

Breast Cancer Prediction by Machine Learning Algorithms - A Comparative Study of Naive Bayes, KNN and J48 in Weka Environment

Trishit Banerjee^a, Geetha Ganesan^b

^a Netaji Subhash Engineering College, Techno City Garia Kolkata- 700152, India

^b Advanced Computing Research Society, Chennai, India

Abstract

Breast cancer is considered one of the most common cancers occurring in women. Each year, it affects 2.1 million women, causing cancer-related deaths. As per the estimation, breast cancer took the lives of 627,000 women in 2018 alone. The disease is mostly observed in the developed areas of the world, with current rates expanding in almost every region across the world. The role of predicting cancer is crucial for the further progress of data mining tools currently available. K-nearest neighbor, J48 algorithm, and Naive Bayes are applied to predict cancer disease. For acquiring an extensive dataset, Naive Bayes is very helpful and easy to design. K-nearest neighbor produces a dataset, separating it into distinct categories. It predicts new points in the classification. Grounded on the Decision Tree, J48 Classifier uses such facts from the datasets of training, which can be utilized to decide the minor subsection. To measure the precision of the cancer dataset, the Weka tool can be applied. Its dataset encompasses nine kinds of cancer. A 70% train and 30% test data set split has been utilized to predict the cancer disease. The exactness of Naive Bayes is 91.81%, whereas the identity of J48 and K-nearest neighbor is respectively 92.98% and 97.07%.

Keywords

Breast Cancer, K-nearest neighbor, j48 algorithm, Naive Bayes

1. Introduction

1.1. Background

Early diagnosis is essential to enhance the results and survival rates of breast cancer. There are two strategies for early detection of breast cancer, including screening and early diagnosis. The settings with limited resources and poor health organizations where most women are detected in advanced stages must focus on the early detection programs grounded on the consciousness of early symptoms and quick recommendations for analysis and treatment. Screening comprises assessing women for identifying the risks of cancer before the appearance of symptoms. These screening tools include breast self-exam, clinical breast exam, and mammography. Since it requires significant investment, the decision to continue with the screening procedure should be made following fundamental breast health amenities that include efficient detection and appropriate treatment. Early diagnosis involves delivering apt access to cancer treatment by decreasing the barriers that come before enhancing access to proper diagnosis services. The purpose is to study the comparison of Naive Bayes, KNN, and J48 classifier accuracy in predicting the proportion of breast cancer identification in the early phases [5, 12]. Cancer is a genetic disorder that happens due to the alterations to the genes and the sudden expansion of cells and division. Metastatic cancer is when the cancer cells spread to another place

CVMLH-2022: Workshop on Computer Vision and Machine Learning for Healthcare, April 22 – 24, 2022, Chennai, India.

EMAIL: gitaskumar@yahoo.com (Geetha Ganesan)

ORCID: 0000-0001-7338-973X (Geetha Ganesan)

 © 2022 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

within the body from the location where it developed. This spread of cells from one part to another is known as metastasis.

1.2. Motivation

The data mining technique is expanding extremely speedily in the medical ground because of its accomplishment in classification and prediction processes that aid the experts in decision-making. We are searching for ways to improve the patients' health and decrease the expense of medicine, and data mining assists a lot in this case. A few papers are accessible to predict the disorder, and the most crucial part is that it is merged with the prediction of the existence of the specific cancer disease. Many types of cancers are still unknown. Doctors are sometimes unable to find out the reason behind the disease. For curing the disorder, early detection is needed, and undertaking the prediction research is critical. Here we utilize the open-source data mining tool Weka, which was created at Waikato University, New Zealand. It assists us in predicting the cancer disease precisely and aids in proper decision-making. We have used renowned classification procedures called K-Nearest, J48, and Naïve Bayes. A dataset of breast cancer diagnosed patients is used to show an apt answer.

1.3. Paper Organization

The present paper has seven sections that initiate with the introduction segment and is followed by Related Work, Problem Statement, Data and Classifier Details, Methodology, Results and Analysis, and Conclusion sections.

