to identify any correlations between age, BMI, glucose, insulin, HOMA, leptin, adiponectin, resistin, and MCP-1 (Appendix B). The matrix indicates a positive correlation between each variable and itself as expected. Although there are some correlations present among some variables, there is a strong correlation between HOMA and insulin. The correlation is reasonable as HOMA is an indication of how much insulin the body needs to keep blood sugar in check. The higher the HOMA, the more insulin the body is using to keep your blood sugar in balance (Biljana Novkovic, 2021). These relationships must be taken into consideration as having both HOMA and insulin included in our analysis could potentially influence our results depending on the type of predictive models we decided to use if the variables are proved to be statistically significant.

Commented [NM5]: "PCA is a rotation of data from one coordinate system to another. A common mistake new data scientists make is to apply PCA to non-continuous variables. While it is technically possible to use PCA on discrete variables, or categorical variables that have been one hot encoded variables, you should not. Simply put, if your variables don't belong on a coordinate plane, then do not apply PCA to them. After application, on our new coordinate system the first dimension has the maximum variance it can, then the second dimension has the most of the remaining variance it can, and so on." https://towardsdatascience.com/pca-is-not-feature-selection-3344fb764ae6

Commented [NM6R5]: our data is one hot encoded data, so this applies to us as well. keeping this here for reference.

2. **Prediction Models:**

   a. **k-Nearest Neighbor:** Since metabolic parameters vary, even if slightly, from person to person, cancer nature should theoretically depend on the closeness of metabolic parameters of a test patient to the known and classified set. Therefore, kNN will be tested, as it compares the Euclidean distances between the parameters of the training and testing sets, and maps out which classification occurs the most for a given optimal k.

      To find the optimal k for our model, a loop was executed such that in each iteration from k = 1 to k = 20, the error rate was plotted, as seen in **Fig. 1**. Theoretically, the optimal k is approximately the square root of the number of data points, which in our case is k = 10. Our results confirm this hypothesis, as the optimal k in the graph was around 9. After the optimal k was computed, a model was constructed using the training data, and both a confusion matrix and the misclassification error were computed.

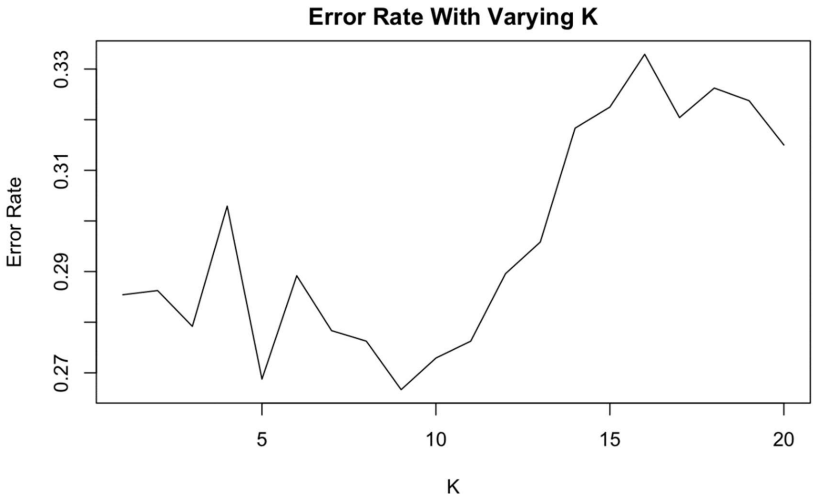


Figure 1: Error rates of our kNN model with varying k values.

b. Support Vector Machines: an SVM model was also surveyed, as it is a classification algorithm known for working well when the number of features is high with respect to the number of data points in a set. In our case, there are 9 features in a set of around a hundred datapoint. Therefore, SVM might be a good option. To construct the model, the e1071 R package was utilized, and the model was constructed and trained before the testing data was predicted. Just like in the case of kNN and all other models surveyed in this paper, a confusion matrix and the misclassification error were computed.

