ML project_Breast Cancer Tumor Classification

Group 32: Danning Tian Xiaoting Chen

2022/4/20

Introduction

The diagnosis of malignant or benign tumors is a vital step in cancer diagnosis and treatment process, without which a specific and effective medical treatment plan cannot be determined. The process of diagnosis is complex, involving observation of clinical presentation, laboratory examination, imaging examination, and then the pathological examination, which is also known as a biopsy. It is generally acknowledged that a biopsy is a gold standard for tumor diagnosis. However, there have been some studies questioning the reliability of the biopsy result for identifying subtle abnormalities. Also, biopsy often raises concerns of patients about the spread of cancer cells.

Breast cancer is the second leading cause of cancer death in women. The chance that a woman will die from breast cancer is about 1 in 39 (about 2.6%). The early diagnosis of breast cancer can improve the prognosis and chance of survival significantly, as it can promote timely clinical treatment to patients. Detecting Malignant Breast Cancer Tumor in the early stage may largely enhance the 5-year and 10-year survival rate of breast cancer patients. Further accurate classification of benign tumors can prevent patients undergoing unnecessary treatments. Thus, In this project, machine learning methods are applied to assist breast cancer tumor classification and diagnosis based on a digitized image of a fine needle aspirate (FNA) of a breast mass to seek solutions for a more precise prediction besides biopsy results.The machine learning classification outcome could be a powerful tool to assist doctors in the diagnostic process. A total of six different machine learning methods are selected in this project, namely Logistic Regression, LASSO, Decision Tree, Bagging, Random Forest, and Boosting. It is hypothesized that our final test outcome would achieve 95% accuracy, as well as 95% sensitivity and specificity in the best model.

Related work

Breast Cancer diagnosis is one of the most intricate processes that requires a lot of medical diagnostic measurements and efforts from oncologists. One of the most widely applied methods to determine Breast Cancer tumor type is Biopsy. A series of lab tests is required to measure the tissue getting from the human body and the process of biopsy is comparatively intrusive. Also, such a method requires a lot of human effort, which is comparatively time-consuming. Also, the result might not be accurate and may not cover all parameters needed to analysis due to potential human error.

Previously, machine learning classification of breast cancer based on RNA sequence of tumor cells have also been launched (Wu and Hicks, 2021). Even though the classification models seem accurate, the RNA sequence analysis is expensive and time consuming. Also, a lot of lab work such as PCR is required for each test sample. Therefore, a method that may simplify steps should be launched.

Also, machine learning classification based on MRI image (Sutton et al, 2020) has been launched as well. This method is non-intrusive but still requires a high quality MRI image for analysis.

Despite the fact that there are many applications of machine learning in the field of breast cancer classification already, our dataset and models used are still considered innovative. Moreover, one strength of our choice of dataset would be generating such dataset does not require professional trained individuals to supervise, fully digitized scales is possible to automate the entire process, which would be a more primitive method that may have wider application.

Methods

Our main objective is developing a machine learning model to predict tumor type based on tumor's observational data. Based on the dataset, detailed data of breast cancer tumors are included. Hence, models constructed based on mean, standard error, and extreme values of the tumor would be evaluated and utilized separately. To give the best tumor classification outcome and to consolidate our understanding of machine learning in the meantime, four classification models, namely decision tree, random forest, boosting tree algorithm, and lasso, would be launched and compared in this project.

Decision tree is considered as one of the classical classification models, which is very straightforward and the pattern may clearly be determined by looking at the tree map. However, the accuracy of this model may be affected by extreme values and variance. We contain this model majorly as a baseline model, which is mainly used for comparison with other models.

Random forest is a classification method based on decision trees. Since our dataset is comparatively small (569 observations), the computational burden of the random forest may not be too heavy to achieve. Random Forests generally provide high accuracy and balance the bias-variance trade-off well. Since the model's principle is to average the results across the multiple decision trees it builds, it averages the variance as well. Also, random forests are not influenced by outliers to a fair degree by bagging the variables. However, Random Forests are not easily interpretable. They provide feature importance but it does not provide complete visibility into the coefficients as linear regression.