2. Related Work

Scientists are striving hard to reduce the consequences of malignancy. Thus, there are multiple queries regarding predicting the survivability of cancer. Thousands of people pass away due to the most frequent breast cancer. It occurs in humans owing to damaged genes caused because of increasing age. Age activates a combination of factors generating mutations, which, in turn, develops tumors. Thus, early diagnosis is necessary; hence, designing a genetic mutation-based strategy to predict cancer has become essential.

The Clustering procedure and diverse classification procedures were utilized by Dona Sara Jacob et al. [6]. The results show that the classification algorithms were superior to the clustering procedure. Grounded on D. Support Vector Machine, studies evaluated all the procedures, indicating that Naïve Bayes was faster than the SVM model. The latter is an ML technique based on K-nearest neighbor and decision tree. Among the well-known data mining processes utilized, four belonged to E.K-means clustering procedures. Research proves that classification procedures predict better than clustering procedures in the case of predicting breast cancer. The most acceptable breast cancer prediction algorithm is decided on the exactness of the procedure.

Alom et al. [1] used the Inception Recurrent Residual Convolutional Neural Network to classify medical images of breast cancer. The specified neural network exhibited higher performance than parallel Inception Networks, RCNNs, and Residual Networks. Satapathi et al. [10] provided a prediction modeling that uses transformed genetic cells due to abrupt abnormalities causing carcinogenic cells in the human body. Saritas & Yasar [9] studied the classification performance of Naïve Bayes classifiers for data containing nine inputs and one output to compare with the Artificial neural network. They worked on ANN and Naïve Bayes classification algorithms-based data classification to predict the preeminent scope for breast cancer detection using anthropometric data and standard blood tests results. Kumar et al. [7] work focused on various forms of data mining technologies to detect malignant and benign breast cancer. The UCI data set was used during clump thickness as an assessment level of Breast Cancer. Twelve algorithms, including J48, Lazy IBK, Logistics Regression, and Naïve Bayes' performances, were evaluated.

According to Rashmi et al. [8], data mining refers to the method of repossessing the details from an enormous dataset. For extracting helpful information, the data ought to be appropriately arranged. The method is utilized to explore a considerable quantity of data for finding some constant patterns. The paper delivers an assessment of the techniques of Prediction and Classification. Breast cancer signifies 12% of fresh cases of this disorder. It is noted to be the second-most occurred cancer globally. It becomes essential to detect the tumor type if diagnosed in the early stages. Pathologists can view the microscopic structures and components of the breast tissue histologically through a breast tissue biopsy. Those histological photographs allow the pathologist to differentiate between normal tissue, non-malignant (benign) tissue, and malignant lesions.

3. Problem Statement

Breast cancer is the most prevalent form of cancer, where a painless, hard lump is the most prominent symptom of the breast. Like most tumors, breast cancer can be best treated when diagnosed early. Early detection of breast cancer raises the number of viable treatment choices and dramatically increases the probability of clinical success and recovery. About 70% of healthy tumors are reported with this method, but between 4% and 34% of breast carcinoma cases cannot be diagnosed by mammography. A number of researches on artificial intelligence, computer learning, and data processing were reviewed to identify breast cancer. Danacı et al. [3] used pattern recognition to detect breast cancer cells using the C4.5 algorithm in the Waikato Environment for Knowledge Analysis (Weka) tool.

The present study aimed to gain a comparative insight into decision-making processes by the classification of breast cancer data, such as naive Bayes, naïve Bayes, and J48 decision trees. In this study, the breast cancer dataset was taken from Kaggle, which Syeda Daraqshn contributed, and was examined using the Weka tool [4]. The effectiveness levels of J48, K-nearest neighborhood, and Naive-Bayes data mining algorithms have been compared.