C. **Logistic Regression:** Before we begin running our predictive analysis in the logistic regression model, it's important to verify whether the data is linearly separable. The boxTidwell test from the 'car' package is used to check for the linearity between the logit and the predictor variables (Leung, 2021). All the variables in the dataset were tested to check whether the variables are linearly related to the logit of the outcome variable. According to our output, the variables are not statistically significant since the p-value is greater than 0.05, implying that the variables are linearly related to the logit of the outcome variable. Hence, the logistic regression model will be surveyed.

The variables in our dataset were fitted to identify which ones are proved to be statistically significant to our model. The summary indicated that 3 out of the 9 features are significant to our model with $p<0.05$. A chi-squared test with anova function (Mervisiano, 2021) is used to compare our first and second models (with

only 3 significant predictors). The output shows a non-significant chi-square value with a p-value of 0.26. This indication means that the second model performs as well as the first model which contains all 9 features. It will be preferable to conduct our analysis with the second model as it will provide a simpler interpretation. The model is fitted with 80% training data and 20% test data. A reasonable prediction would be to assign "Yes", indicating malignancy of breast tumors if the predicted probability of yes is greater than 50%. The classification performance is described using the confusion matrix and misclassification error.

d. **Classification Tree:** Decision trees, particularly classification trees, are constructed based on the ruling that each observation corresponds to the most commonly occurring class of training observations in the region to which it belongs (Le, 2018). Decision trees are ideal for classification problems, and their flowchart-like structure is easily interpretable by people who are not versed in machine learning or technical analytic terms. In this case, the classification tree model will be analyzed to predict the accuracy of prediction on the test data.

   The training data is fitted into the classification tree using the tree function. The initial tree was constructed with 11 terminal nodes, using 6 out of 9 features, including glucose, age, bmi, resistin, MCP.1, and adiponectin. A cross-validation approach to finding the optimum number of nodes was also implemented. As tree-based models are not considered the best-supervised learning approaches in prediction accuracy, a random forest model will be surveyed and discussed next.

e. **Random Forest:** Random Forest models have the advantage of increasing prediction accuracy and preventing overfitting. As the steps to create the train and test data have been completed while preparing for the classification tree model, we began by fitting the training data on the random forest model with mtry = 2. A confusion matrix is computed along with the misclassification rate. The variable importance plot (Appendix C) is created to identify the features that have the largest influence on the prediction accuracy of the model. The mean decrease in Gini impurity is used to measure how each variable contributes to the nodes and leaves in the random forest model. The higher the mean decrease Gini, the higher the importance of the feature in the model.

Results

a. **Models:** All prediction models were completed successfully, with the confusion matrices and misclassification error computed. The results can be seen in Table 1.

Prediction Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	Most significant feature	Age importance (p-value)	Misclassification Error (%)	False Negative (%)
kNN	79.1	80.4	78.1	glucose	0.135	~ 40.0	21.9
SVM	79.1	74.5	82.8	glucose	-	14.8	17.2

Commented [HA7]: This formula is just for reference: FN rate = FN/(FN+TP)

Logistic Regression	77.4	80.0	76.2	glucose	0.135	22.6	20.0
Classification Tree	86.1	80.4	90.6	glucose	-	13.9	19.6
Random Forest	94.8	94.1	95.3	glucose	-	19.4	5.9

Table 1: Prediction model metric results from the confusion matrices, misclassification errors, and variable importance analyses.

b. **Age as a marker for cancer:** In other words, descriptive statistics results indicated that age had a p-value (~ 0.5) well over the value that deems a feature statistically significant (< 0.05). In our models, the same results were replicated with almost all the models, where age had a higher p-value deeming it not as statically significant as social constructs imply. These p-values can be seen in **Table 1**.

c. **Variable importance in cancer proliferation:** A variable importance analysis showed that instead of age being a marker for cancer onset, glucose levels were more influential, across all the models' descriptive statistics methods, as seen in **Table 1**. Results from the original paper indicate that glucose, resistin, HOMA, and insulin, in order, were the most statistically significant in this classification problem. In our case, results were similar, with glucose consistently being the most statistically significant variable. Even in the Random Forest Gini impurity analysis, where results differed slightly for other variables when compared to previous works, glucose still had the highest mean impurity decrease, followed by BMI, age, and resistin. This could be further supported by a research paper on pre-diagnosis blood glucose and prognosis in women with breast cancer, which indicated that patients with elevated random blood glucose levels have significantly shorter overall survival and time to tumor recurrence (Monzavi-Karbassi, 2016).

