Method Based on Data Mining Techniques for Breast Cancer Recurrence Analysis

Morales-Ortega Roberto Cesar[1], Lozano-Bernal German[2], Ariza-Colpas Paola Patricia[1](✉), Arrieta-Rodriguez Eugenia[3], Ospino-Mendoza Elisa Clementina[1], Caicedo-Ortiz Jose[1], Piñeres-Melo Marlon Alberto[4], Mendoza-Palechor Fabio Enrique[1], and Roca-Vides Margarita[1]

[1] Universidad de la Costa, CUC, Barranquilla, Colombia
{rmorales1,pariza1,eospino14,jcaicedo,fmendoza,mroca}@cuc.edu.co
[2] Universidad Simón Bolívar, Barranquilla, Colombia
glozano3@unisimonbolivar.edu.co
[3] Universidad del Sinú, Cartagena, Colombia
Investigacionsistemas@unisinucartagena.edu.co
[4] Universidad del Norte, Barranquilla, Colombia
pineresm@uninorte.edu.co

Abstract. Cancer is a constantly evolving disease, which affects a large number of people worldwide. Great efforts have been made at the research level for the development of tools based on data mining techniques that allow to detect or prevent breast cancer. The large volumes of data play a fundamental role according to the literature consulted, a great variety of dataset oriented to the analysis of the disease has been generated, in this research the Breast Cancer dataset was used, the purpose of the proposed research is to submit comparison of the J48 and randomforest, NaiveBayes and NaiveBayes Simple, SMO Poli-kernel and SMO RBF-Kernel classification algorithms, integrated with the Simple K-Means cluster algorithm for the generation of a model that allows the successful classification of patients who are or Non-recurring breast cancer after having previously undergone surgery for the treatment of said disease, finally the methods that obtained the best levels were SMO Poly-Kernel + Simple K-Means 98.5% of Precision, 98.5% recall, 98.5% TPRATE and 0.2% FPRATE. The results obtained suggest the possibility of using intelligent computational tools based on data mining methods for the detection of breast cancer recurrence in patients who had previously undergone surgery.

Keywords: Breast cancer · Data mining · Classification · Cluster · Dataset

1 Introduction

Cancer is a disease in constant evolution, which affects a large number of people worldwide. According to the world health organization 8.2 million people died from this disease where you can see the usual occurrence of cancer such as lung, liver, colon and

Y. Tan et al. (Eds.): ICSI 2020, LNCS 12145, pp. 584–596, 2020.
https://doi.org/10.1007/978-3-030-53956-6_54

breast stomach, approximately 30% of cancer deaths are due to five behavioral and food risk factors (high body mass index, insufficient consumption of fruits and vegetables, lack of physical activity and consumption of tobacco and alcohol) [1]. In Colombia, according to the report issued by the National Cancer Observatory, the mortality rate for breast cancer in Colombia on average is 11.49 * 100,000 for the year 2014 in women, in addition there is a tendency to increase for the last 10 years [2].

The diagnosis and prognosis of breast cancer is a great challenge for researchers. The implementation of machine learning and data mining methods have generated great changes revolutionized the entire process of breast cancer detection and prediction [3]. Different authors have made great efforts at the research level for the implementation of innovative methods based on artificial intelligence, machine learning, data mining, big data among others, where it is necessary to highlight the innovations made in biomedical technologies, Software and Hardware which allows the collection of information for the construction of large volumes of data as mentioned in [4], which allows the design of methods based on computational intelligence for the prediction or efficient detection of the disease as can be seen in the contributions made by [3, 5–10]. The large volumes of data play a fundamental role for the design of tools that allow the detection of breast cancer, according to the literature consulted, a wide variety of dataset oriented to the analysis of the disease has been generated, among which we can highlight the sets of data used in different experiments such as: SEER [4], Breast Cancer Wisconsin [8], Breast Cancer [11]. Data mining, understood as the discipline responsible for analyzing large volumes of data, is used as an alternative to support decision-making processes for the early and successful detection of breast cancer. In the proposed research, the implementation of classification methods, which are integrated with segmentation methods to detect the recurrence of breast cancer in patients whose information is collected in the Breast cancer Wisconsin dataset, is taken as a fundamental axis. The structure of the proposed work is presented below: Sect. 2 Previous Works, where you can find the relevant literary sources associated with breast cancer screening, Sect. 3. Materials and methods, in this section you will be able to appreciate the description of the data set used as well as the conceptual framework of the data mining methods used, Sect. 4 methodology, the process followed for the design and implementation of the proposed model is presented, Sect. 5 Results, contains the findings result of the exploratory process of the data, Sect. 7 conclusion, highlights the results achieved by the proposed model for the detection of breast cancer recurrence.

