RESEARCH ARTICLE

SPECIAL ISSUE

RSP Science Hub

International Research Journal on Advanced Science Hub
2582-4376
Vol. 05 May, Issue 05S
www.rspsciencehub.com

Check for updates

http://dx.doi.org/10.47392/irjash.2023.S012

International Conference on intelligent COMPUting TEchnologies and Research (i-COMPUTER) 2023

Breast Cancer Classification Using Machine Learning

Kiran R [1], T M Rajesh[2], Manav G Krishna[1], Nanda Gopal[1], Kishan G [1]

[1]Department of Computer Science and Engineering, Dayananda Sagar University, Karnataka, India

[2]Associate Professor, Department of Computer Science and Engineering, Dayananda Sagar University, Karnataka, India

Email: kiranmadhu1239@gmail.com

Article History

Received: 26 February 2023
Accepted: 9 March 2023

Keywords:

breast cancer;
benign;
malignant;
accuracy

Abstract

Manual determination of breast cancer takes a lot of time sometimes weeks or months and is perilous with a high pace of morbidity, mortality and can involve human error. an early assurance will help the treatment of this sickness. So, determination of breast cancer using machine learning model result in quick identification of tumor. This research paper focuses on early determination of breast cancer as benign or malignant. with the help breast cancer dataset, the research paper aims to produce a better decision-making visualization pattern through swarm plots and heat maps. To accomplish this, we utilized Light GBM Calculation and furthermore contrasted our model's exhibition and other surviving ML models specifically Logistic Regression, Gradient Boosting Algorithm, Random Forest Algorithm and XG Boost Algorithm. We were able to achieve the highest accuracy of 97.07% with the Light GBM Algorithm.

1. Introduction

Quite possibly one of the most common tumor's affecting people today, particularly women, is breast cancer. An early location would greatly lessen the harm that this cancer causes to its victims. Family ancestry heftiness, chemicals, radiation treatment and surprisingly, regenerative factors are among the multifactorial reasons for breast cancer. As per a study by the World Health Organization, a big part of 1,000,000 women dies because of the determination of the infections. (Simon et al.)

Found lately. There are two types of breast cancer: malignant breast cancer and benign breast cancer. By carefully examining the characteristics of breast tumors, lumps, or other anomalies seen in the breast, breast cancer can be classified as benign or malignant. Cancer that is delegated benign has a lower endanger and is not hazardous, yet disease that is named malignant is dangerous. Dissimilar to harmless masses, which cannot spread to different tissues and can develop inside the harmless mass, dangerous growths spread to the adjoining cells and can hence spread to different regions. scientists have utilized AI, a part of computerized reasoning (man-made intelligence), to accurately classify breast cancer as harmless or dangerous. AI calculations are used train the model that accepts dataset features as input. And with the help of the features determining the tumors and giving the label which is 1 for harmless and 0 for dangerous that is malignant. (Bardou, Zhang, and Ahmad) The motivation of the paper is to find a way for early diagnosis of the tumor because the conventional way of diagnosis is time consuming and can involve human errors. So, we implement machine learning model to diagnose it at the early stage.

1.1. Problem Definition

At present around 1,78,000 cases of breast cancer cases are reported around India. So manual determination of cancer is very time consuming, dragging, it can involve human errors. So, we are going to build a predictive model and classify the tumors as malignant or benign. And this is achieved by using Machine Learning models in which we will see the features correlation and remove the redundant data and get the highest accuracy model as possible.

1.2. Objectives

The first objective of this research is to analyze breast cancer data from a diagnostic dataset which contain 30 column that is features and around 570 rows. The goal is to find similar characteristics in the groups that indicate good sharp characteristics between benign and malignant. And next objective is to visualize the heat map and remove the redundant features and the final objective is to develop a machine learning model that helps the user to classify whether it is a benign or malignant.

1.3. Scope

Our project identifies difficulties and offers solutions to improve the accuracy. One of the biggest challenges in classification is accuracy. Inaccurate model usually results in poor output. In addition, the research focuses on solving problems related to increasing the accuracy of targeting different algorithm namely Logistic Regression, Gradient Boosting Algorithm, Random Forest Algorithm (Octaviani and Z Rustam), XG Boost Algorithm, Light GBM Algorithm to achieve the best accuracy of the model.

