IOSR Journal of Biotechnology and Biochemistry (IOSR-JBB)
ISSN: 2455-264X, Volume 1, Issue 5 (Jul. – Aug. 2015), PP 01-05
www.iosrjournals.org

Knowledge Discovery from Breast Cancer Database

Ms. N.Vijayalakshmi[1], Ms. G.Divya lakshmi[2]

1. *Asst. Professor, Dept. of M.C.A., Shrimati Indira Gandhi College, Trichy - 2.*
2. *Research Scholar in Computer Science, Shrimati Indira Gandhi College, Trichy-2*

Abstract: *In this paper, we study various factors leading to breast cancer and also a few symptoms that act as biomarkers for the occurrence of breast cancer in women. Totally 18 factors are taken for study. Statistical techniques are used to analyze the influence of various factors towards the disease and test for significance of factors is also done. Besides association rule mining is attempted to generate possible factors that may lead to breast cancer. An attempt to classify the given dataset using information gain techniques and CHAID techniques was done. Clustering was also done to predict the occurrence of breast cancer. The results show that there is more possibility of developing breast cancer among married working women who have breast fed less than 2.5 years in total.*

Keywords: *Breast cancer, Data mining, decision tree, classification, clustering*

I. Introduction

Breast cancer is a chronic and fatal disease that occurs mostly in women. According to the American Cancer Society(Cancer Facts & Figures 2015), it is estimated that more than 1.6 million new cases of cancer will be diagnosed in 2015. Among women, breast (29%), lung (13%), and colon & rectum (8%) cancers are the most common cancers. The lifetime probability of a woman developing cancer has been found to be 1 in 8. 15% of deaths due to cancer in the U.S are attributed to breast cancer. However it has been observed that Asians are less prone to cancer in general.

Even though the occurrence of breast cancer in women is lower in India compared to other countries, the mortality rate is very low. One out every two women diagnosed with breast cancer dies. Hence the government has to take aggressive steps to prevent this trend. It is also very much required to bring awareness to people about possible factors that may greatly lead to breast cancer. In this paper an attempt has been made to identify the highly determining factors leading to breast cancer using statistical and data mining techniques. Hidden medical information from data of normal and breast cancer patients is a useful approach to find the possible influence of each factor on the disease. Data mining techniques also play a vital role in determining the major factors that cause the disease.

II. Literature Review

Study of various research activities related to breast cancer reveal that there have been very few attempts to use data mining techniques to predict the occurrence of breast cancer based on social, economic, demographic, and biological factors relating to the same. However a few studies relating to breast cancer by Dursen et al.,[1] reveal that decision tree (C5) is the best predictor with 93.6% accuracy on the holdout sample, ANN with 91.2% accuracy and logistic regression models with 89.2% accuracy. However another study by Pendharkar et al.,[2], based on data obtained from a large medical facility in western Pennysylvania shows that data mining can be a viable tool for breast cancer diagnosis. In another study by Bernard Rosner et al.[3], the authors found that ages at first and subsequent births have a long-term influence on breast cancer incidence and that child bearing women are more prone to breast cancer.

Paffenbarger et al [4] have found that common factors to increased risk of this disease in both periods of womanhood(before and after menopause) were: early menarche and late menopause; delayed marriage and first childbirth; more nulliparity or reduced gravidity and parity; reduced frequency of abortions; shorter overall child-bearing interval; more advanced education, higher socioeconomic status, and more contraceptive usage; and familial tendencies toward the disease. Bruce et al., [5] have found that prolonged OC use may accelerate the onset of breast cancer for a small group of susceptible women.

III. Methodology

3.1 Data Sources

A sample population consisting of 202 women from various areas in and around Trichy was taken. Socio-economic factors and certain biological factors were considered. A questionnaire was used to elicit the required data from the sample population. Data on various factors were collected from G.V.N Cancer Care Hospital, Trichy from patients undergoing treatment for breast cancer in 2015.Out of 202, 116 women were diagnosed with breast cancer. Out of these, 86 were at the benign stage and 30 were at advanced stages of cancer.