2 Brief Review of Literature

Different authors seeking to propose solutions for the early and efficient detection of breast cancer have made great contributions through the use of data mining, machine learning, artificial intelligence and bigdata methods. According to the literary analysis different studies are associated with the analysis of breast cancer using different Dataset, SEER is used in [4] where a study is presented for the prediction of breast cancer survival, a study where mining methods are compared of data DT decision trees, artificial neural networks and the statistical method of logistic regression, the best results achieved are associated with the DT method reaching a level of 93.6% accuracy, then the second best result is obtained by RNA with 91.2% and finally Logistic Regression with 89.2%. In [5] the authors present an analysis of the prediction of the survival rate in patients using data mining methods, the methods used for the experimentation process were Naive Bayes NB, Back-Propagated RNA and DT the best results were achieved by the DT method reaching an accuracy level of 86.7%, while Back-Propagated RNA achieved an 86.5% accuracy level and finally the NB method obtained an 84.5%. In [6] an analysis of breast cancer using statistical and data mining methods is presented, according to the authors there are three methods for the diagnosis of said disease which correspond to mammography, FNA (fine needle aspirate) and biopsy, which sometimes are usually expensive and unpleasant, the method that presents the best results is the biopsy with an accuracy level of approximately 100%, however it is possible to obtain better results easily through the implementation of integrated FNA with Data mining methods such as attribute selection, DT, AR association rules and statistical methods such as Principal component PCA analysis, PLS linear regression analysis. In [7] it is presented to the implementation of data mining for the discovery of breast cancer patterns based on the use of RNA and multivariable adaptive regression splines, according to the authors the RNAs have been very popular for prediction tasks and classification, the basis of the analysis is, first, to use MARS to model the classification problem, then the significant variables obtained are used as input variables of the designed neural network model. To demonstrate that the inclusion of important variables obtained from MARS would improve the accuracy of the classification of networks, diagnostic tasks are performed in a breast cancer data set with fine needle aspiration cytology.

In [8] a system for the automatic diagnosis of breast cancer based on the AR Association Rules method as attribute reduction technique and Neural Network NN as classification technique is presented, the data set used corresponds to Wisconsin Breast Cancer During the training and validation process, the 3-fold cross-validation method was used, the findings resulting from the experimentation carried out indicate that the AR + NN method achieves a correct classification rate of 95.6%. In [12] the same data set is used, where the authors propose a model for the prediction of benign and malignant ash cancer through the implementation of the Naive Bayes NB, RBF Network and J48 algorithms, the results obtained indicate that NB is the best predictor with 97.3 accuracy while RBF Network obtained 96.77% and finally the j48 algorithm achieves an accuracy level of 93.41%, during the experimentation process cross validation with 10 folds was used. In [13] an application of ML machine learning algorithms using the breast cancer Wisconsin data set for breast cancer detection is presented, for which 6 ML algorithms were submitted for comparison which correspond to GRU-SVM, Linear Regression, Multilayer Perceptron MLP, NN, Softmax Regression and Support Vector Machine SVM, in the proposed experimentation the data set is divided into 70% for the training phase and 30% for the test phase, the algorithms that obtained the best results It corresponds to MLP who achieved an accuracy level of 99.04%.