2. Methodology

The methodology of this research helps to understand the differences between benign and malignant cancer. So firstly, we collect the breast cancer data from diagnostic dataset. we will preprocess the dataset check for null values and remove null values if any present. then we will visualize and compare the features with the help of swarm plots and see whether there is a sharp difference between the benign and malignant and remove the outliers that is features and then with the help of heat map we will remove the redundant features and drop it off. After removal of outliers, we split the preprocessed data into train data and test data and then by training the data with different machine learning models like Logistic Regression (Sultana and Jilani), Gradient Boosting Algorithm, Random Forest Algorithm, XG Boost Algorithm, Light GBM Algorithm and check for the highest accuracy giving model and lastly build a predictive system based on the highest accuracy giving model.

2.1. Breast Cancer Dataset

The diagnostic dataset used for this research consists of 569 rows and 30 columns. The 30 parameters considered for this research based on the dataset. These attributes in the data helps to produce visualization patterns easily and to do heat maps for the visualization of the features.

2.2. Data Cleaning Procedure

After importing the dataset using the panda's library, it is important to find the existence of any missing values in the dataset. The data cleaning procedure involves removing the entire row having any missing value. By this procedure, the subsequent activities like visualization can be performed efficiently with good accuracy. The heat maps show the outliers which are removed to improve the accuracy.

3. Results and Discussion

The results show that Light GBM algorithm is best suited for classification of the cancer as benign or malignant. It gives an accuracy score of 97.07%. The machine learning models that were used to find the accuracy for our dataset included Logistic Regression, Gradient Boosting, Random Forest, XG Boost, Light GBM Algorithm (Derangula et al.) to achieve the best accuracy of the model.

The count plot shows the count of benign and malignant in the dataset are displayed as shown in the FIGURE 1.

The output of swarm plots graphs for visualization of the first 5 features namely mean area, mean radius, mean texture, mean perimeter, mean smoothness out of the 30 features to check out the relation between them.

The output of swarm plots graphs for visualization of the 5 to 10 features namely mean fractal dimension, mean compactness, mean concavity, mean concave points, mean symmetry. out of the 30 features to check out the relation between them.

The output of swarm plots graphs for visualization of the 10 to 15 features namely texture error,

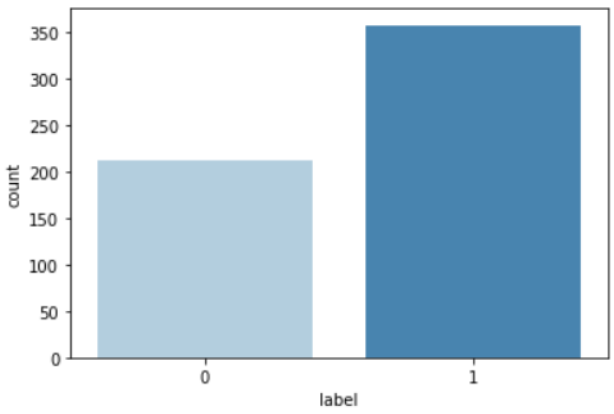

FIGURE 1. Count chart for benign (1) and malignant (0)

