Breast cancer detection using machine learning algorithms

Pavel Makarov

Bellevue University

DSC680: Applied Data Science

Amirfarrokh Iranitalab

06/23/2024

**Breast cancer detection using machine learning algorithms**

# Introduction

Breast cancer is the most common type of cancer in women, accounting for approximately 30% of all new female cancer cases each year. Detecting this type of cancer at the early stages is extremely critical. Modern techniques, such as laboratory cell and blood tests, physical condition assessments, and breast scan analysis, allow specialists to detect and determine whether a tumor is benign (non-cancerous) or malignant (cancerous). The process of laboratory tests is done manually, which can be very time-consuming and error-prone. Modern computational techniques enable the development of machine learning models that can accurately and precisely determine if a cell is benign or malignant. By using historical results with labeled outcomes, a classification algorithm can be trained to analyze future unseen data.

The model developed in this project aims to offload the manual labor involved in reviewing patients' analyses and test results by classifying tumors as malignant or benign. This supervised machine learning model can assist in the automated detection of breast cancer based on blood and cell analysis results, providing oncologists with a tool to support the detection of breast cancer and minimize human involvement, thus reducing the possibility of human errors.

# Data Exploration

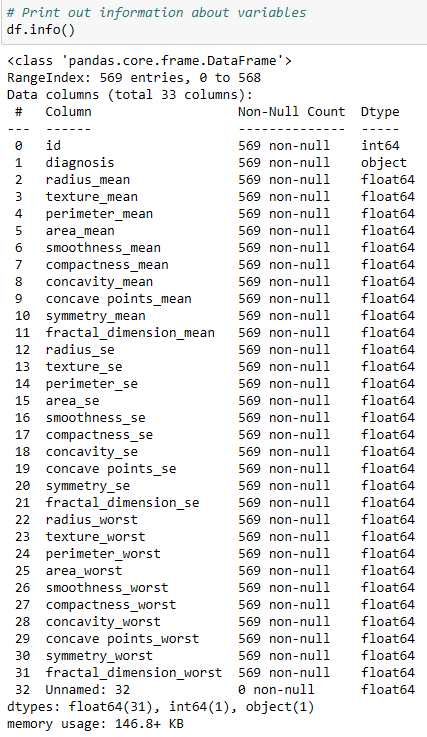
An Original Wisconsin Breast Cancer database dataset from Kaggle.com was used to develop the classification model. Below is a description of all variables in the dataset:

*Table 1 - Data description*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Variable Name** | **Role** | **Type** | **Units** | **Missing Values** | **ID number** |
| Sample\_code\_number | ID | Categorical |  | no | 1-10 |
| Clump\_thickness | Feature | Integer |  | no | 1-10 |
| Uniformity\_of\_cell\_size | Feature | Integer |  | no | 1-10 |
| Uniformity\_of\_cell\_shape | Feature | Integer |  | no | 1-10 |
| Marginal\_adhesion | Feature | Integer |  | no | 1-10 |
| Single\_epithelial\_cell\_size | Feature | Integer |  | no | 1-10 |
| Bare\_nuclei | Feature | Integer |  | yes | 1-10 |
| Bland\_chromatin | Feature | Integer |  | no | 1-10 |
| Normal\_nucleoli | Feature | Integer |  | no | 1-10 |
| Mitoses | Feature | Integer |  | no | 1-10 |
| Class | Target | Binary |  | no | 2 = benign, 4 = malignant |

The dataset does not require extensive preparation as no missing values were detected. The class feature, coded as 2 = benign and 4 = malignant, was re-coded into 0 = benign and 1 = malignant for model training purposes. The .info() function was applied to the entire dataframe for an initial overview. The first step was to remove irrelevant columns such as “id” and “Unnamed:32”. Then, the isna() function was applied to ensure no missing values in the dataframe.

*Figure 1 – Snapshot of the data frame description*



Since all features are numeric, their normality and the presence of outliers were tested. Most features have a minor number of values on the higher side of the distribution, but they are not considered outliers because they are not far from the mode. These values were not removed or imputed before modeling.

*Figure 2 – Histograms for all features in the data frame*

A collage of blue and white graphs

Description automatically generated

The bar graph below demonstrates the balance between target classes. The ratio between classes is approximately 40% to 60%, so the target cannot be considered significantly imbalanced. In this cases class balancing technique were not applied.

