The International Journal Of Engineering And Science (IJES)
|| Volume || 5 || Issue || 11 || Pages || PP 29-33 || 2016 ||
ISSN (e): 2319 – 1813 ISSN (p): 2319 – 1805

Classification and Predication of Breast Cancer Risk Factors Using Id3

Rawaa Abdulridha Kadhim
Electrical technical engineering collage/Middle Technical University

----------ABSTRACT----------
Breast cancer considered as the first or second cause of death of women each year. There are many risk factors give an indication for the possibility of future cancer development. Therefore, a predictive system depends on risk factors will help in early detection of this disease. Here we used an ID3, by WEKA software, J48 algorithm to build a predictive system of a classifier tree to classify the dataset groups of risk factors according to its priority and gives a future likelihood to develop cancer, the results show that the family history is the most affecting factor in addition to other factors

Keywords: *Breast cancer, predictive systems, tree classifiers*

Date of Submission: 02 November 2016 Date of Accepted: 22 November 2016

I. INTRODUCTION

For the last several decades breast Cancer was one of the dominant causes of death of women around the world and the first cause in some developing countries [1]. Many risk factors increase the possibility to have breast cancer such as age, gender, no of births and age of first birth. Despite the fact that cancer is not an inherited disease, last studies show that a mutation (loss of genetic function based on a certain change in genetic code) in a particular genes called BRCA1 and BRCA2 will increase the possibility to have breast cancer. These mutations may inherited from parents, sisters (first-degree relatives), although genetic mutation not necessarily leads to have breast cancer but it is an important risk factor especially when it related to other risk factor, the family history [2]. So that this genetic factor can be used to give information about the ability of the woman to develop breast cancer in the future. Predictive system translates the available factors and signs to a clear vision about the readiness of this woman to develop cancer in the future [3].

Id3 tree classifier algorithm will use to classify a several breast cancer risk factors according to their priority to develop cancer, and to see the effect of genetic mutations of BRCA1, BRCA2, and family history. This tree classification based on data sets of probability tables constructed according to several studies that gives an accurate percentages of how the risk factors are effective[7].

II. RELATED WORK

S. Syed shajahaan [4] explored the applicability of decision trees to predict the presence of breast cancer; he also analyzed the performance of conventional supervised learning algorithms visa random tree, ID3, CART and C4, 5. He proved that random trees servs to be the best one with high accuracy.

Shweta Kharya [5] summarizes various reviews and technical articles on breast cancer diagnosis and prognosis.

Abdelghani Bellaachia [6] presented an analysis of the prediction of survivability of breast cancer patients using data mining techniques. He investigated three techniques naive Bayes, the Back propagation neural network and 4.5decision tree algorithm, they found that c4.5 algorithm has much better performance than the two other techniques

Three groups of data set used; the first data set group contains age, BRCA1, BRCA2 and family history each risk factor will give a percentage the whole percentage will be the average of all these factors.

The second data set group contains in addition to the risk factors in-group 1 the no of births for woman before age 30 that will decrease the probability percentage.

An Id3 classifier tree algorithm used to classify the risk factors and determine which factor is more effective than others are and to see the effect of genetic mutation with family history to develop breast cancer.

ID3

Is an algorithm used to classify data so that a group of data set enters to the algorithm and the output is a base classifier through which new dataset group that have not used before will be classified. this classifier is a tree structure so called decision tree and can be converted into a set of rules called the rules so also the name of decision rules[7]

A decision tree will be build based on the best choice Attribute property. This property should classify the Training set so that the depth of the tree will be less and as the classification of the data is correct at the same time [8]

It isused because it is easy to understand and implement, It simply deals with data as (if_ else) condition. this classification tree composed of nodes with single input and a root node with no input these nodes are called leaves ,and there is another type of nodes that have only output is called decision nodes , the algorithm used by this classification tree is Id3 .

The dataset table1 includes factors of age, family history and no of births Will entered tothe classifier system the result tree shown in fig (1)

WEKA is a java machine learning software; it provides choices of many algorithms and tools to analyze data and predictive systems [8].

III. MATERIALS AND METHODS

The software used for classification of risk factors is WEKA GUI chooser software, J48 algorithm. Dataset 1 group was loaded as shown in fig (1)(a and b) including many attributes (the risk factors) such as age, family history, BRCA1, BRCA2, and no of births after 30 years. This data used for training process of the classification, that the output classification tree shown in fig (2)