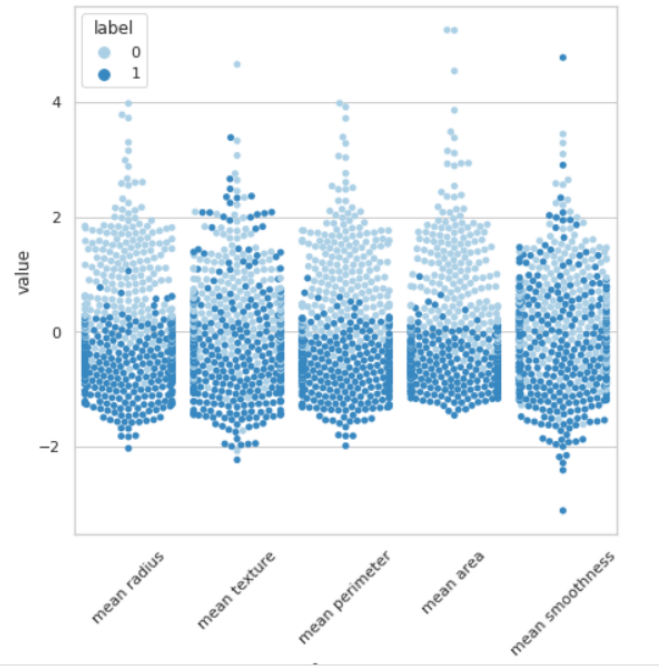

FIGURE 2. Swarm plot for 5 features of dataset

radius error, smoothness error, perimeter error, area error out of the 30 features to check out the relation between them.

The output of swarm plots graphs for visualization of the 15 to 20 features namely symmetry error, compactness error, concavity error, fractal dimension error, concave points error. out of the 30 features to check out the relation between them.

The output of swarm plots graphs for visualization of the 20 to 25 features namely worst radius, worst texture, worst perimeter, worst area, worst smoothness. out of the 30 features to check out the relation between them.

The output of swarm plots graphs for visualization of the 25 to 30 features namely worst compactness, worst concavity, worst symmetry, worst fractal dimension, worst concave points. out of the 30 fea-