The following factors were considered for study

1.	Present Age of patient	
2.	Age at Menarche	
3.	Marital Status	
4.	Age at Menopause	
5.	Currently pregnant	
6.	No. of Children	
7.	Total Period of Lactation	
8.	Whether working during lactation	
9.	Occurrence of pain in breast	
10.	Is Breast pain related to menstruation	
11.	Discharge from nipple	
12.	Swelling in breast	
13.	Swelling in armpit	
14.	Family history of breast cancer	
15.	Itching/dryness of breast	
16.	Oral contraceptive pills taken continuously	
17.	Diagnosis of breast cancer	
18.	Stage of breast cancer if applicable	

The collected patient data was processed using statistical, and data mining techniques to discover certain facts that may help in predicting the occurrence of breast cancer in patients and also in preventing the same

3.2 Statistical Analysis

Use of statistical analysis on the sample data revealed the following facts:

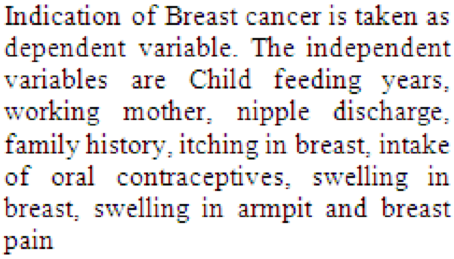
1. 93.5% of our sample population have their age between 30 and 60. Menarche age ranges between 13 and 19, 13-16 being 50% and 17 to 19 being 50%
2. 200 women were married, 6 did not have any children, while out of the others 146 women had 1 or 2 children and the rest had more than 2 children. We can also see that out of those patients with cancer at stages 2 or 3, 26 have less than 3 children. This implies that the incidence of cancer decreases with increase in the no. of child births.
3. Out of 202 patients, 87 did not have breast cancer, 115 had breast cancer. Out of those who have breast cancer, 24 ie 21% have pain in any one breast only, while 11% have pain in both breasts. 68% do not have any pain in their breasts. However out of those who had cancer at stage 2 and stage 3 ie 30 patients, 15 that is nearly 50% had pain in their breasts. However the chi-square statistic shows that breast pain is a significant factor leading to breast cancer.
4. We can see that 93% of patients suffering from breast cancer do not have any kind of discharge from their nipples. So this cannot be taken as a significant factor for having breast cancer.
5. We can see that 87% of patients suffering from breast cancer do not have any kind of swelling in their breasts. So it appears that this cannot be taken as a significant factor for having breast cancer. However 8 out of 28 patients at stage 2 cancer have swelling in their breasts, which is significant.
6. We can see that 89% of patients suffering from breast cancer do not have any kind of swelling in their armpits. So this cannot be taken as a significant factor for having breast cancer. 50% of those who have swelling in their armpits are diagnosed at stage1, 35% at stage 2 and 15% at stage 3.
7. We can see that 98% of patients suffering from breast cancer do not have any family history of breast cancer. Similarly one person who has a family history of cancer is not currently a cancer patient. So this cannot be taken as a significant factor for having breast cancer.
8. Itching or dryness in the breast cannot be taken as a factor leading to breast cancer. However it is also seen that all patients who have itching are married and have fed children.
9. Even though regular use of oral contraceptives has been attributed to cause breast cancer, only 3.4% of breast cancer patients have regularly taken contraceptives, as per our sample data.
10. Only 2 unmarried women out of 117 have cancer. This tells us that married women are more prone to get breast cancer. But among all married woman(200), 43% do not have cancer while 56% have cancer.
11. It is also observed that out of 115 cancer patients who were scanned to find that there was a lump or cist in their breasts, 30 were found to have stage 2 or stage 3 cancer and the rest ie., 85 out of 115 were at stage 1, ie., they had a possibility of developing cancer later. Therefore marital status is one of the significant factors influencing occurrence of breast cancer.

12. It is found that 68% of people who have not reached menopause are prone to get breast cancer. It is also observed that those who still have their periods are more prone to cancer at stage 2 and 3 than those who have reached menopause
13. It can also be seen that we have 127 working mothers, out of which 71 have been diagnosed with cancer(56%). Among stage 2 and stage 3 patients, there is an equal amount of both working and non-working women having breast cancer. But comparing the number of working and non-working women who do not have cancer or who are prone to develop cancer later, we see that the ratio is 60:110 that is 35% : 65%. Therefore working women should be more careful.