4. Data and Classifier Details

4.1. Data Details

The cancer patients' data were collected from the Kaggle website created on 28/08/2020 [4]. There are data from 569 instances that contain 30 attributes and a single class attribute. The 30 attributes include symptoms and the stages of breast cancers (benign/ malignant) in .csv format. The dataset was divided into train and test datasets with a 70-30 split in Weka. The 30 attributes can be classified in mean, se, and worst categories for the ten specified attributes (radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension).

4.2. Classifier Details

The most heard term in Weka is a classification that assists in decision-making. Classification can be divided into two parts: multi-class targets and binary. The latter comprises two kinds of outcomes, whereas the former aims at delivering superior two values. The key objective is to categorize and forecast accurately. Forecasting relies on some associated data that can inform us about further happenings. Prediction develops a connection between the data that people are aware of and the data that people require forecasting in the future. Among several types of classifiers, three different types are utilized in this paper K-Nearest Neighbor, J48 algorithm, and Naïve Bayes. We used Weka 3.9 version and Windows 10 OS.

Naïve Bayes is a straightforward technique regarded as "probabilistic classifiers" [11]. It encompasses an amalgamation of procedures that share standings, on which each aspect is categorized

self-reliantly from other presented values. It is executed in the supposition that the impression of a specific value does not rely upon other characteristic values. This particular algorithm is very helpful to acquire a big dataset and extremely easy to develop.

K-Nearest Neighbor (KNN) is a non-parametric and indolent procedure that utilizes data sets and produces a new dataset, separating it into distinct categories and forecasting new points in classification [2]. K-nearest neighbor, also known as instance-based learning, memorizes the opinion to categorize the unnoticed test data. It makes a comparison between the training observations and the test observations.

J48 utilizes a non-proprietary Java incorporation C4.5 procedure. For instance, if we have a dataset that comprises dependent variables and the other encompasses independent variables, the application of the decision tree such as J48 enables us to forecast new records. It acts fine on incessant and distinct also missed out data values. It even provides a choice for cropping trees after generation. The classifier avails of the avaricious process also decreased the fault. The test condition outcome is displayed as a branch, and a class tag is given for every final node. Typically, the root node is the largest decision tree node. Any path is an adjective concept in a decision tree. In two steps, it typically uses the depth-first method. Tree preparation in the top-down technique takes place. The layout is separated repetitively at this point before the data elements belong to the same class plate.

5. Methodology

The following figure presents the process of the classification algorithms with the training dataset. A biopsy has to be performed if symptoms characteristics indicate malignancy. Otherwise, there will be no requirement for the biopsy test in the case of benign characteristics. The data mining process involves two sections. First, the classification models were trained with the 70% split train dataset. Finally, prediction using the test dataset has been noted.

