**Providing an optimal predictive machine learning model for the breast cancer tumor diagnosis**

Sahar Nouri

BDA 104, Predictive Modeling and Data Mining

# Summary

Breast cancer is the second leading cause of cancer deaths among women in the United States. Although mortality rates have been decreasing over the past decade, it is important to continue to make advances in diagnostic procedures as early detection vastly improves chances for survival. This project aims to identify a classification machine learning model that accurately predicts the presence of a malignant tumor using data from fine-needle aspiration (FNA) with visual interpretation.

The Wisconsin Breast Cancer dataset was used that contains 569 patients (63.7% benign and 37.3% malignant). Redundant variables were specified and dropped to reduce the dimensionality of data. Alternative classification machine learning models include logistic regression, decision tree, and random forest were implemented to find the best model with the lowest classification error.

The Random Forest model was selected as the best predictive model with an accuracy score of 99.12%, precision 100%, recall 98%, and AUC 98.78%. The five most important features were radius\_mean, concavity\_mean, concavity\_worst, redius\_se, compactness\_worst. These results suggest that future research should attempt to refine the techniques used to determine these specific model inputs. Greater accuracy in characterizing the FNA attributes will allow researchers to develop more promising models for early detection.

# Introduction

Breast cancer is the most common cancer in women after skin cancer in the United States. Studies estimate that approximately 15% of newly-diagnosed cancer patients will die in 2015. Breast cancer occurs as a result of abnormal growth of cells in the breast tissue commonly named as a tumor. A tumor does not mean cancer. It can be benign (not cancerous), pre-malignant (pre-cancerous) or malignant (cancerous). Currently, 3 methods are commonly used to diagnose cancer, including mammography, fine needle aspiration (FNA) with visual interpretation, and surgical biopsy. Fine needle aspiration (FNA) is a cost-effective method of diagnosis. It examines a small amount of tissue from the tumor and then carefully examines both the characteristics of individual cells and important contextual features. However, this process is highly subjective, depending on the skill and experience of the physician.

Researchers have been working diligently over the past decade to improve the sensitivity of this process. In addition to advances in technology and visual interpretation, researchers have been using various data mining methods to identify the key factors that can help doctors correctly diagnose malignant tumors. Nowadays, machine learning has become more accurate thank human medical professionals in the diagnosis of cancerous and non-cancerous tumors. The goal of this project is to use alternative methods of predictive machine learning algorithms and evaluating their performances to provide an optimal diagnostic model. The model can thereby help clinicians make better decisions for the treatment of the breast tumor. In the next section, a description of the data set was provided. In the third section, explanatory analyses for a better understanding of the data at hand and methods for dimensionality reduction are proposed. The fourth section shows the results of the implementation of alternative machine learning methods and their comparisons.

# Data Description

This project utilized the Wisconsin Breast Cancer Data set. The data was created at the University Of Wisconsin Hospital at Madison, Wisconsin, USA, and is available at  <https://www.kaggle.com/uciml/breast-cancer-wisconsin-data>. The dataset contains 569 patients assessing the features of FNAs taken from patients’ breasts. There are 33 attributes per observation; including the ID and binary target variable. The target variable diagnoses whether the tumor is benign or malignant. To create the features they used fluid samples, taken from patients with solid breast masses, and an easy-to-use graphical computer program called Xcyt, which is capable of performing the analysis of cytological features based on a digital scan. The program uses a curve-fitting algorithm, to compute ten features from each one of the cells in the sample, then it calculates the mean value, extreme value, and standard error of each feature for the image, returning a 30 real-valuated vector. Table 1 shows the variable names, descriptions, and types.

**Table 1. List of variables, their descriptions and types.**

|  |  |  |
| --- | --- | --- |
| Variable | Description | Type |
| ID number ​ | Patient code number | Numeric (Int64) / Input |
| Diagnosis ​ | The diagnosis of breast tissues (M=Malignant, B=Benign) ​ | Categorical (object)/output |
| radius ​ | Distances from the center to points on the perimeter ​ | Numeric (float64) ​/ Input |
| texture ​ | standard deviation of gray-scale values ​ | Numeric (float64) ​/ Input |
| perimeter ​ | ​ size of the core tumor | Numeric (float64) ​/ Input |
| area ​ | ​ | Numeric (float64) ​/ Input |
| smoothness  ​ | local variation in radius lengths ​ | Numeric (float64) ​/ Input |
| compactness ​ | perimeter^2 / area - 1.0 ​ | Numeric (float64) ​​/ Input |
| concavity ​ | the severity of concave portions of the contour ​ | Numeric (float64) ​​/ Input |
| Concave points ​ | number of concave portions of the contour ​ | Numeric (float64) ​​/ Input |
| symmetry ​ | ​ | Numeric (float64) ​​/ Input |
| Fractal dimension ​ | "coastline approximation" - 1 ​ | Numeric (float64) ​​/ Input |

There were no missing values in the data set. The id column was redundant and not useful. Also Unnamed: 33 feature includes NaN. They were dropped from the analysis.

# Explanatory Analyses

Table 2. shows descriptive statistics of the features include means, standard deviations, minimum and maximum values, and their percentiles. The ranges of variables were highly different and using the original scale for model building may put more weights on the variables with large values. So, feature scaling method i.e. normalization was used to make feature scales almost the same. In this way, each feature is equally important and makes it easier to process by machine learning algorithms.

A screenshot of a cell phone

Description automatically generated

The target variable consists of 357 (62.7%) benign and 212 (37.3%) malignant patients. (Figure 1)

**A picture containing screenshot

Description automatically generated**

Figure1. Frequency of cancer diagnosis.

## Dimensionality Reduction

The correlation heat map was calculated to show the linear relationships between 30 features (Figure 2). Any feature with the correlation value more than .9 was dropped because they almost had the same information. Ten features were dropped.