In [14] a novel approach to the detection of breast cancer using data mining techniques is presented, the objective of the proposed study is to compare three classification techniques using the Weka tool where the algorithms of SMO, IBK and BF Tree are used, the data set used corresponds to Breast Cancer Wisconsin, the results obtained show that

SMO achieves the best 96.2% accuracy results. A comparative study between the methods of K-means and fuzzy C-Means FCM for the detection of breast cancer is presented in [15], said study is focused first on comparing the performance of K-clustering algorithms means and FCM and, secondly, the integration of different computational measures is considered that allow to improve the grouping accuracy of the aforementioned techniques, FCM obtains better results compared to K-means considering that it achieves a 97% level of accuracy compared to the other technique that achieves 92%. A study for the prediction of breast cancer recurrence using data mining techniques is presented in [16], the study proposed proposes the use of different classification algorithms such as C5.0, KNN, Naive Bayes, SVM and as K-Means, EM, PAM, Fuzzy C-means clusterirng method, the experimentation performed evidence that the best results are achieved by C5.0 with an accuracy level of 81.03%.

3 Materials and Methods

3.1 Dataset Description

The set of breast cancer is provided in [17], according to the literature consulted, it is important to highlight the use of various data sets which have been made with the aim of generating significant contributions in the successful and early detection of the disease of breast cancer based on the application of artificial intelligence methods, machine learning, data mining, for the realization of classification, prediction, segmentation and association processes taking into account the heterogeneity of the variables related to the pathology in question, as You can see in [4, 18–20] where the authors present the studies of different data set alternatives for the analysis of the disease under study. For the investigation carried out, the data set called Breast Cancer taken from [17] is used, which contains 286 records of people who after having undergone surgeries present or not recurrences in the occurrence of breast cancer disease, the set of data has 201 records of people who do not show recurrence in the disease and 85 records in which there is evidence of recurrence as can be seen in Fig. 1.

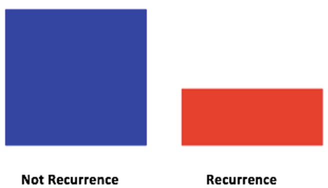

Fig. 1. Distribution of patients classified as recurring and non-recurring.

The data set has 9 attributes plus 1 class variable which limbs are described in Table 1 below.

Table 1. Description of data set attributes

Attribute	Value
Age	Ranges between: 10–19, 20–29, 30–39, 40–49, 50–59, 60–69, 70–79, 80–89, 90–99
Menopause (Time when menopause occurs)	Lt40, Ge40, Premeno
Size Tumor (tumor size expressed in mm)	Ranges between: 0–4, 5–9, 10–14, 15–19, 20–24, 25–29, 30–34, 34–39, 40–44, 45–49, 50–54, 55–59
PCCNL (Presence of cancer cells in lymph nodes)	Ranges between: 0–2, 3–5, 6–8, 9–11, 12–14, 15–17, 18–20, 21–23, 24–26, 27–29, 30–32, 33–35, 36–39
CCACNL (Cancer cells passed through lymph node capsule)	Yes, no
GHT (Histological Grade of the tumor)	1, 2, 3
Breast (Affected Breast)	Left, Right
Breast (Quadrant of the affected breast)	Left – Top, Left – Bottom, Right – Superior, Right – Bottom, Center
RT (Radio therapy)	Yes, No
Recurrencia (Variable de Clase)	Recurrence, Not recurrence

In the investigation carried out, the use of different classification and segmentation methods is proposed through the implementation of the DT, SVM, Naive Bayes, Simple K-Means algorithms, which is why in Sect. 4 the procedure and the changes made are described about the data set for its optimal behavior with the techniques mentioned.

3.2 Decision Trees

Decision tree is responsible for the recursive partition of a data set for which subdivisions are generated, generally the tree structure is composed of a main node called root and a node set that grow from the root called child nodes, Finally, within the tree structure there are the terminal nodes, it should be noted that each node of the tree has only one parent node and two or more descendant nodes [21]. According to [22] some DT algorithms are: classification and regression tree (CART), iterative Dichotomiser 3 (ID3), C4.5 and C5.0, Automatic detection of Chi-square interaction (CHAID), Decision stump, M5, Conditional decision trees. The following is a diagram of the structure of a decision tree.