*Figure 3 – Target class balance*

A graph of a number of people

Description automatically generated

The correlation matrix heatmap shows the correlation between features and the target value. Most features, except fractal\_dimension\_se, symmetry\_dimension\_se, texture\_se, fractal\_dimension\_mean, and smoothness\_se, strongly correlate with the target variable.

*Figure 4 – Correlation heatmap*

A red and white chart

Description automatically generated with medium confidence

A recursive feature elimination (RFE) algorithm was applied to further investigate feature importance and perform feature selection. This algorithm ranks features based on their impact on accuracy using a logistic regression model as the base. Based on the accuracy change and feature ranking, it was decided to keep all features for modeling purposes.

*Figure 5 – Feature ranking using RFE*

A colorful graph with text

Description automatically generated with medium confidence

*Figure 6 – Model accuracy vs number of features used*

A graph with blue lines and numbers

Description automatically generated

# Methods

Logistic regression, random forest classification, and XGBoost classification algorithms were chosen based on the nature of the problem. Since the features are ordinal numerical values and the target is a binary feature, these models are the most suitable. Accuracy, precision, recall, and F1 score are typical metrics used to assess model performance. One of the challenges is that the breast cancer detection model must have high sensitivity to detect as many positive cancer cases as possible. Therefore, the focus is on sensitivity and accuracy for this model.

Additionally, a confusion matrix was built to visualize False Positive and False Negative cases better. Random Forest and XGBoost classifications have advanced performance improvement steps called hyperparameter tuning. This technique allows for iteratively applying different model parameters to find the parameters with the highest model accuracy. For the final model, a Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) was built to assess the model's ability to predict positive and negative classes. The dataset was split into training and test subsets with a 75/25 ratio. Due to its computational intensity, hyperparameter tuning using grid search cross-validation from the Sklearn library was conducted using GPU hardware.

# Analysis

All suggested models were successfully built, and their performance was assessed using the metrics described. The table below provides a comparison of these metrics:

*Table 2 – Classification models metrics*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Precision | | Recall | | F1 score | | Support | |
|  | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 |
| Logistic Regression | 0.96 | 0.96 | 0.98 | 0.93 | 0.97 | 0.94 | 89 | 54 |
| Random Forest Classifier | 0.97 | 0.96 | 0.98 | 0.94 | 0.97 | 0.95 | 89 | 54 |
| XG Boost | **0.98** | **0.98** | **0.99** | **0.96** | **0.98** | **0.97** | 89 | 54 |

It is evident that XGBoost, with fine-tuned hyperparameters, performs best among all trained models. The trained model was saved for future use. Although the Random Forest model was not selected as the final model for breast cancer detection, it provides insightful information into feature importance ranking. This ranking correlates with the RFE ranking and offers an orthogonal assessment of feature importance, which is useful for experts considering what lab tests are most valuable for cancer detection.

*Figure 7 – Random Forest Classifier features a ranking*

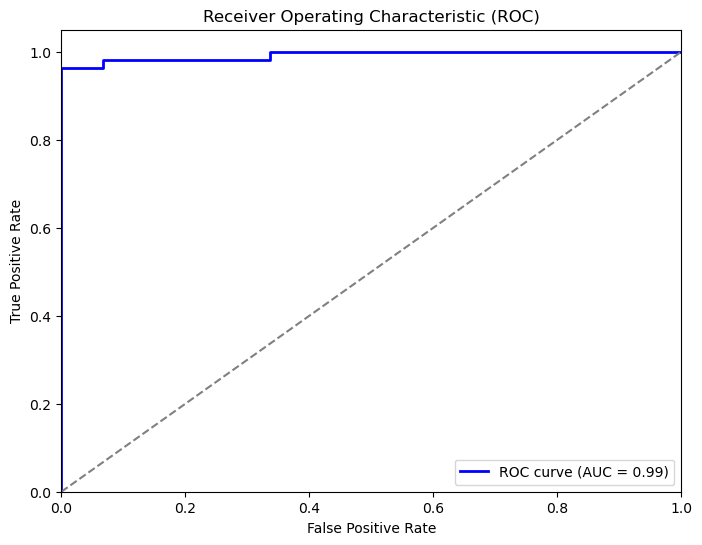
A graph with blue and black text

Description automatically generated

# Conclusion

The selected model has high values of precision, recall, and F1 score for the test data subset – Precision = 0.98, Recall = 0.96, F1 Score = 0.97. The ROC curve displayed below shows the relation between false positive and true positive rates.