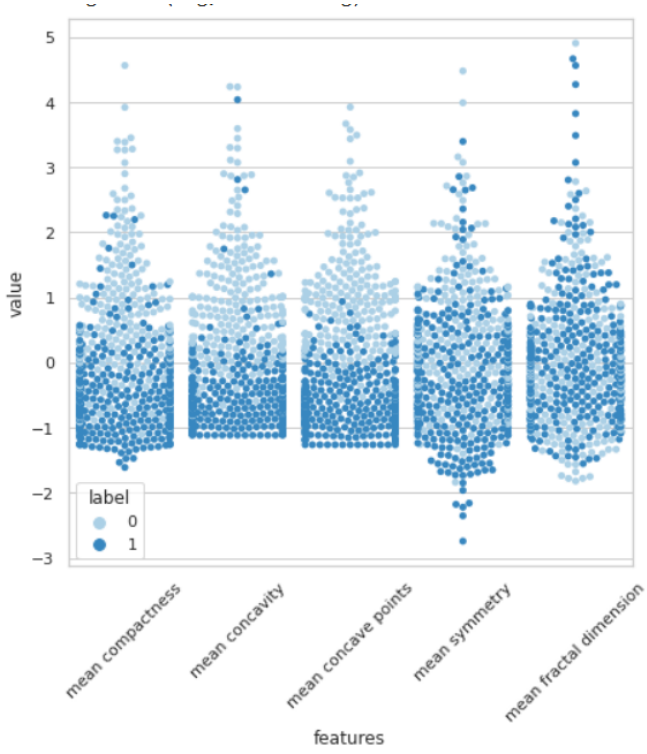

FIGURE 3. Swarm plot for 5-10 features of dataset

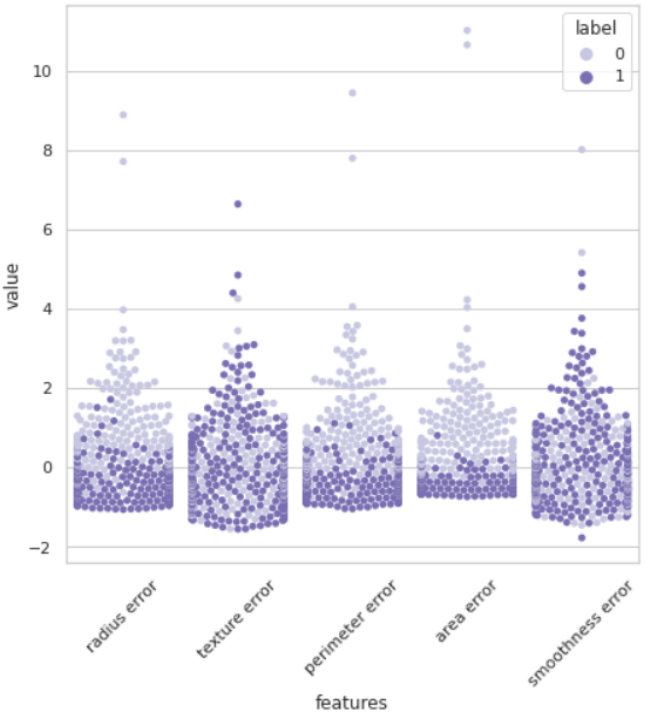

FIGURE 4. Swarm plot for 10-15 features of dataset

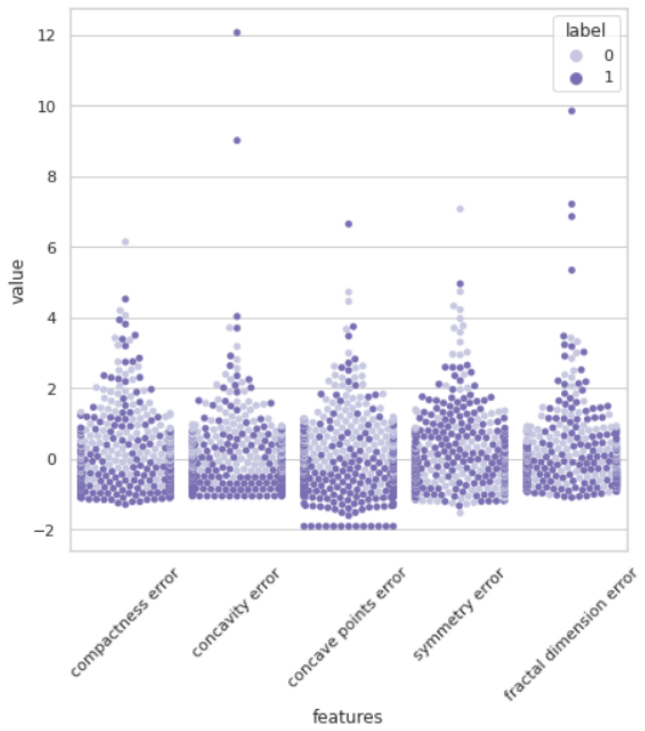

FIGURE 5. Swarm plot for 15 - 20 features of dataset

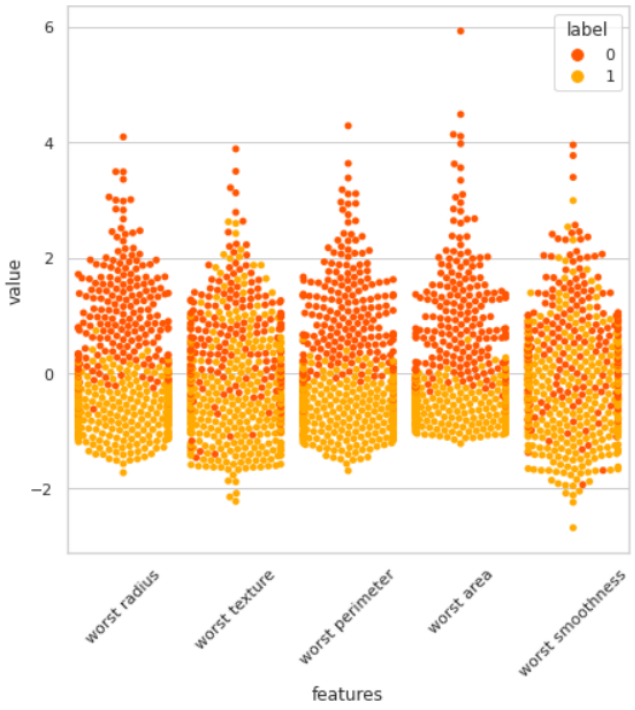

FIGURE 6. Swarm plot 20 -25 features of dataset

tures to check out the relation between them.

