Application of Machine Learning Classifiers for Breast Cancer Diagnosis

Homayoon Fotros

Statistical Machine Learning (STAT-627)

American University
Spring 2022

Background and Motivation

Between 2009 and 2013, the Gynaecology Department of the University Center of Coimbra in Portugal conducted a study on newly diagnosed women with breast cancer (BC). For each of these 64 patients, the diagnosis was confirmed through a medically reliable process. These women had no other acute disease or comorbidities at the time of the study. The study also recruited 52 healthy women as control subjects. Under the approval of the university's ethical research procedure, all the participants have provided written consent to the principal investigators for anonymized use of the collected data.

The original study's goal was to investigate the relationship between *hyperresistinemia* (a metabolic issue related to the amount of Resistin in blood) along with other metabolic anomalies and breast cancer. For this purpose, the participants' blood samples were collected and analyzed to extract the measures of interest. These measures include the level of *Glucose*, *Leptin*, *Adiponectin*, *Resistin*, and the *Chemokine Monocyte Chemoattractant Protein 1* (*MCP-1*) in the blood. The data set also includes participants' plasma level of *Insulin*, which is used to calculate the *Homeostasis Model Assessment* (*HOMA*) index (*HOMA* = log((If) x (Gf)) / 22.5, where If is the fasting insulin level (μU/mL) and Gf is the fasting Glucose level (mmol/L). In addition, the participants' age (year) and BMI (weight (kg)/squared height (m2)) values were recorded in the data set. Altogether, the data set includes 9 features for all 116 participants. The response variable is whether a participant is diagnosed with BC (=1) or not (=0).

In this paper, I use 9 classification methods and fine-tune them on a training set that includes the above features and the response. Then, I will evaluate and compare the models' performance on a holdout set. The goal is to achieve a model that has robust predictive power while maintaining complexity, computational feasibility, and interpretability among other considerations in developing statistical machine learning models. In the following, I first provide a summary statistic of the variables. After setting aside a random portion of the data as the holdout validation set (15% of the data set), I outline a brief exploratory data analysis (EDA) on the training set. Then, I explain the models and their respective hyper-parameters for the training phase. Finally, I will apply the fine-tuned models on the holdout set to obtain the models' performance and discuss the findings.

Table-1 provides a summary statistic of the features and their descriptions. All the features in the data set have continuous quantitative values. The response–"Classification" is coded as a binary variable with values of 0 and 1, representing 'not diagnosed with BC' and 'diagnosed with BC' respectively. The data set is complete and has no missing value. The holdout set includes 15% of the data and I proceed with the rest of the analysis on the training set that includes N=98 instances.

Table 1 - Descriptive Summary of Data

	Min	Median	Mean	Max	Description
Age	24.0	56.0	57.3	89.0	Participant's Age (year)
BMI	18.4	27.7	27.6	38.6	Body Mass Index (kg/m2)
Glucose	60.0	92.0	97.8	201.0	Glucose level in blood (mg/dL)
Insulin	2.4	4.4	10.0	58.5	Insulin level in blood (mg/dL)
HOMA	0.5	1.4	2.7	25.1	*Homeostasis Model Assessment* (*HOMA*) index
Leptin	4.3	20.3	26.6	90.3	Leptin level in blood (ng/mL)
Adiponectin	1.7	8.4	10.2	38.0	Adiponectin level in blood (ng/mL)
Resistin	3.2	10.8	14.7	82.1	Resistin level in blood (ng/mL)
MCP-1	45.8	471.3	534.7	1698.4	Monocyte Chemoattractant Protein-1 (MCP-1) (pg/dL)
Classification	0 - Not Diagnosed with Breast Cancer (n=52) 1 - Diagnosed with Breast Cancer (n=64)				

Exploratory Data Analysis

The right panel in Figures-1 show the frequency distribution of features for each class of the response, which highlights several important patterns. First, the distribution of blood-related variables such as *Glucose* and *Adiponectin* is skewed. Based on this observation, I use the log transformation of these variables in the classification models. Second, the interquartile range of each variable versus the response class (Figure-1, left panel) has considerable overlaps that makes it largely impossible to predict the response only by one variable. However, the overlaps for *Glucose* and *Resistin* are relatively thinner than others. The scatter plot in Figure-2 illustrates the relationship between *Glucose* and *Resistin*, where each instance is colored by the response type. This figure shows despite some observable patterns and clusters, the two types are not simply separable by the values of these two features. Finally, the correlation matrix in the right panel of Figure-2 indicates that Classification (response) has a high correlation with *Glucose* (0.41), while its correlation with other features is not significant. We also notice that *Glucose*, *Insulin*, and the *HOMA* index are highly correlated with each other. This is in part because of the close relationship between *Glucose* and *Insulin* levels and the way *HOMA* is calculated. Therefore, I expect one or two of these three variables may ultimately not be present in the fine-tuned classification model.