*Figure 8 – ROC curve*



The area under the curve confirms good separability between positive and negative classes, performing well in classifying positive cases. Despite the high accuracy and precision, the model falsely labeled four benign cases as malignant and two as benign. This presents a significant challenge for deploying the model in actual applications, as these misclassifications could impact patients' health if professionals make incorrect decisions based on the model's predictions. Therefore, it is recommended not to rely on the prediction results overly; each patient should receive careful assessment and treatment in combination with the model's outcomes. This model can be improved by introducing a larger data set. The data set on which it was trained is insufficient to ensure its robustness with new unseen data. Suppose good robustness can be proved by validating the model on new larger data. In that case, this model can be deployed in cancer clinics and research centers to support accurate and automated breast cancer detection.

Ethical considerations for dealing with patients' data include maintaining the confidentiality of their results. The data used for model training should be of high quality and integrity to avoid unfair treatment of certain groups and inaccurate treatment of patients. Patients should provide informed consent and be educated on how their data might be used. Hospitals must operate with legal compliance in handling patient data.

**Appendix**

**Supplemental visualizations**

Confusion matrix for Logistic Regression model

A blue squares with white text

Description automatically generated

Confusion matrix for Random Forest Classification model

A blue squares with white text

Description automatically generated

Confusion matrix for XGBoost model

A blue squares with white text

Description automatically generated

**Questions and answers**

1. How can this model be deployed?

There are several steps required for the model deployment. First, all required files should be packaged together using Docker for example. Then, an API should be created for the model inference requests. Then, depending on the scale of the model, it can be deployed on a cloud or internal resources. The model certainly will require maintenance and monitoring to ensure robustness and performance.

1. Are the data chosen for the modeling enough to build an accurate model?

In this case, specific parameters of the cell nuclei dictate what is important to determine the accurate outcome of the analysis. Based on the results, we can say that this data is enough; however, the model was trained on a relatively small data set, about 500 entries. To ensure generalization and model performance it can be retrained on a larger data set.

1. Will normalizing features help with model performance improvement?

Normalization certainly can improve model performance by speeding up the convergence rate. It potentially can improve the model accuracy by normalizing the features to the same scale. However, the model accuracy was assessed pre and post-normalization, and accuracy for this specific model is higher without normalizing the features. This might be due to the amplification of the less relevant features.

1. Can further broadening the grid search for hyperparameter tuning improve model performance?

It is possible that several hyperparameters were not included in the grid search of the XGboost and Random Forest models. However, the accuracy of the model is already at its highest value. Further hyperparameter tuning might be more costly but not necessarily beneficial.

1. Why does the XGBoost model perform better than the Random Forest model for this data set?

Perhaps the XGboost is a more advanced model than the random forest because it utilizes algorithms that focus on error correction from previous trees during training. Also, it does regularization that prevents overfitting and can handle imbalanced data.

1. Are there any additional features that can be added to improve accuracy?

This subject can be researched deeply because it requires subject matter experts to identify any other parameters of the nuclei that will support more accurate outcome prediction.

1. Are there any other algorithms that can be considered for modeling?

Those three were selected as the most suitable for the purpose. However, Neural Networks can also be utilized for classification problems.

1. What impact does this model have on society?

This model helps improve human health by providing accurate early-stage cancer detection. The model can be scaled and implemented in any cancer research laboratory.

1. Will balancing classes to a 50/50 ratio help with accuracy improvement?

Based on the target class ratio of 40/60, it is balanced.

No further balancing was needed in this case. The algorithms are advanced enough to handle this case, so balancing the classes to a 50/50 ratio would not significantly improve the model’s performance.

1. Can you look into falsely classified cases and identify why the model labeled them incorrectly?

For future work, it might be beneficial to see what causes the false positive and false negative cases. Looking further into those details may help with model adjustments to achieve higher accuracy.

References

U. M. L. (n.d.). *Breast Cancer Wisconsin (Diagnostic) Data Set*. Www.Kaggle.com. https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data

Rabiei, R., Ayyoubzadeh, S. M., Sohrabei, S., Esmaeil, M., & Atashi, A. (2022, June 1). *Prediction of Breast Cancer using Machine Learning Approaches*. Www.Kaggle.com. Retrieved June 20, 2024, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9175124/