Figure 8 shows the heat map with all the features name and correlation between them

Figure 9 shows the heat map with all the redundant feature name removed and correlation between

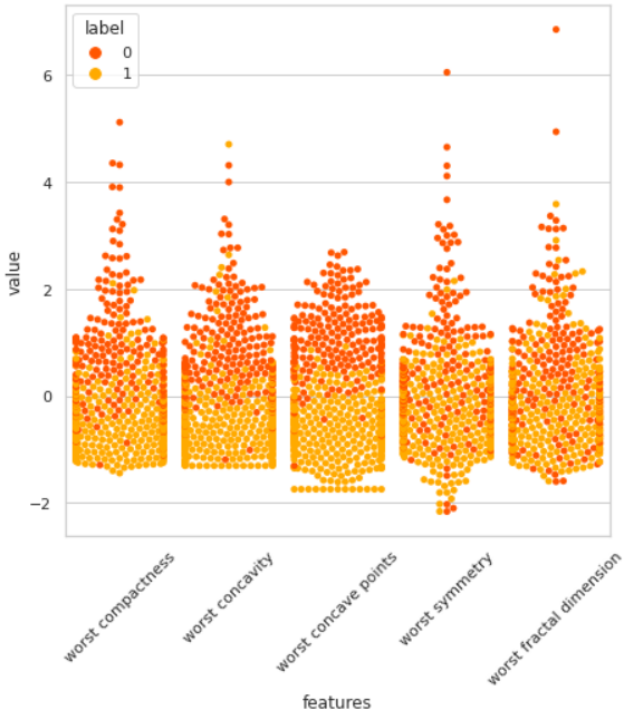

FIGURE 7. Swarm plot for 25 - 30 features of dataset

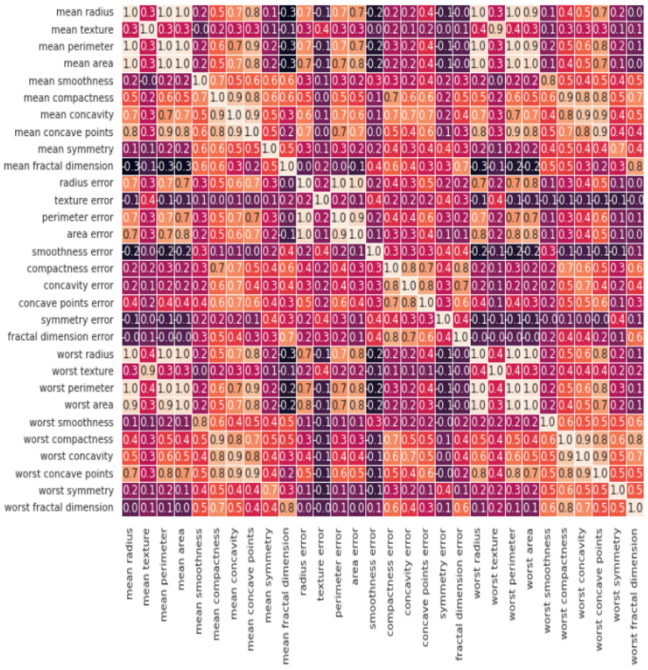

FIGURE 8. Heat map for 30 features of dataset

them and the features considered for training the data are namely mean fractal dimension, worst symmetry, mean texture, mean area, mean smoothness, mean concavity, worst fractal dimension, mean symmetry, area error, smoothness error, concavity error, symmetry error, fractal dimension error, worst smoothness, worst concave points, texture error.