Boosting is also a machine learning method based on decision trees. Advantage of this model is the method reduces variance and bias in a machine learning ensemble. However, boosting is a slow learning algorithm, hence, our comparatively small dataset may not give a favorable classification outcome.

LASSO regression is also a powerful tool in logistic models. The reason we choose LASSO is mainly because it is more advanced than traditional regression models to address collinearity. However, this model may not be as straightforward as the previous two models based on decision trees.

By launching those four models, train error and test error would be compared to determine the best model for predicting the tumor type based on diagnosis data of breast cancer. The standard to determine the effectiveness of the model would be high training classification accuracy (small training error) and high testing classification accuracy (small training error). Besides, sensitivity and specificity are also important measurements in model selection.

Data and Experiment setup

The data we plan to use is Breast Cancer Wisconsin (Diagnostic) Data. The dataset includes the attribute information of diagnosis of breast tissues and features describing the cell nuclei present in the image, which is digitized from the breast mass fine needle aspiration (FNA) products. There are a total of 10 categories of predictors, namely radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. The mean, standard error, and "worst" or largest of these features were computed, resulting in 30 features. Histograms shows each feature's distribution is showns as followed. 569 observations in breast cancer tumors were included, among which 63% (357 observations) were diagnosed as benign and the rest 37% (212 observations) malignant.

For the 569 observations, 70% of those observations are randomly selected as train sets for model building and the rest 30% of data are included in the test set as our initial settings. In each set, the proportion and distribution of malignant observation and benign observations shall be the same. Train set is used to construct each model and a test set is used to evaluate the model. To avoid overfitting, each training error and testing error are both important measuring standards to choose the best model.