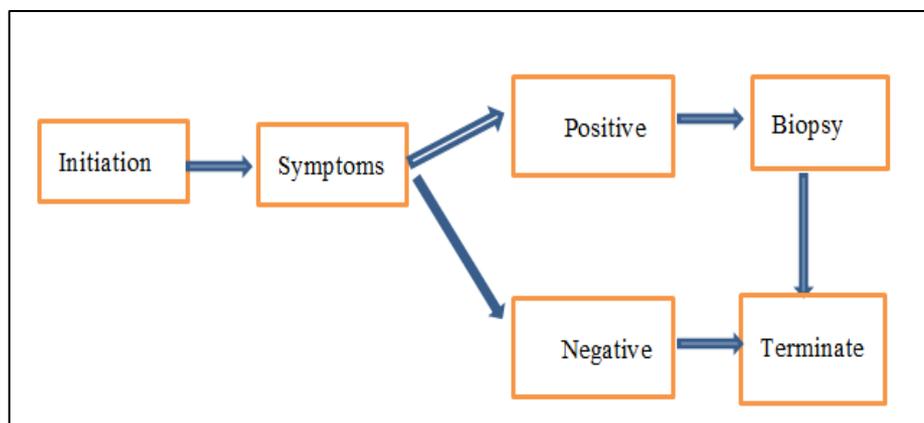


Figure 1: Classification flow of Prediction of Breast Cancer

6. Results and Analysis

In this study, three classification algorithms have been used to compare their performances regarding accuracy in correctly predicting the instances in Weka 3.9 version. The best algorithm has been chosen based on the best accuracy of prediction. The benign and malignant classes have been analyzed based on error rate, accuracy, precision, F-score, sensitivity, and specificity. Whereas sensitivity will indicate the true value, specificity indicates the correct negative cases. A cost-benefit analysis has discussed the confusion matrices for all three cases. The confusion matrices were interpreted in terms of TP, FP, TN, FN (T = True, P = Positive, N = Negative, F = False).

6.1. Data Exploration

Exploration of 569 instances revealed no missing value present for any attribute. There were 212 malignant and 357 benign cases.

6.2. Performance Exploration of the Classifiers

Naïve Bayes classifiers are non-deterministic in nature that can handle noisy big data. The algorithm is generally faster than the KNN classifier and is based on the hypothesis that all the factors correlate among them a contribution to the classification. In the present scenario, the algorithm correctly predicted 91.81% instances (n = 157) from the test dataset of 171 total instances. The precision of prediction was higher for benign cases compared to malignant tumors. The confusion matrix revealed that 59 instances were predicted correctly out of 65 for malignant cases, whereas 98 instances for benign cases were properly predicted out of 106. The following table provides a detailed, accurate description of the two classes. The root mean square (RMSE) error was 0.29, and the relative absolute error (RAE) was 17.49%.

Table 1: Detailed Accuracy of Naïve Bayes Classifier on Breast Cancer Data

Class	TP-Rate	FP-Rate	Precision	Recall	F-Measure	MCC	ROC-Area	PRC-Area
M	0.91	0.08	0.88	0.91	0.89	0.83	0.97	0.93
B	0.93	0.09	0.94	0.93	0.93	0.83	0.97	0.99

K-Nearest Neighbor (Lazy IBK) is a non-deterministic classifier that is generally slower for huge data sets and refuses to deal with the noisy dataset. The KNN algorithm for N = 1 predicted 95.91% of instances from the test dataset. The classifier produced an optimum prediction of 97.07% correctly classified instances at N = 4. The precision of prediction was higher for benign cases compared to malignant tumors. The confusion matrix revealed that 63 instances were predicted correctly out of 65 for malignant cases, whereas 103 instances for benign cases were properly predicted out of 106 instances. The following table provides a detailed, accurate description of the two classes. The RMSE was 0.16, and the RAE was equal to 10.10%.

Table 2: Detailed Accuracy of KNN (N = 4) Classifier on Breast Cancer Data

Class	TP-Rate	FP-Rate	Precision	Recall	F-Measure	MCC	ROC-Area	PRC-Area
M	0.97	0.03	0.96	0.97	0.96	0.94	0.98	0.98
B	0.97	0.03	0.98	0.97	0.98	0.94	0.98	0.98

J48 Decision Tree is a deterministic classifier that can deal with a noisy large dataset with high accuracy. The J48 is a depth-first or breadth-first approach technique consisting of roots nodes and internal and leaf nodes. In the present study, the classifier produced an optimum prediction of 92.98% correctly classified instances. The precision of prediction was higher for benign cases compared to malignant tumors. The confusion matrix revealed that 58 instances were predicted correctly out of 65 for malignant cases, whereas 101 instances for benign cases were properly predicted out of 106. The following table provides a detailed, accurate description of the two classes. The RMSE was 0.26, and the RAE was equal to 17.72%.