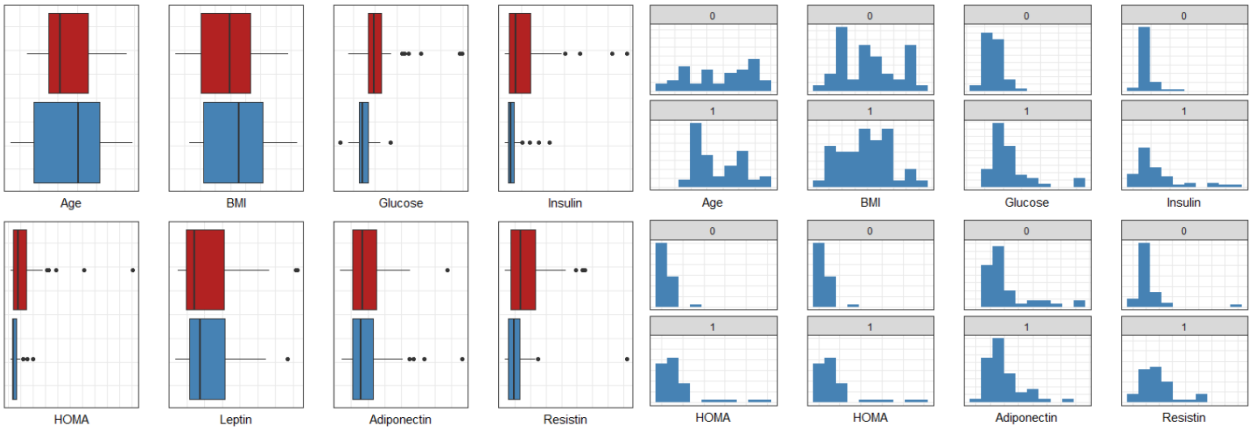

Figure 1 – Left: ***Comparison of features' interquartile range for each class.*** **Right:** ***Distribution of instances by features separated for each class***

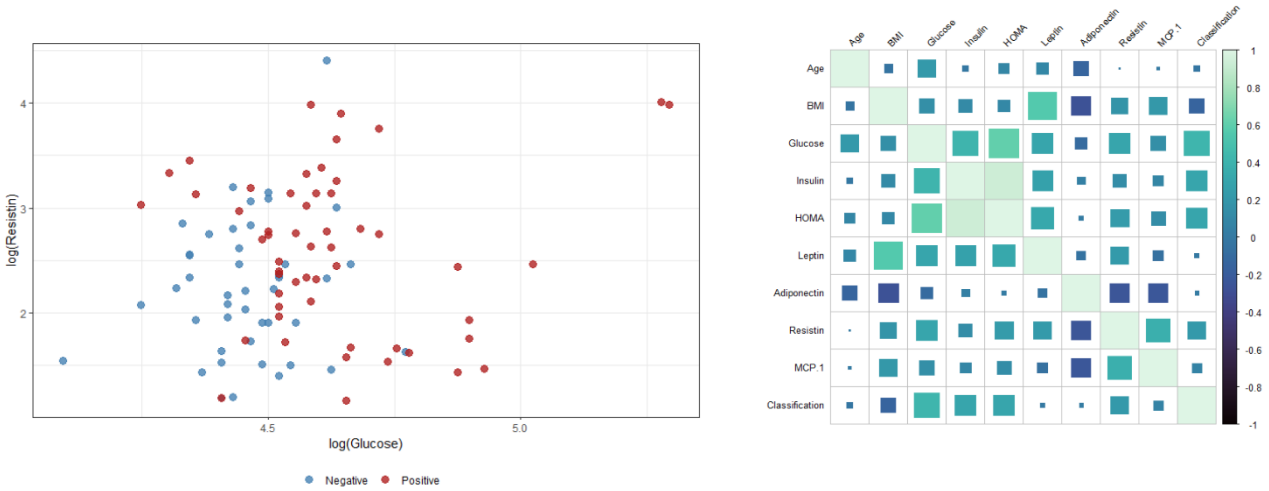

Figure 2 – Left*: Relationship between Resistin and Glucose in the training data set.* **Right:** *Correlation matrix for the training data set.*

Considering the skewed distribution of blood-related variables, I use their log transformation in the following classification models. This transformation can help recognize the patterns in the data and might help increase the interpretability. For *Age* and *BMI*, I examine whether these features have a quadratic relationship with the response. Including the quadratic forms of these variables would not complicate the models because the design (e.g., the lasso, best subset selection method) and fine-tuning will take care of excessive model complexity.