According to [23] a decision tree supports the decision-making process. In [24] they mention that decision trees represent a supervised approach to classification that has a simple structure composed of terminal nodes or nodes, the nodes represent tests on one or more attributes and the terminal nodes present the results of the decisions.

3.3 Vector Support Machines

It is a method of classification that was proposed by Vapnik [25], It is based on the use of a hyperplane separator or a decision. The plane is responsible for defining the limits of decision between a set of data points classified with different labels according to what is mentioned in [22], they express the simple geometric explanation of this approach is to determine an optimal separation plane or hyperplane that separates the two classes or groups of data points fairly and is equidistant from both of them. SVM was first defined for the linear distribution of data points, additionally through the use of the kernel function it is possible to implement it to address problems with non-linear data. SVM has been applied by different agents in tasks such as recognition of Vapnik manuscript digits [25], used in object recognition problems [26], text classification [27]. According to [28], one of the advantages of SVM is considered to be the availability of powerful tools and algorithms to find the solution quickly and efficiently. SVM vector support machines have a strong theoretical foundation and excellent empirically results [24].

3.4 Naive Bayes

It is a supervised classification method developed using the Conditional Probability Theorem, they perform well in different situations, such as text classification and spam detection. Only a small amount of training data is necessary to estimate certain parameters [22]. Bayesian networks are considered an alternative to classic expert systems oriented to decision making and prediction under uncertainty in probabilistic terms [29]. The NB algorithm is a probabilistic classifier that calculates a set of probabilities based on the frequency and combination of the values given on the dataset [30], the algorithm uses the Bayes theorem and assumes that all data are independent of the values of class variable [31], rarely the assumption of conditional independence is met in real world applications, this is a naïve characterization, but the algorithm tends to work properly and learn fast in several supervised classification problems [32]. The most common algorithms that implement this method are: Naive Bayes, Gaussian Naive Bayes, Multinomial Naive Bayes, Averaged One-Dependence Estimators (AODE), Bayesian Belief Network (BBN), Bayesian Network (BN) [22].

3.5 Cluster

The cluster method is considered the most popular unsupervised learning task, which is based on the construction of a set of physical or abstract objects that have matching characteristics or attributes [33]. According to [34] the product of a good grouping, high quality groups with high intra-class similarities and low class similarities should be generated. The implementation of clustering algorithms for unsupervised data analysis has become a useful tool to explore and solve different problems in data mining. According to [22] Some of the algorithms implemented for cluster methods are: K-Means, K-Medians, Affinity Propagation, Spectral Clustering, Ward hierarchical clustering, Agglomerative clustering. DBSCAN, Gaussian Mixtures, Birch, Mean Shift, Expectation Maximization (EM).

4 Methodology

For the development of the proposed research, we initially start with obtaining the data set called Breast Cancer Wisconsin taken from [17], within which it was necessary to perform the data preprocessing stage called phase number 1 where we highlight the realization of a analysis of the balancing of the data, later in phase number 2 the training process and test of the classification methods used where the DT, NB and SVM algorithms were compared through the metrics of precision, coverage evaluation, true positive rate, false positive rate, finally in phase 3 the best classification method is taken with the simple cluster method k-means and the result obtained is compared in relation to the results obtained only by the classification. The experimentation process was performed using the WEKA data mining tool. The aspects mentioned above are detailed below:

4.1 Phase No 1. Preprocessing

Within the problems found in the data set we can see that the data is not balanced with respect to the class variable called recurrence which contains two possible recurring and non-recurring values, the foregoing considering that the non-recurring variable contains 201 records representing 70.27% of the total data, on the other hand, the recurring variable only has 85 data which corresponds to 29.7% of the records, the above can be evidenced in Fig. 1. the Previous figures mentioned are a problem for the method since it would learn to correctly identify the non-recurrence variable, which is why balancing the data is proposed. Next, in Fig. 2 the configuration used in the SMOTE filter for the data balancing process is presented.