Figure 3. shows the distribution of (continuous) features among malignant and benign patients. There was no perfect separation between benign and malignant patients for texture\_se, smoothness\_se, and fractal\_dimension\_se and symmetry\_se, and they were dropped.

A picture containing circuit

Description automatically generated

Figure 2. The correlation heat map of independent variables.

A close up of a map

Description automatically generated

Figure 3. The correlation heat map of independent variables.

# Predictive Analyses

Classification machine learning algorithms are supervised learning approaches in which the computer program learns from the input data and then uses this learning to classify new observations. Different machine learning algorithms were implemented and evaluated to find the best model with the lowest classification errors. The data was split to train and test data with a ratio of 80/20. Training data was used for model training and test data was used for the evaluation of model performances and predictions. Then models were then tuned to optimize their performances.

## Logistic Regression

It is a statistical method for analyzing a data set in which there are one or more features that determine the target variable. The target is measured with a dichotomous variable (in which there are only two possible outcomes). The goal of logistic regression is to find the best fitting model to describe the relationship between the dichotomous characteristic of interest (dependent variable = response or outcome variable) and a set of independent (predictor or explanatory) variables. Before hyperparameter tuning the accuracy of logistic regression 93.68% (Figure 4)

A screenshot of a cell phone

Description automatically generated

Figure 4. Results of logistic regression before hyperparameter tuning.

### Hyperparameter tuning for Logistic Regression

For improving the performance of the model the Gridsearch Cross Validation method was used. Using this method cross-validation and hyperparameter tuning are done at the same time. Cross-validation is the process of training model using one set of data and testing it using different data. Parameter tuning is the process of selecting the values for model parameters that maximize the accuracy of the model. The plots of accuracy values for different numbers of cross-validation and penalties methods are shown in figure 5. The highest accuracy was 97.58%with the penalty method l2 and cv=5. The accuracy score of this model for the test data was 97.37% (figure 6).

A close up of text on a white background

Description automatically generated

Figure 5. Plots of accuracy values for a different number of cross-validation and penalty methods.

A screenshot of a cell phone

Description automatically generated

Figure 6. The results of logistic regression after hyperparameter tuning.

### Feature Selection for logistic regression

Using recursive features elimination the best subset of features was selected. The number of selected features was 10 with an accuracy of 93.86% (figure 7).

A screenshot of a cell phone

Description automatically generated

Figure 7. Results of logistic regression after feature selection.

## Classification Tree

Figure 8 shows the initial results of the classification tree before hyperparameter tuning. The model had an accuracy of 91.23%. Using GridSearch cross-validation the accuracy of the classification tree increased to 96.49% (figure 9).

A screenshot of a cell phone

Description automatically generated

Figure 7. The results of the classification tree before hyperparameter tuning.

A screenshot of a cell phone

Description automatically generated

Figure 8. The results of the classification tree after hyperparameter tuning.

### Feature selection using classification tree

Using the classification tree the feature importances were found (figure 9). The five most important features were radius\_mean, concavity\_worst, concavity\_se, smoothness\_worst, and radius\_se. The accuracy of the classification model after feature selection was 93.86% (Figure 10).

A screenshot of a cell phone

Description automatically generated

Figure 9. Feature importances using the classification tree.

A screenshot of a cell phone

Description automatically generated

Figure 10. Results of classification tree after feature selection.

## Random Forests

The random forest method is an ensemble learning method for classification and regression. It constructs a multitude of decision trees at training time and outputs the class or the mean prediction. The accuracy of the random forest model before hyperparameter tuning was 98.24% (figure 11). Using GridSearch cross-validation for hyperparameter tuning the accuracy of the model increased to 99.12% (figure 12).

### Feature selection using Random Forest

Feature importances were determined based on the random first method (figure 13). The five most important features were radius\_mean, concavity\_mean, concavity\_worst, redius\_se, and compactness\_worst. Using feature selection, the accuracy of the random forest was 97.37% (Figure 14).

A screenshot of a cell phone

Description automatically generated

Figure 11. The results of the random forest before hyperparameter tuning.

A screenshot of a cell phone

Description automatically generated

Figure 12. The results of the random forest after hyperparameter tuning.

A screenshot of a cell phone

Description automatically generated

Figure 13. Feature importances using the random forest.

A screenshot of a cell phone

Description automatically generated

Figure 14. Results of the random forest after feature selection.

# Conclusion

The accuracies of all three models reduced after feature selection. Normally, accuracy reduces when the number of features decreases. In some cases where the number of features is very large using feature selection or principal components can improve the performance of the model. However, for this data, the total number of remained features after dropping redundant ones was 16. Considering 212 malignant samples it was an acceptable number for features (statistically we need 10 samples of the class of interest, that here is malignant, for each feature). So we used all features for prediction in this project. Table 2 shows the results of all three models after hyperparameter tuning. The Random Forest model was selected as the best predictive model with an accuracy score of 99.12%, precision 100%, recall 98%, and AUC 98.78%. The five most important features were radius\_mean, concavity\_mean, concavity\_worst, redius\_se, compactness\_worst. These results suggest that future research should attempt to refine the techniques used to determine these specific model inputs. Greater accuracy in characterizing the FNA attributes will allow researchers to develop more promising models for early detection.

Table 2. Final results of the three machine learning models on the Wisconsin Breast Cancer Data set.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Logistic regression | Classification Tree | Random Forest |
| Accuracy | 97.37% | 96.49% | 99.12% |
| Precision ​ | ​100% | 100% | 100% |
| Recall​ | 93% | 90% | 98% |
| F1-score ​ | ​96% | 95% | 99% |
| AUC​ | ​96.34% | 95.65% | 98.78% |