Classification Models

I implement two categories of models. First, a logistic regression model and fine-tuned versions of that using the lasso and best subset methods. The logistic regression model can be formally represented as

$$log\left(p(X)/\left(1-p(X)\right)\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

where *p(X)* is the probability of being in the positive class for values of $X = (X_1, \ldots, X_p)$ as the models' p predictors. Here, the predictors are *Age*, *BMI,* and their quadratic forms, plus the log transformation of blood-related variables (e.g., *Glucose*, *Resistin*, etc.). This amounts to a total of $p = 11$ variables in the full logistic model. Using the Anova test, I also verify that adding the quadratic forms of *Age* and *BMI* significantly improves the model fit (F=9.15, p-value=0.0001). Notably, the result shows Age^2 is among the statistically significant variables (p-value=0.001) along with *Resistin* and *BMI*. The pseudo R-squared (Cox & Snell) value for this model is 0.52. The logistic model's accuracy on the training set is 0.87, which may seem significant but at the same time could have been driven by the small size of the training set and overfitting.

In addition to a simple logistic regression model as the baseline, I estimate the coefficients (β_i) by bootstrapping (N=1000). The bootstrap method does not perform well on the training set (accuracy = 0.64). Nonetheless, I will evaluate its performance on the holdout set to compare it with other models. Next, I apply the lasso using k-fold (k=10) and LOOV (leave-one-out) cross-validation methods. The lasso's advantage is that it can shrink the number of features and improve the model's parsimony. In this case, the lasso chooses only 6 or 7 variables (*BMI, Glucose, HOMA, Resistin,* Age^2, and BMI^2) and sets the others to zero. Notice that this result provides further support for the relationship between the response and the quadratic forms of *Age* and *BMI*. I also implement the best subset and stepwise model selection considering AIC and BIC criteria. The subset selection

Table 2 – Coefficients of Linear Logistic Regression with Various Methods ***(exponentiated values)***

Var	Logistic (Base)	Lasso (k-fold)	Lasso (LOOV)	Subset (Forward Sel.)
Age	0.525	0	0	0
BMI	0.200	-0.245	-0.293	0.267
Glucose	0	0.739	0.783	0
Insulin	0	0	0	0
HOMA	-	0.149	0.176	-
Leptin	1.592	0	0	0
Adiponectin	1.757	0	0	1.810
Resistin	2.283	0.218	0.256	2.670
MCP.1	1.253	0	0	0
Age^2	0.213	-0.490	-0.526	0.236
BMI^2	0.698	0	-0.010	0

models result in similar performance metrics on the training set, so I only pick the best of them that belongs to the forward selection method using AIC.

For the second category, I implement several non-parametric models: the k-Nearest Neighbors (k-NN), Random Forest, and Support Vector Machine (SVM). I use repeated cross-validation (10 folds, 3 repetitions) for the training step to fine-tune the models' hyperparameter. For k-NN, the fine-tuning involves choosing the number of neighbors (*k*) with the highest accuracy, which occurred at *k*=9. For the Random Forest model, I used a grid search to find the optimal number of randomly selected variables (*mtry*) and the number of 'trees to grow' (*n_trees*). This search yields that the best accuracy was obtained by *mtry* = 6 and *n_tree* = 1000, which was close to the accuracy score resulted by several other configurations of these two hyperparameters. However, I chose *mtry* as close as possible to the recommended value (square root of the number of features) and kept the value of *n_trees* low for computational feasibility. Figure-3 shows the importance of features derived from the Random Forest model. We can infer from this model that *Glucose, Age*^{2}, and Resistin are more salient in determining the classification task compared to other features.

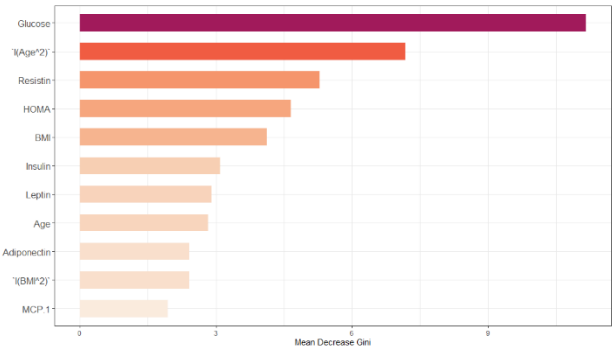

Figure 3 – ***Feature Importance in the Random Forest***

Finally, I employ linear and radial kernels of the SVM model through a grid search to fine-tune the models' *C* and *sigma* parameters (*sigma* applies only to radial kernel). The *C* parameter controls the classification's error penalty, where higher values provide a more relaxed decision boundary. *Sigma* is related to the amount of decision boundary's curvature, wit h lower values corresponding to a wigglier boundary. *I* pursued a narrow-