Table 3: Detailed Accuracy of J48 Classifier on Breast Cancer Data

Class	TP-Rate	FP-Rate	Precision	Recall	F-Measure	MCC	ROC-Area	PRC-Area
M	0.89	0.05	0.92	0.89	0.91	0.85	0.91	0.87

7. Conclusion

This paper is chiefly grounded on the medical dataset that can forecast Cancer prevalence from the symptoms. In this case, three different classification procedures are used to validate the dataset. For fulfilling this purpose, we utilized the Weka 3.9 tool. The application consists of classification and data exploration. The dataset does not create any viciousness as it does not consist of any personal details. It comprises a few medical info. From the evaluation, we can state that KNN did the most accurate classification for N= 4, which could correctly classify almost 97% of instances. For detecting the confusion matrix, three distinct algorithms are utilized. It comprises details regarding the actual and forecasted classification. There are two stages of this disorder- the one is malignant, and the other one is benign stages. For having a clear perception, tables are needed to be created. It compares the three different classification procedures, namely K-Nearest Neighbor, Naïve Bayes, and J48 algorithms. The contrast table lucidly states which model of classification will be better. All three algorithms function outstandingly; however, in this study, K-Nearest Neighbor has worked more supremely than the other two procedures: the J48 algorithm and Naive Bayes. In the upcoming days, we will strive to revolutionize the technologies and equipment for more significant improvement, expand datasets, and implement distinct preparation, clustering, classification, visualization, and regression. In addition, we will act in advancing the new models' survivability and predictive capabilities.

8. References

- [1] Alom, M., Yakopcic, C., Nasrin, M., Taha, T., Asari, V.: Breast Cancer Classification from Histopathological Images with Inception Recurrent Residual Convolutional Neural Network. *Journal of Digital Imaging*. 32, 605-617 (2019).
- [2] Bharati, S., Rahman, M., Podder, P.: Breast cancer prediction applying different classification algorithm with comparative analysis using WEKA. 4th International Conference on Electrical Engineering and Information & Communication Technology (iCEEICT). IEEE (2018).
- [3] Danacı, M., Çelik, M., Akkaya, A.: Prediction and diagnosis of breast cancer cells using data mining methods. ASYU'2010. pp. 9-12. , Kayseri, Turkey (2010).
- [4] Daraqshan, S.: Kaggle: Your Machine Learning and Data Science Community, 8. <https://www.kaggle.com/syedadaraqshan/breast-cancer-predicition-using-machine-learning>.
- [5] Dubey, A., Gupta, U., Jain, S.: Comparative Study of K-means and Fuzzy C-means Algorithms on The Breast Cancer Data. *International Journal on Advanced Science, Engineering and Information Technology*. 8, 18 (2018).
- [6] Jacob, D., Viswan, R., Manju, V., PadmaSuresh, L., Raj, S.: A Survey on Breast Cancer Prediction Using Data Mining Techniques. In 2018 Conference on Emerging Devices and Smart Systems (ICEDSS). pp. 256-258. IEEE (2018).
- [7] Kumar, V., Mishra, B., Mazzara, M., Thanh, D., Verma, A.: Prediction of Malignant and Benign Breast Cancer: A Data Mining Approach in Healthcare Applications. *Advances in Data Science and Management*. pp. 435-442. Springer, Singapore (2020).
- [8] Rashmi, G., Lekha, A., Bawane, N.: Analysis of Efficiency of Classification and Prediction Algorithms (Naïve Bayes) for Breast Cancer Dataset. 2015 International Conference on Emerging Research in Electronics, Computer Science, and Technology (ICERECT). pp. 108-113. IEEE (2015).

- [9] Saritas, M., Yasar, A.: Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification. *International Journal of Intelligent Systems and Applications in Engineering*. 7, 88-91 (2019).
- [10] Satapathi, G., Srihari, P., Jyothi, A., Lavanya, S.: Prediction of cancer cells using DSP techniques. *International Conference on Communication and Signal Processing*. pp. 149-153. IEEE (2013).
- [11] Xu, S.: Bayesian Naïve Bayes classifiers to text classification. *Journal of Information Science*. 44, 48-59 (2018).
- [12] Verma, D., Mishra, N.: Analysis and prediction of breast cancer and diabetes disease datasets using data mining classification techniques. In *2017 International Conference on Intelligent Sustainable Systems (ICISS)*. pp. 533-538. IEEE (2017).