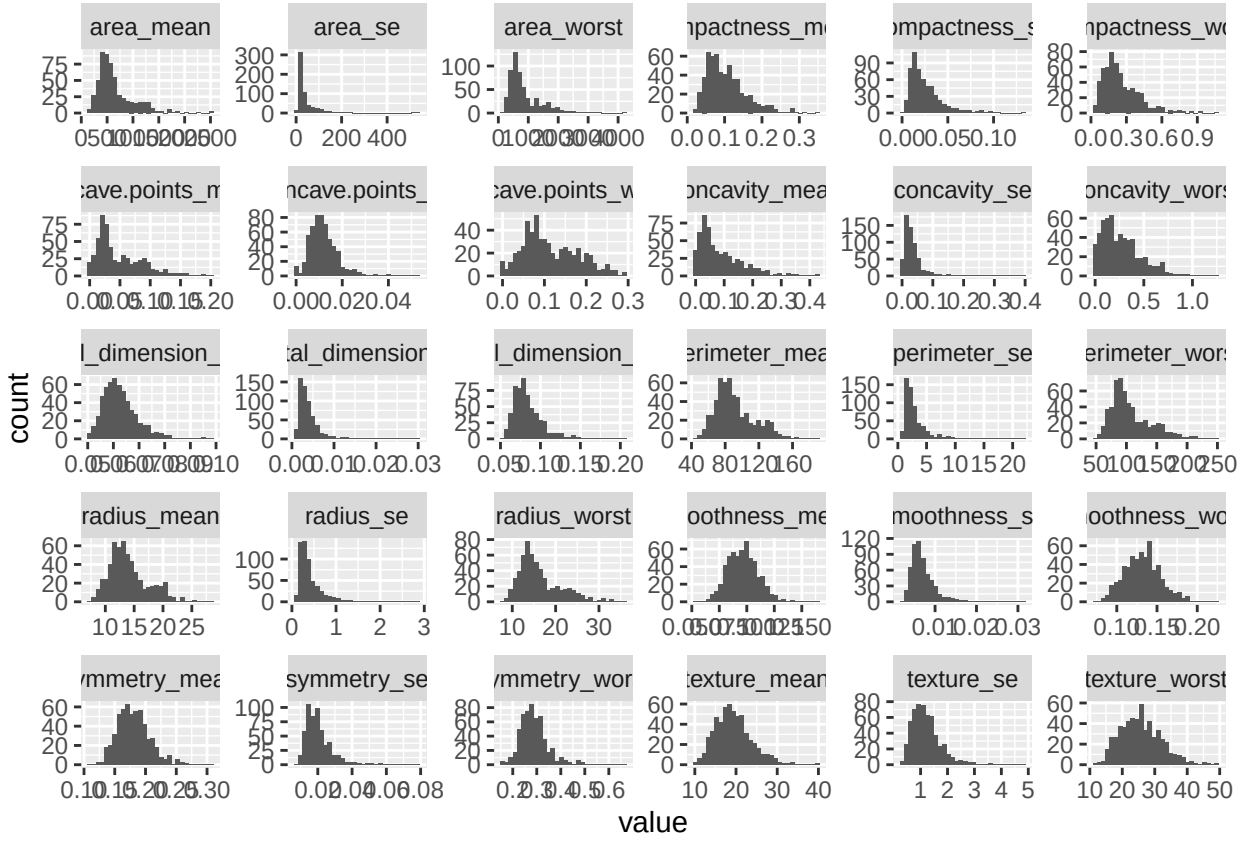


Figure 1: Feature Histogram Summary

Results

lasso

The first model is a cross-validated lasso regression classification model, with `nfolds` equal to 10. The CV result plot is shown below, and the value of lambda giving the minimum binomial deviance is 0.0027284.

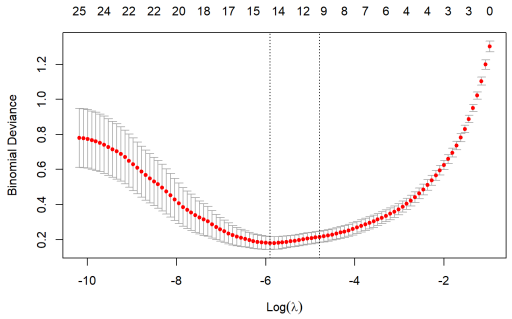


Figure 2: Lasso cross-validation result

The regression model summary using the chosen lambda is as follows. After the shrinkage process, some predictors are showing more importance in the model, such as the mean texture, the mean concave points, and the standard error of fractal dimension.

	s1
(Intercept)	-21.63783061
radius_mean	.
texture_mean	0.06528292
perimeter_mean	.
area_mean	.
smoothness_mean	.
compactness_mean	.
concavity_mean	.
concave.points_mean	14.26319404
symmetry_mean	.
fractal_dimension_mean	.
radius_se	4.79156202
texture_se	.
perimeter_se	.
area_se	.
smoothness_se	.
compactness_se	.
concavity_se	.
concave.points_se	.
symmetry_se	.
fractal_dimension_se	-52.56139372
radius_worst	0.49703237
texture_worst	0.12240116
perimeter_worst	.
area_worst	.
smoothness_worst	16.15526400
compactness_worst	.
concavity_worst	0.06730865
concave.points_worst	21.27160201
symmetry_worst	5.30571963
fractal_dimension_worst	.

Figure 3: Lasso model summary

Predictions are made on training and test sets using the chosen model, and training and test accuracy are both around 98%, which indicates the good performance of the LASSO model.

```
##       Accuracy Error Sensitivity Specificity
## Train    0.987 0.013       0.972       0.996
## Test     0.982 0.018       0.972       0.990
```

decision tree

The second method is the decision tree. Both Entropy D and Gini G were applied as the evaluation index and the model using Entropy D got a slightly higher accuracy, so only the results using Entropy D will be discussed here. Using cross-validation, the best size of the tree turns out to be 5, and the pruned tree with 5 nodes is shown in Fig-4. According to the pruned decision tree, it is intuitive that the predictor `worst area` plays a leading role in cancer diagnosis, and `concavity` is another essential indicator. Although the decision tree model is easily interpreted, its prediction performance is not that good, with test accuracy only around 92%.