Table 3 – Classification Model Hyper-parameters

Model	Best Tune Parameter
k-NN	k (#Neighbors) = 9
Random Forest	mtry = 6, n_trees = 1000
SVM (Linear)	C = 1.5
SVM (Radial)	C = 45, sigma = 0.01

down strategy for fine-tuning: first, I ran the grid search to obtain an optimal zone for *C* and *sigma*. Then, I ran the grid search in that zone to obtain the parameters' optimal values. I also kept an eye on the cross-validated accuracy score to not exceed above 95 percent, which often signals the model's being overfitted to the training data set. Specifically, I limited the lower range of *sigma* to 0.01 to avoid overfitting. The optimal parameters were found at *C=45* and *sigma*=0.01, which offered a significant accuracy score at 0.94 on the training set.

Results

Table-3 and Table-4 provide the classification reports for the training and the test set respectively. Several noteworthy points emerge in the results. First, the performance of linear logistic models on the test set is considerably worse than non-parametric models. Aside from the Lasso, the logistic models' AUC scores are below 0.5, which makes them inferior to random choice. This means the very good performance of logistic models (except for the bootstrap model) on the training set are likely because of overfitting.

In contrast, the non-parametric models have done a great job of predicting the response in the holdout set. Specifically, both SVM models obtained above 80 percent accuracy and in the case of SVM with a radial kernel, the recall (sensitivity) score is equal to 1. We should treat this superbly high scores with caution, particularly for the small size of the test set (*N=18)*. Nonetheless, achieving this accuracy with a limited number of training

Table 3 – Performance Scores of Classification Models on the Training Set

Model	Accuracy	Recall	Specificity	F1	AUC
Logistic (Base)	0.87	0.87	0.86	0.85	0.86
Logistic (Boot)	0.64	0.73	0.58	0.65	0.65
Logistic (Fwd. Sel.)	0.87	0.87	0.86	0.85	0.86
Lasso (LOOV)	0.83	0.80	0.86	0.79	0.82
Lasso (k-fold)	0.82	0.79	0.86	0.78	0.81
K-NN	0.86	0.87	0.84	0.84	0.86
Random Forest	0.76	0.75	0.76	0.71	0.75
SVM (Linear)	0.88	0.86	0.90	0.86	0.87
SVM (Radial)	0.94	0.93	0.95	0.93	0.94

Table 4 – Performance Scores of Classification Models on the Holdout Test Set

Model	Accuracy	Recall	Specificity	F1	AUC
Logistic (Base)	0.50	0.55	0.43	0.40	0.49
Logistic (Boot)	0.44	0.50	0.33	0.29	0.42
Logistic (Fwd. Sel.)	0.50	0.55	0.43	0.40	0.49
Lasso (LOOV)	0.67	0.67	0.67	0.57	0.65
Lasso (k-fold)	0.67	0.67	0.67	0.57	0.65
K-NN	0.83	0.82	0.86	0.80	0.82
Random Forest	0.72	0.86	0.64	0.74	0.74
SVM (Linear)	0.72	0.73	0.71	0.67	0.71
SVM (Radial)	0.89	1.00	0.80	0.89	0.90

instances and minimal fine-tuning can inspire further research and model development. Lastly, while non-parametric models appear to be highly superior to logistic models in predicting the response, they lack the interpretability aspect that are offered by the latter models' coefficients. An exception here could be the Random Forest model that we may use the feature importance for interpretation.

Conclusion

Achieving a high level of accuracy in data-driven classification tasks such as breast cancer diagnosis serves as a salient example of machine learning applications. This has become possible by advancements in statistical classification models and major breakthroughs in high-capacity and ultra-fast computing technologies. Taking advantage of this progress, I applied several classification models on the Coimbra's Breast Cancer data set to evaluate the models' performance on the validation set. Specifically, I used logistic regression model, the lasso, k-NN, the Random Forest, and Support Vector Machines (SVM) on the training set. The best result was obtained by SVM and k-NN, followed by the Random Forest, the lasso, and the logistic model. Notably, the SVM's radial model achieved a sensitivity score of 1 on the test set, which is impressive and significant in the cancer diagnosis domain.

There are several avenues to expand the present work and improve the results. First, applying the models on a larger data set would certainly benefit the training and testing process. A larger data set could particularly make it possible to analyze the data on various subsets of patients (e.g., based on BMI, age range, etc.). Secondly, adding more relevant features and removing redundant ones that may also cause multi-collinearity issue would help the models' performance, particularly in the case of the logistic regression. Providing a better set of features would also help extracting better interpretations out of the outcomes. Another area to explore is to test other transformations of the features in the models such as the polynomial forms of blood-related variables. Finally, employing more sophisticated models such as Neural Networks may achieve higher accuracy levels while also help improve interpretability by providing layers that reveal the features' weights and importance with respect to the response.