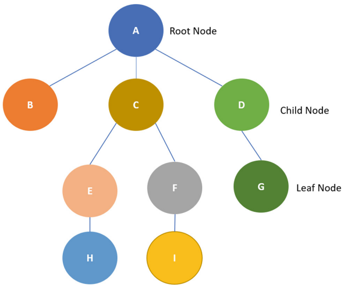

Fig. 2. DT structure representation.

Once the SMOTE filter has been implemented for the balancing process of the data, the recurrence class variable has a homogeneous distribution in terms of the number of records, so in the non-recurrence value there are 201 records (50.7%) and in recurrence 195 records (49.3%), in Fig. 3 the number of records per class is presented after the data balancing process. The data were subjected to verification where it was evidenced that it was not necessary to replace atypical data and missing data (Fig. 4).

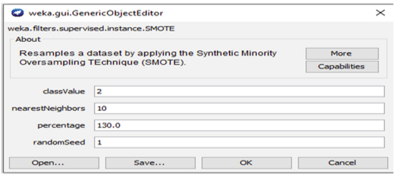

Fig. 3. SMOTE filter configuration for data balancing

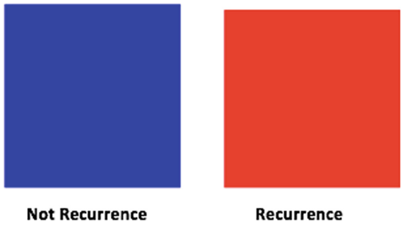

Fig. 4. Balanced data distribution with SMOTE filter

4.2 Phase No 2. Training and Testing of Classification Algorithms

During the training and testing process of the selected methods, the cross-validation method was used as a test option where a part of the data set is randomly taken for the training process and the other part for the test this procedure is repeated according to the number of stipulated folds which in this case correspond to 10, then in Fig. 5. A scheme of the cross-validation process is presented.

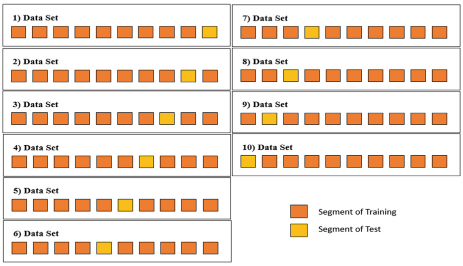

Fig. 5. Cross validation process scheme

Additionally, during this phase different evaluation metrics were taken into account such as level of accuracy, coverage, true positive rate, false positive rate.

4.3 Phase No 3. Integration of Classification and Segmentation Techniques

During this phase, the Simple K-means cluster method was integrated with the different classification algorithms, with the purpose of validating whether, by integrating both methods, better results are obtained in the evaluation metrics used during the experimentation process. In session V results, a description of the centroids corresponding to the clusters will be made.

5 Results

The methods used for the classification process correspond to DT which was implemented through the J48 and RandomForest algorithms, NB implemented the NaiveBayes and NaiveBayesSimple algorithms and finally the SVM method which is implemented using the SMO algorithms with Polikernel and SMO with RBFKernel. The results obtained are presented in Table 2 below.

Table 2. Results classification methods

Method	Algorithm	TPRATE	FPRATE	PRECISION	RECALL
DT	J48	70,7%	29,2%	72,0%	69,2%
	RandomForest	73,7%	26,2%	73,8%	73,7%
NB	NaiveBayes	69,7%	30,4%	69,7%	69,7%
	NaiveBayesSimple	69,7%	30,4%	69,7%	69,7%
SVM	SMO - PoliKernel	70,7%	29,3%	70,7%	70,7%
	SMO - RBF Kernel	69,9%	30,3%	70,3%	69,9%

In Table 2, you can see the different results obtained by the classification methods where we must highlight the result achieved by the RandomForest algorithm who obtained a level of Accuracy of 73.8%, Recall 73.7%, TPRATE 73.7% and FPRATE 26.2%, however, the results obtained by the J48 and SMO algorithms with polikernel are promising given that their level of accuracy is approximate to the level achieved by RandomForest. After the results obtained, it can be mentioned that these are acceptable, additionally the classifiers are integrated with the cluster method in search of improving the levels in the evaluation metrics used. Next, in Fig. 6 the distribution of the data with respect to the generated clusters which are described in Table 3 can be seen.