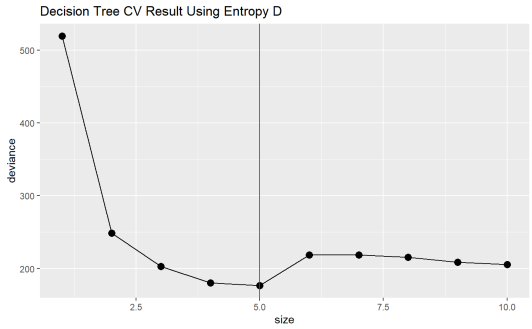


Figure 4: Tree size

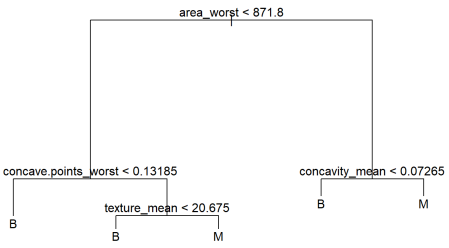
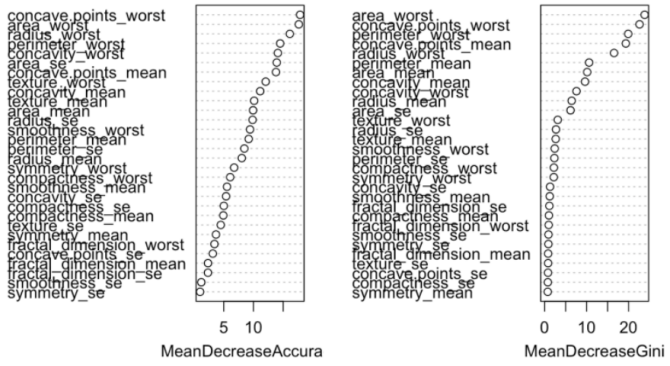


Figure 5: The pruned decision tree

```
##       Accuracy Error Sensitivity Specificity
## Train    0.962 0.038       0.894           1
## Test     0.924 0.076       0.817           1
```

random forest

The third model is Random Forest, which is a tree based model. When consider the variable importance indicators, concave point worst and area worst are the most important variables indicated by both mean decrease accuracy chart and mean decrease gini chart. The sensitivity of the the train set is 1, which indicated that the model could been over fitted. However, the sensitivity and specificity for the test model is both over 95% as well, which indicated that this model is actually accurate and perform better when compared to the decision tree model. /



	B	M	MeanDecreaseAccuracy
radius_mean	6.860695	5.2973331	8.041860
texture_mean	7.105749	8.0088955	10.119776
perimeter_mean	7.317963	5.1572938	9.202921
area_mean	8.611097	6.0463730	9.938129
smoothness_mean	2.911747	5.3320704	5.549062
compactness_mean	3.835610	2.9278306	4.929921
concavity_mean	6.664262	8.6337870	11.130907
concave.points_mean	9.346590	11.2631484	13.791466
symmetry_mean	2.166076	2.7629276	3.698981
fractal_dimension_mean	2.163711	0.5983889	2.351034
radius_se	7.950978	5.4800377	9.857496
texture_se	4.451598	1.6204776	4.499202
perimeter_se	6.379117	5.4248153	8.463178
area_se	11.695399	6.6852663	13.920850
smoothness_se	2.106388	-2.0772719	1.224612
compactness_se	3.984064	2.3028218	5.015273
concavity_se	3.064729	4.3242695	5.396903
concave.points_se	2.810519	1.7385847	3.166781
symmetry_se	1.809913	-1.0161959	1.024535
fractal_dimension_se	2.494355	0.3822174	2.343065
radius_worst	13.159203	10.0804416	16.121003
texture_worst	9.870154	8.9813737	12.061817
perimeter_worst	12.011375	10.1685548	14.493320
area_worst	14.145218	10.7818722	17.629191
smoothness_worst	6.725258	7.5419555	9.424146
compactness_worst	4.294738	3.8793750	6.085037
concavity_worst	7.683524	11.6494937	14.103388
concave.points_worst	14.532940	12.0072818	17.841619
symmetry_worst	5.138609	4.8610940	6.752209