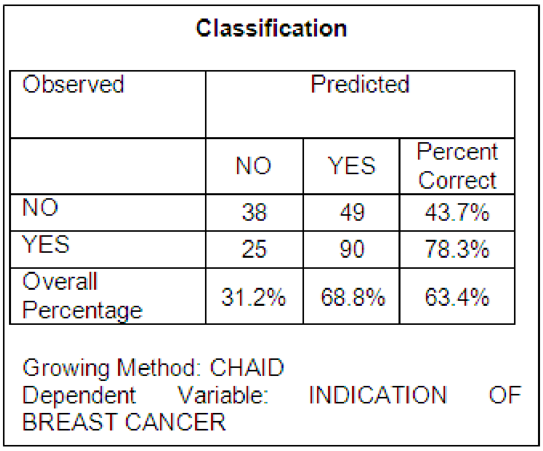
3.3 Data mining

3.3.1 Use of Chaid for classification:

Chaid or Chi-Square Automatic Interaction Detector is an exploratory method or more precisely an algorithm to study the relationship between a dependent variable and a series of predictor variables. This algorithm selects a set of predictors and their interactions and predicts the optimal value of the dependent variable. In the end what we get is a classification tree. Use of decision trees for classification of the given dataset using CHAID method produced the following result:



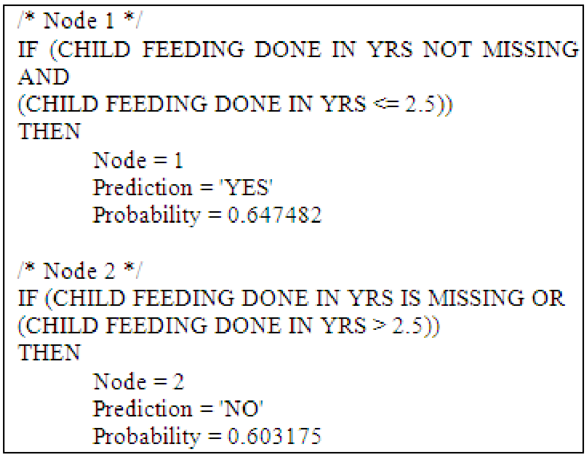
Indication of Breast cancer is taken as dependent variable. The independent variables are Child feeding years, working mother, nipple discharge, family history, itching in breast, intake of oral contraceptives, swelling in breast, swelling in armpit and breast pain



Classification

Observed	Predicted		
	NO	YES	Percent Correct
NO	38	49	43.7%
YES	25	90	78.3%
Overall Percentage	31.2%	68.8%	63.4%

Growing Method: CHAID
Dependent Variable: INDICATION OF BREAST CANCER



```
/* Node 1 */
IF (CHILD FEEDING DONE IN YRS NOT MISSING
AND
(CHILD FEEDING DONE IN YRS <= 2.5))
THEN
        Node = 1
        Prediction = 'YES'
        Probability = 0.647482

/* Node 2 */
IF (CHILD FEEDING DONE IN YRS IS MISSING OR
(CHILD FEEDING DONE IN YRS > 2.5))
THEN
        Node = 2
        Prediction = 'NO'
        Probability = 0.603175
```

Figure. 1 Decision rule generated using CHAID method

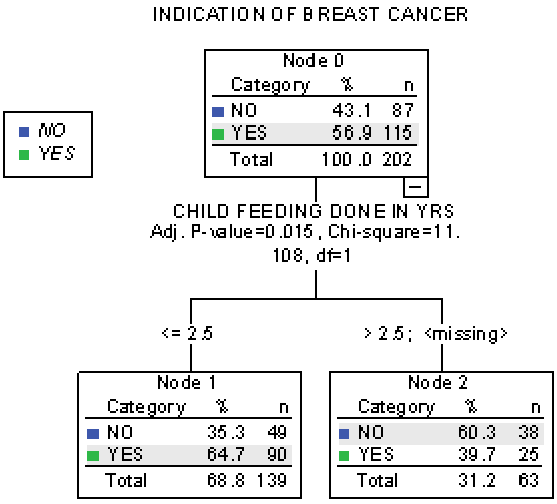


Figure. 2 Decision tree generated using CHAID method

3.3.2 Use of k-means clustering based on years of child feeding yielded the following results:

Table 1. k-means clustering based on years of child feeding

Initial Cluster Centers

	Cluster	
	1	2
CHILD FEEDING DONE IN YRS	.0	5.0

Iteration History(a)

Iteration	Change in Cluster Centers	
	1	2
1	1.216	1.491
2	.009	.018
3	.000	.000

Number of Cases in each Cluster

Cluster	1	138.000
	2	56.000
Valid		194.000
Missing		8.000

Final Cluster Centers

	Cluster	
	1	2
CHILD FEEDING DONE IN YRS	1.2	3.5

Thus we can see that most women have either fed children for around 1.2 years or around 3.5 years.