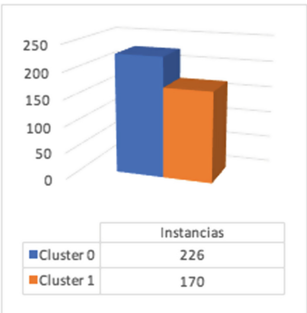

Fig. 6. Clustered data distribution

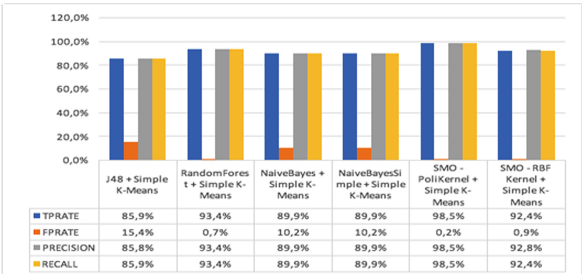

Fig. 7. Classification methods results.

The result of the classification methods used during the experimentation with the cluster algorithm which corresponds to Simple K-means is presented in Fig. 7.

Table 3. Cluster description

Cluster	Description
Cluster 0	Represents those whose age range is approximately between 40 and 49 years, the time at which menopause occurs is generally at 40, the size of the tumor expressed in mm corresponds to 30 and 34, whose presence of cancer cells in lymph nodes is 0 and 2.3, do not have cancer cells that cross the lymph node capsule, with a histological grade of the tumor 3 (high level), which has the affected left breast in the lower left quadrant, was not subjected to radius therapy and is recurrent
Clúster 1	It represents those whose age range corresponds to 30 to 39 years, who are in premenopausal condition, whose tumor size expressed in mm corresponds to 25 to 29, with presence between 0 and 2.3 cancer cells in lymph nodes, they do not have cancer cells that have passed through lymph node capsules, the affected breast corresponds to the right in the lower left quadrant, they have not undergone radiotherapy and are non-recurring patients

6 Discussion

Breast cancer is a problem that affects many women worldwide, which has generated great interest in the scientific field in order to take advantage of technological advances and thus generate intelligent tools or methods that allow diagnosis or prevention of Successful form of said disease. Data mining is a discipline in constant development, through this a large number of solutions have been generated for the analysis of different diseases that society suffers worldwide. In the proposed work, a model based on data mining techniques is proposed for the detection of recurrent or non-recurring persons after having undergone surgery in a period prior to 5 years.

According to the results, an analysis was initially carried out through classification methods such as DT, NB and SVM where different algorithms were used for the implementation of the aforementioned techniques. Taking into account the results, the levels of precision achieved are not the best since randomforest only reached a 73.8% level of accuracy, followed by J48 and SMO with 72% and 70.7% respectively, consequently, it was determined the alternative of integrating the classification methods with the simple cluster method k-means with the purpose of obtaining improvements with respect to the previously obtained results.

After the integration of the classification methods with the cluster algorithm, a significant improvement is observed in the levels of precision achieved, taking into account that, for example, the randomforest method went from having an accuracy of 73.8% to 93.4%, on the other hand, all the algorithms increased their values in the precision assessment metric, which indicates that an alternative for the successful identification of breast cancer is the joint use of classification methods with cluster methods. In previous works such as [8, 12–14] the study of breast cancer is proposed using classification methods which show good results, however it is possible to explore the possibility of integrating the proposed methods with cluster methods. This cluster method approach with classification method is also explored prior research for the determination of other pathologies that affect different societies [35, 36].