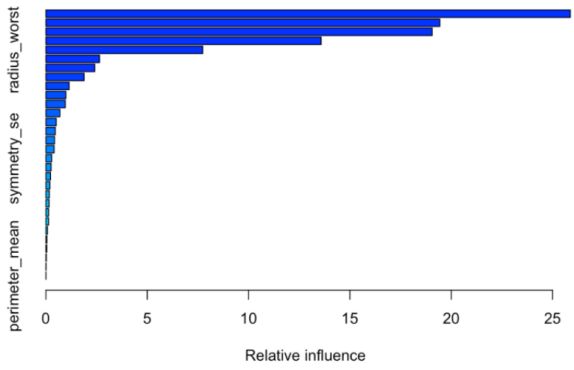
Figure 6: Random Forest Variable Importance Summary

	Accuracy	Error	Sensitivity	Specificity
Train	1.000	0.000	1.000	1.00
Test	0.977	0.023	0.958	0.99

boosting

The last model is Boosting. When consider the variable importance indicator, concave point worst, perimeter worst and area worst are the three most important variables in this model, which is consistent with the decision tree and random forest results. Also, by looking at the accuracy, sensitivity, and specificity of both train and test data, this would be considered as the best model with reasonable trade-offs. Also, the sensitivity for both train and test data are over 97%, which makes this method favorable.

	Accuracy	Error	Sensitivity	Specificity
Train	0.992	0.008	0.979	1.00
Test	0.982	0.018	0.972	0.99




```
## concave.points_worst       concave.points_worst 5.486143e+01
## perimeter_worst                 perimeter_worst 2.130670e+01
## area_worst                           area_worst 7.031576e+00
## texture_mean                       texture_mean 4.430712e+00
## radius_worst                       radius_worst 2.714392e+00
## symmetry_worst                   symmetry_worst 1.851826e+00
## concave.points_mean         concave.points_mean 1.815164e+00
## smoothness_worst               smoothness_worst 1.340364e+00
## symmetry_se                         symmetry_se 1.011408e+00
## fractal_dimension_mean   fractal_dimension_mean 6.580202e-01
## area_se                                 area_se 5.002595e-01
## concavity_worst                 concavity_worst 4.697911e-01
## texture_worst                     texture_worst 3.456081e-01
## compactness_se                   compactness_se 2.910071e-01
## concavity_se                       concavity_se 2.886500e-01
## symmetry_mean                     symmetry_mean 2.466849e-01
## smoothness_mean                 smoothness_mean 2.166568e-01
## fractal_dimension_worst fractal_dimension_worst 2.126867e-01
## radius_se                             radius_se 1.150422e-01
## compactness_worst           compactness_worst 8.833850e-02
## texture_se                           texture_se 7.272067e-02
## fractal_dimension_se       fractal_dimension_se 6.298122e-02
## compactness_mean             compactness_mean 4.086112e-02
## perimeter_se                     perimeter_se 1.027425e-02
## concave.points_se           concave.points_se 8.067201e-03
## smoothness_se                   smoothness_se 4.594327e-03
## area_mean                           area_mean 4.190136e-03
## perimeter_mean                 perimeter_mean 3.954440e-07
## concavity_mean                 concavity_mean 1.227173e-09
## radius_mean                       radius_mean 9.409109e-13
```