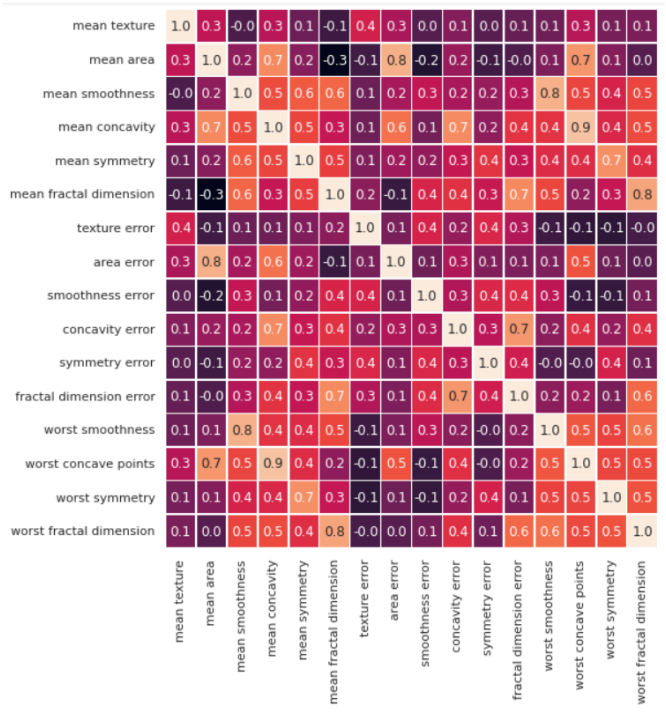

FIGURE 9. Heat map for 16 features of dataset

These above features are taken into consideration for the training of the model that is algorithm and train the model and then to get the highest accuracy possible.

4. Conclusion

The result that we came to know is that Light GBM algorithm is efficient and easy to implement while working on diagnostic dataset. Where from 30 features on removing the outliers, we obtained 16 features were the ones that were relevant to contributing immensely to the final accuracy of the model. Among all the five algorithms that were used, Light GBM provided us with the highest accuracy of 97.07%. Logistic Regression which gave accuracy of 92%, Random Forest Algorithm which gave accuracy 80%, and XG Boost Algorithm which gave accuracy 83%. Which compared to other literatures out there which gave around 78.6 % accuracy achieved using a MLP classifier where the dataset is diagnostic dataset.

The other comparison where Afshar the author studied about the survival of breast cancer patients using a dataset with 856 rows and 15 columns using machine learning models. The obtained accuracy is 84% (Afshar et al.).

5. Authors' Note

We declare that there is no conflict of interest regarding the publication of this article. We confirm that the paper is free of plagiarism.

References

Afshar, Lotfnezhad, et al. "Prediction of Breast Cancer Survival by Machine Learning Methods: An Application of Multiple Imputation". *Iran J Public Health* (2021). 10.18502/ijph.v50i3.5606.

Bardou, Dalal, Kun Zhang, and Sayed Mohammad Ahmad. "Classification of Breast Cancer Based on Histology Images Using Convolutional Neural Networks". *IEEE Access* 6 (2018): 24680–24693. 10.1109/ACCESS.2018.2831280.

Derangula, Anusha, et al. "Feature Selection of Breast Cancer Data Using Gradient Boosting". (2020). https : / / ejmcm . com / article_2569_a7c3766658c69272f70820a964a9cd36.pdf.

Octaviani, T L and Z Rustam. "Random forest for breast cancer prediction". *PROCEEDINGS OF THE 4TH INTERNATIONAL SYMPOSIUM ON CURRENT PROGRESS IN MATHEMATICS AND SCIENCES (ISCPMS2018)* (2019). 10 . 1063/1.5132477.

Simon, Michael S, et al. "Cardiometabolic risk factors and survival after cancer in the Women's Health Initiative". *Cancer* 127.4 (2021): 598–608. 10.1002/cncr.33295.

Sultana, Jabeen and Abdul Khader Jilani. "Predicting Breast Cancer Using Logistic Regression and Multi-Class Classifiers". *International Journal of Engineering & Technology* 7.4.20 (2018): 22–22. https : / / www . researchgate . net / publication / 331233978_Predicting_Breast_Cancer_using_Logistic_Regression_and_Multi - Class_Classifiers.

Embargo period: The article has no embargo period.

To cite this Article: , Kiran R, T M Rajesh, Manav G Krishna, Nanda Gopal, and Kishan G . "**Breast Cancer Classification Using Machine Learning**." International Research Journal on Advanced Science Hub 05 May.05S (2023): 88–93. http://dx.doi.org/10.47392/irjash.2023.S012