7 Conclusions

Cancer is a disease in constant evolution, which affects a large number of people worldwide. According to the world health organization 8.2 million people died from this disease where you can see the usual occurrence of cancer such as lung, liver, colon and breast stomach [2], large volumes of data play a role Fundamental for the design of tools that allow the detection of breast cancer, according to the literature consulted, a wide variety of dataset oriented to the analysis of the disease has been generated. In this investigation, the Breast Cancer data set taken from [17] was used, which was preprocessed for the validation of the data quality where it was necessary to perform data balancing, the algorithms implemented for the classification process correspond to J48 and randomforest (DT), NaiveBayes and NaiveBayes Simple (NB), SMO Poly-kernel and SMO RBF-Kernel (SVM), the Simple K-Means algorithm was used as a cluster method which was integrated with the classification methods In order to obtain the best results, finally the methods that obtained the best levels were SMO Poly-Kernel + Simple K-Means 98.5% Precision, 98.5% recall, 98.5% TPRATE and 0.2% deFPRATE. The results obtained suggest the possibility of using intelligent computational tools based on data mining methods for the detection of breast cancer recurrence in patients who had previously undergone surgery.

References

1. Facts and figures of cancer. https://www.who.int/cancer/about/facts/es/
2. Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., Fotiadis, D.I.: Machine learning applications in cancer prognosis and prediction. Comput. Struct. Biotechnol. J. **13**, 8–17 (2015)
3. Gupta, S., Kumar, D., Sharma, A.: Data mining classification techniques applied for breast cancer diagnosis and prognosis. Ind. J. Comput. Sci. Eng. (IJCSE) **2**(2), 188–195 (2011)
4. Delen, D., Walker, G., Kadam, A.: Predicting breast cancer survivability: a comparison of three data mining methods. Artif. Intell. Med. **34**(2), 113–127 (2005)
5. Bellaachia, A., Guven, E.: Predicting breast cancer survivability using data mining techniques. Age **58**(13), 10–110 (2006)
6. Xiong, X., Kim, Y., Baek, Y., Rhee, D.W., Kim, S.H.: Analysis of breast cancer using data mining & statistical techniques. In: Sixth International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing and First ACIS International Workshop on Self-Assembling Wireless Network, pp. 82–87. IEEE (2005)
7. Chou, S.M., Lee, T.S., Shao, Y.E., Chen, I.F.: Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines. Expert Syst. Appl. **27**(1), 133–142 (2004)
8. Karabatak, M., Ince, M.C.: An expert system for detection of breast cancer based on association rules and neural network. Expert Syst. Appl. **36**(2), 3465–3469 (2009)
9. Hung, P.D., Hanh, T.D., Diep, V.T.: Breast cancer prediction using spark MLlib and ML packages. In: Proceedings of the 2018 5th International Conference on Bioinformatics Research and Applications, pp. 52–59. ACM (2018)
10. Shadman, T.M., Akash, F.S., Ahmed, M.: Machine learning as an indicator for breast cancer prediction, Doctoral dissertation, BRAC University (2018)
11. Alwidian, J., Hammo, B.H., Obeid, N.: WCBA: weighted classification based on association rules algorithm for breast cancer disease. Appl. Soft Comput. **62**, 536–549 (2018)