d. **Model Performance:** To assess our models, accuracy, sensitivity, specificity, misclassification error, and false-negative ratios were computed. The most key factors to consider in the cases of diagnosis are lower false-negative rates, as diagnosing a patient with a false negative when they truly have the illness is much more dangerous than being falsely diagnosed, as the true positive patient will go undiagnosed and unmedicated. Random Forest was able to achieve the highest accuracy, sensitivity, and specificity, while also maintaining the lowest false-negative rate. The initial classification tree is constructed with 11 nodes, it is ideal to see if tree pruning will reduce the size of the tree using cross-validation. To find the optimal number of nodes, the cross-validation error rate is plotted (Appendix D), and it indicated that the tree with 9 nodes will have the lowest error rate, which isn't too far off from what we already have. However, using 9 nodes instead of 11 did not impact our model accuracy, as it remained at 86.1%, which means the tree cannot be improved any further. Logistic regression had lower accuracy, sensitivity, and specificity (~ 10% less than random forest), but the lowest misclassification rate of all models, with SVM's misclassification rate following shortly after, but better accuracy, specificity, and false-negative rates.

Commented [HA8]: did you run the classification tree on your end?

Commented [NM9]: if you don't mind filling these in real quick just so i can write and finish the results section.

Commented [NM10]: no results appear in the methods. these should all be moved to the results section

Conclusion

Breast cancer is one of the most prevalent cancers in women, and very often goes undiagnosed due to the high cost of mammography screenings, and the lack of outwardly visible symptoms. To remedy this, classification models were used to predict the nature of breast cancer from routine blood test data. The random forest model was able to predict the response variable with the highest accuracy and with the lowest type II error rate. These models also proved that while old age is commonly stressed as a significant factor influencing breast cancer proliferation, glucose levels are far more statistically significant. Future improvements to this work may include an increase in the number of metabolic parameters measured during blood tests, to increase prediction specificity, and increasing the number of data points to avoid overfitting.

References

Biljana Novkovic, P. (2021, January 15). *HOMA-IR: A Test of Insulin Resistance + Ways to Decrease It.* Retrieved from SelfDecode: https://labs.selfdecode.com/blog/homa-ir/

Le, J. (2018, June 18). *R Decision Trees Tutorial*. Retrieved from datacamp: https://www.datacamp.com/tutorial/decision-trees-R

Leung, K. (2021, October 4). *Assumptions of Logistic Regression, Clearly Explained*. Retrieved from Towards Data Science: https://towardsdatascience.com/assumptions-of-logistic-regression-clearly-explained-44d85a22b290

Mervisiano, M. (2021, January 2). *How to do Logistic Regression in R*. Retrieved from Towards Data Science: https://towardsdatascience.com/how-to-do-logistic-regression-in-r-456e9cfec7cd

Monzavi-Karbassi, B. G. (2016). Pre-diagnosis blood glucose and prognosis in women with breast cancer. *Cancer Metab*.