3.3.3 Use of Information gain for classification

Next we try to generate a decision tree using another method. The type of decision tree used is the gain ratio decision tree. The gain ratio decision tree is based on the entropy (information gain) approach, which selects the splitting attribute that minimizes the value of entropy, thus maximizing the information gain [6]. Information gain (Leung et al., 2011) is the difference between the original information content and the amount of information needed. The features are ranked by the information gains, and then the top-ranked features are chosen as the potential attributes used in the classifier. To identify the splitting attribute of the decision tree, one must calculate the information gain for each attribute and then select the attribute that maximizes the information gain. The information gain for each attribute is calculated using the following formula [7]

$$E = \sum_{i=1}^{k} p_i \ \log_2 p_i$$

Where k is the number of classes of the target attribute, and p_i is the number of occurrences of class i divided by the total number of instances. Gain ratio adjusts the information gain for each attribute to allow for the breadth and uniformity of the attribute values.

Gain ratio = Information Gain / Split information

Where the split information is the value based on the column sums of the frequency table(Bramer, 2007).

Use of information gain method to generate decision tree yielded the following results

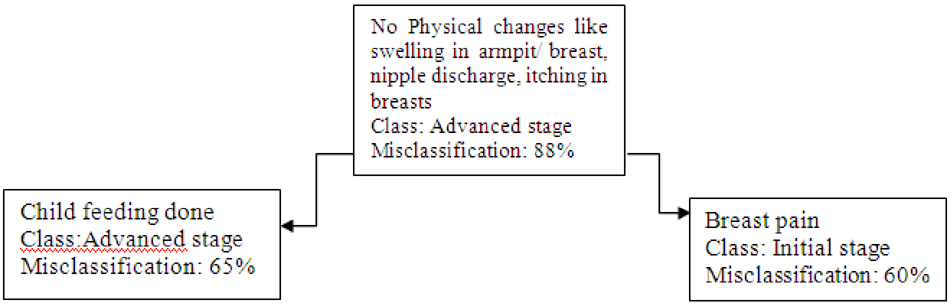


Figure 3. Decision tree based on information gain method.

IV. Results and Discussion

Using statistical analysis, it was found that working women who were married and who have fed children for less than 2.5 years were more prone to get breast cancer. The same results were also obtained through data mining techniques. Use of CHAID and information gain method for decision tree generation also emphasized that child feeding was the most significant factor in determining the occurrence of breast cancer. However, breast pain and physical changes in the body like swelling in the breast and armpit, nipple discharge, itching in the breasts were also important factors that led to breast cancer, as observed by the information gain method. The k-means clustering method generated two clusters, one with median 1.2 and the other with median 3.5, based on period of lactation.

V. Conclusion

As already suggested by researchers, decision tree method accurately predicts factors determining breast cancer. In this work also two decision tree methods have been used to predict the probability of getting breast cancer based on significant factors. We get an accuracy of 64% in the results. We have also used statistical methods to acquire knowledge about various factors. These also correlate with the results produced by data mining techniques.

References

Journal Papers

[1]. Dursen Delen, Glenn Walker, Amit Kadam, "Predicting breast cancer survivability: A comparison of three data mining methods", Artificial Intelligence in Medicine, Volume 34, June 2005.

[2]. P.C. Pendharkar, J.A.Rodger, G.J. Yaverbaum, N. Herman, M.Benner, "Association, statistical, mathematical and neural approaches for mining breast cancer patterns", Expert Systems with Applications, Volume 17, Issue 3, October 1999.

[3]. Bemard Rosner, Graham A. Colditz and Walter C. Willett , "Reproductive Risk Factors in a Prospective Study of Breast Cancer: The Nurses' Health Study", Oxford Journals Medicine & Health, American Journal of Epidemiology, Volume 139, Issue 8, 1994

[4]. Paffenbarger RS Jr, Kampert JB, Chang HG,"Characteristics that predict risk of breast cancer before nd after the menopause", American Journal of Epidemiology, 112(2):258-268,1980

[5]. Bruce V. Stadel, M.D.,M.P.H, Lai Shenghan, M.M.,Ph.D., James J. Schlesselman, Ph.D., Pamela Murray, M.S. "Oral contraceptives and premenopausal breast cancer in nulliparous women",Volume 38, Issue 3, 287-299, Sep 1988

Books

[6]. Bramer M., Principles of Data Mining (Springer, London, 2007)

[7]. Han J and Kamber M, Data Mining: Concepts and techniques, 2nd Ediction, (Morgan Kaufmann, San Fancisco, CA)