12. Chaurasia, V., Pal, S., Tiwari, B.B.: Prediction of benign and malignant breast cancer using data mining techniques. J. Algorithms Comput. Technol. **12**(2), 119–126 (2018)
13. Agarap, A.F.M.: On breast cancer detection: an application of machine learning algorithms on the wisconsin diagnostic dataset. In: Proceedings of the 2nd International Conference on Machine Learning and Soft Computing, pp. 5–9. ACM (2018)
14. Chaurasia, V., Pal, S.: A novel approach for breast cancer detection using data mining techniques (2017)
15. Dubey, A.K., Gupta, U., Jain, S.: Comparative study of K-means and fuzzy C-means algorithms on the breast cancer data. Int. J. Adv. Sci. Eng. Inf. Technol. **8**(1), 18–29 (2018)
16. Ojha, U., Goel, S.: A study on prediction of breast cancer recurrence using data mining techniques. In: 2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence, pp. 527–530. IEEE (2017)
17. Lichman, M.: UCI machine learning repository, University of California, School of Information and Computer Science, Irvine, CA (2019). http://archive.ics.uci.edu/ml/datasets/breast+cancer
18. Abbass, H.A.: An evolutionary artificial neural networks approach for breast cancer diagnosis. Artif. Intell. Med. **25**(3), 265–281 (2002)
19. Akay, M.F.: Support vector machines combined with feature selection for breast cancer diagnosis. Expert Syst. Appl. **36**(2), 3240–3247 (2009)
20. Polat, K., Güneş, S.: Breast cancer diagnosis using least square support vector machine. Digit. Sig. Proc. **17**(4), 694–701 (2007)
21. Friedl, M.A., Brodley, C.E.: Decision tree classification of land cover from remotely sensed data. Remote Sens. Environ. **61**(3), 399–409 (1997)
22. Das, K., Behera, R.N.: A survey on machine learning: concept, algorithms and applications. Int. J. Innov. Res. Comput. Commun. Eng. **5**(2), 1301–1309 (2017)
23. Magerman, D.M.: Statistical decision-tree models for parsing. In: Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics, pp. 276–283. Association for Computational Linguistics (1995)
24. Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. J. Mach. Learn. Res. **2**(Nov), 45–66 (2001)
25. Vapnik, V.: Statistical Learning Theory. Wiley, Hoboken (1998)
26. Papageorgiou, C., Oren, M., Poggio, T.: A general framework for object detection. In: Proceedings of the International Conference on Computer Vision (1998)
27. Joachims, T.: Text categorization with support vector machines: learning with many relevant features. In: Nédellec, C., Rouveirol, C. (eds.) ECML 1998. LNCS, vol. 1398, pp. 137–142. Springer, Heidelberg (1998). https://doi.org/10.1007/BFb0026683
28. Bekele, E., et al.: Multimodal adaptive social interaction in virtual environment (MASI-VR) for children with Autism spectrum disorders (ASD). In: 2016 IEEE virtual reality (VR), pp 121–130 (2016). https://doi.org/10.1109/vr.2016.7504695
29. Picard, R.W., et al.: Affective learning—a manifesto. BT Technol. J. **22**(4), 253–269 (2004). https://doi.org/10.1023/B:BTTJ.0000047603.37042.33
30. Patil, T.R., Sherekar, S.S.: Performance analysis of Naive Bayes and J48 classification algorithm for data classification. Int. J. Comput. Sci. Appl. **6**(2), 256–261 (2013)
31. O'Reilly, K.M.A., Mclaughlin, A.M., Beckett, W.S., Sime, P.J.: Asbestos-related lung disease. Am. Fam. Phys. **75**(5), 683–688 (2007)
32. Peddabachigari, S., Abraham, A., Grosan, G., Thomas, J.: Modeling intrusion detection system using hybrid intelligent systems. J. Netw. Comput. Appl. **30**(1), 114–132 (2007)
33. Han, J., Kamber, M.: Data mining: concepts and techniques, 2nd edn. Morgan Kaufmann Publishers, San Francisco (2001)

34. Hastie, T., Tibshirani, R., Friedman, J.H.: The Elements of Statistical Learning: Data Mining, Inference and Prediction. Springer, New York (2001). https://doi.org/10.1007/978-0-387-84858-7
35. Palechor, F.M., De la Hoz Manotas, A., Colpas, P.A., Ojeda, J.S., Ortega, R.M., Melo, M.P.: Cardiovascular disease analysis using supervised and unsupervised data mining techniques. JSW **12**(2), 81–90 (2017)
36. Mendoza-Palechor, F.E., Ariza-Colpas, P.P., Sepulveda-Ojeda, J.A., De-la-Hoz-Manotas, A., Piñeres Melo, M.: Fertility analysis method based on supervised and unsupervised data mining techniques (2016)