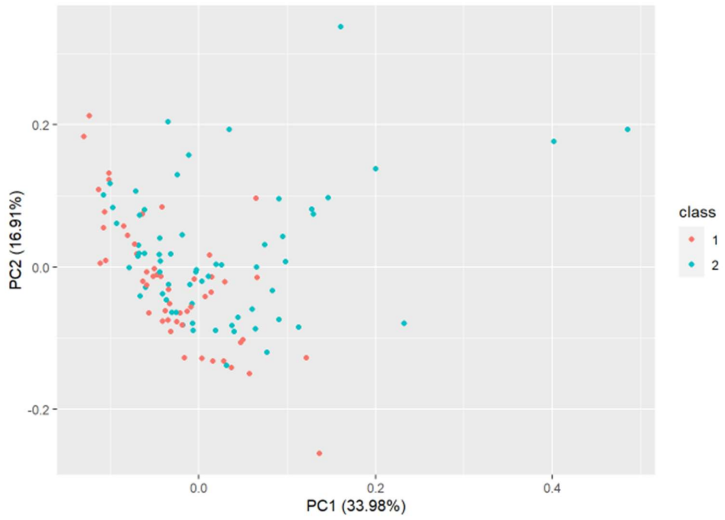
R CRAN. (n.d.). *Plotting PCA (Principal Component Analysis).* Retrieved from https://cran.r-project.org/web/packages/ggfortify/vignettes/plot_pca.html

Appendix

Appendix A

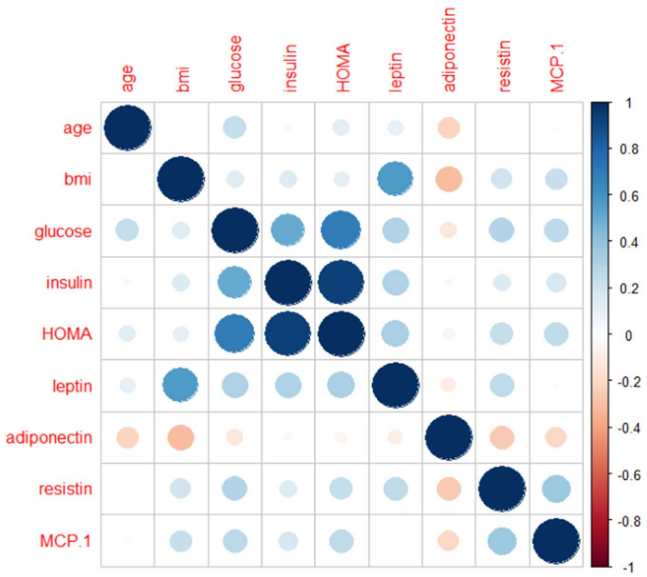
```
bPCA<- prcomp(bdata[,-10],scale.=TRUE)
summary(bPCA)
bPCA$sdev
bPCA$rotation

plot(bPCA)
autoplot(bPCA, data=bdata, colour='class')
```



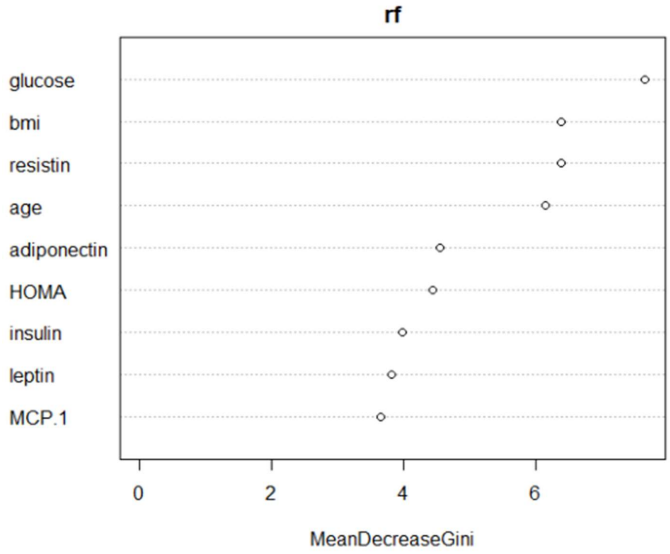
Appendix B

```
library(corrplot)
correlations <- cor(bdata[,1:9])
corrplot(correlations, method="circle")
```



Appendix C

```
library(randomForest)
set.seed(1)
rf <- randomForest(class~.,data=bdata, subset=train, mtry=2)
pred = predict(rf, test, type = "class")
confusionMatrix(pred,test$class)
importance(rf)
varImpPlot(rf)
```



Appendix C

```
set.seed(10)
cv.class=cv.tree(tree.bdata, FUN=prune.misclass)
plot(cv.class$size,cv.class$dev,type="b")
```