Figure 7: Boosting Variable Importance Summary

Discussion

Below is the summary of model performance on the training and test set. According to the performance summary, the decision tree model has relatively lower accuracy compared to the other models. So out of four models, Lasso, Random Forest, and Boosting are recommended, they all have high accuracy for both training and test sets' prediction. More specifically, the random forest model fits the training data best. Lasso and Boosting have the highest sensitivity, which is the probability of correctly identifying malignant tumors in this case. And Lasso, Random Forest, and Boosting models all have high specificity, which is to correctly identify benign tumors. In practice, the model selection should be based on the result of previous examinations. For example, if the previous examination result suggests that there's a high probability that the tumor is benign, then we might tend to use the model with high specificity. To sum up, the obtained models have pretty good performance in both the training and testing sets, high sensitivity and specificity are achieved.

```
##           model  accuracy       error sensitivity specificity
## 1         Lasso 0.9874372 0.012562814   0.9716312   0.9961089
## 2 Decision Tree 0.9623116 0.037688442   0.8936170   1.0000000
## 3 Random Forest 1.0000000 0.000000000   1.0000000   1.0000000
## 4      Boosting 0.9924623 0.007537688   0.9787234   1.0000000
```

```
##           model  accuracy      error sensitivity specificity
## 1         Lasso 0.9824561 0.01754386   0.9718310        0.99
## 2 Decision Tree 0.9239766 0.07602339   0.8169014        1.00
## 3 Random Forest 0.9766082 0.02339181   0.9577465        0.99
## 4      Boosting 0.9824561 0.01754386   0.9718310        0.99
```

In terms of the significant predictors of tumor diagnosis, here are the predictors showing importance in more than or equal to 3 models out of 4: `texture_mean`, `radius_worst`, `area_worst`, and `concave.points_worst`. This is only a subjective conclusion to suggest these may be the primary aspects of concerns in future tumor diagnosis, not a rigorously deduced result.

For further analysis, firstly, since we have roughly identified the important predictors, it is worth considering building the model with only the selected predictors, and see whether the accuracy would still be good while the model becomes more concise. Secondly, out of all the models we have built, no model has achieved 100% sensitivity. That is to say, if applied in practical use, there's risk that a malignant tumor could be diagnosed as benign, which would lead to a fatal consequence. Further analysis should be made to improve the model sensitivity.

Reference

About breast cancer: Breast cancer overview and basics. American Cancer Society. (n.d.). Retrieved March 29, 2022, from https://www.cancer.org/cancer/breast-cancer/about.html

Bhattarai, S., Klimov, S., Aleskandarany, M. A., Burrell, H., Wormall, A., Green, A. R., Rida, P., Ellis, I. O., Osan, R. M., Rakha, E. A., & Aneja, R. (2019, August 9). Machine learning-based prediction of breast cancer growth rate in vivo. Nature News. Retrieved May 9, 2022, from https://www.nature.com/articles/s41416-019-0539-x

breastcancer.org. (n.d.). Breast Cancer Facts and Statistics . Breast cancer facts and statistics. Retrieved March 29, 2022, from https://www.breastcancer.org/facts-statistics

Jeremy Jordan. (2018, August 25). Evaluating a machine learning model. Jeremy Jordan. Retrieved March 29, 2022, from https://www.jeremyjordan.me/evaluating-a-machine-learning-model/

Learning, U. C. I. M. (2016, September 25). Breast cancer wisconsin (diagnostic) data set. Kaggle. Retrieved March 29, 2022, from https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

Mayo Foundation for Medical Education and Research. (2022, January 13). Breast Cancer. Mayo Clinic. Retrieved May 9, 2022, from https://www.mayoclinic.org/zh-hans/diseases-conditions/breast-cancer/diagnosis-treatment/drc-20352475

Sutton, E. J., Onishi, N., Fehr, D. A., Dashevsky, B. Z., Sadinski, M., Pinker, K., Martinez, D. F., Brogi, E., Braunstein, L., Razavi, P., El-Tamer, M., Sacchini, V., Deasy, J. O., Morris, E. A., & Veeraraghavan, H. (2020, May 28). A machine learning model that classifies breast cancer pathologic complete response on MRI post-neoadjuvant chemotherapy - breast cancer research. BioMed Central. Retrieved May 9, 2022, from https://breast-cancer-research.biomedcentral.com/articles/10.1186/s13058-020-01291-w

Wu, J., & Hicks, C. (2021, January 20). Breast cancer type classification using machine learning. Journal of personalized medicine. Retrieved May 9, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7909418/