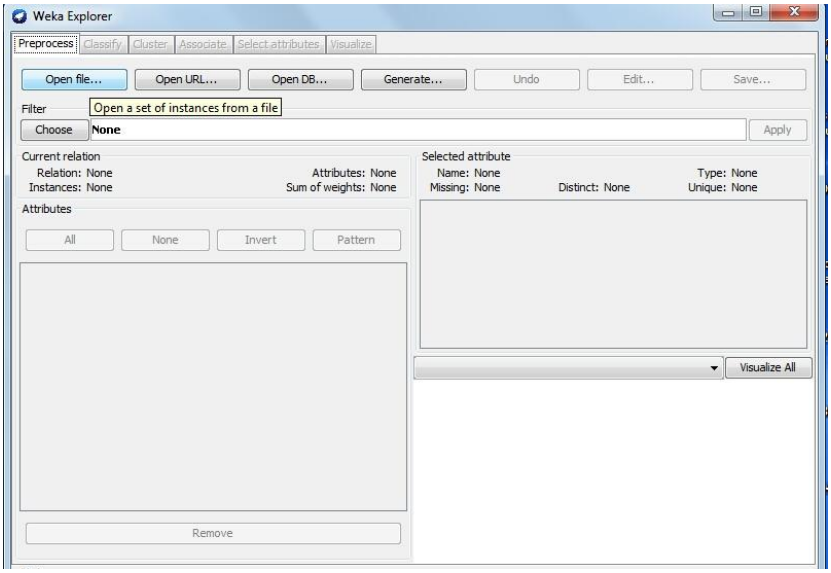


Fig (1_a) the software page of choosing dataset

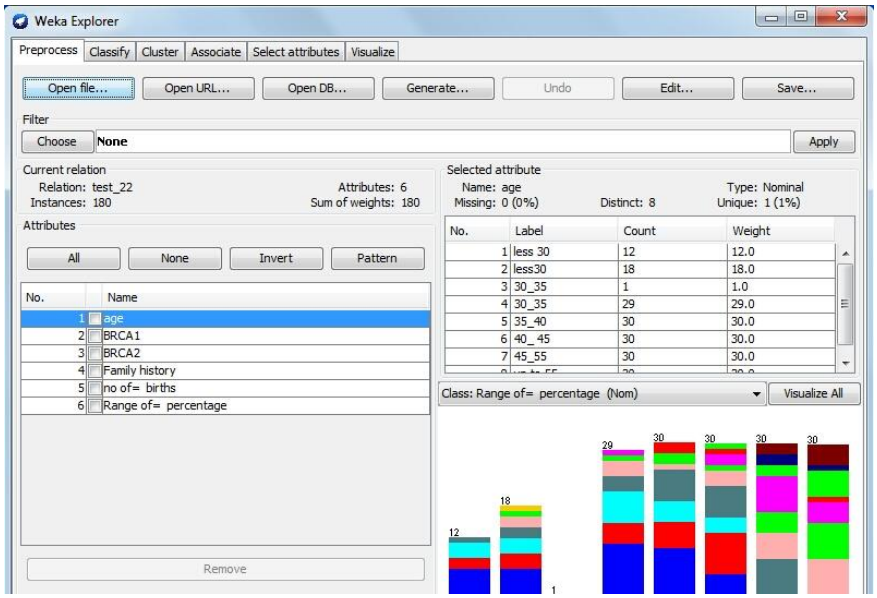


Fig (1_b) Weka software page downloading the dataset and classifier

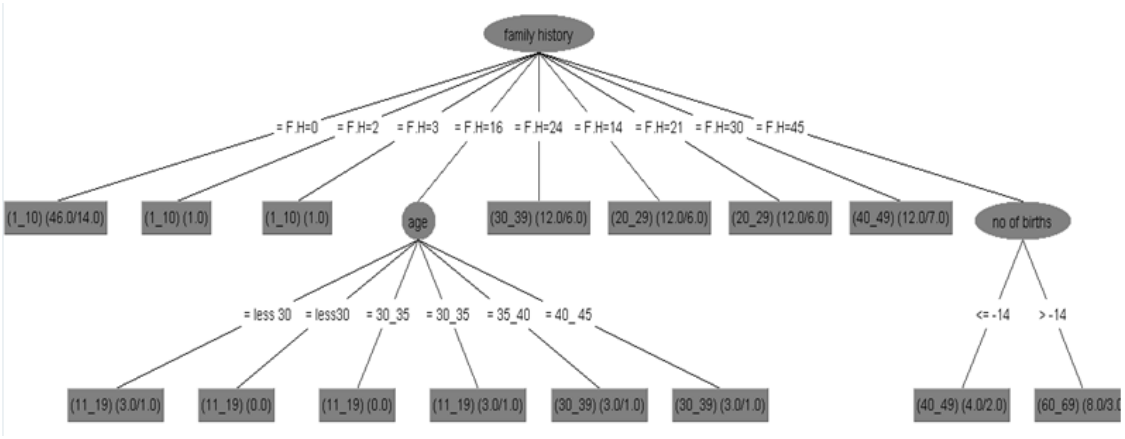


Fig (2) classification tree by using Id3 algorithm

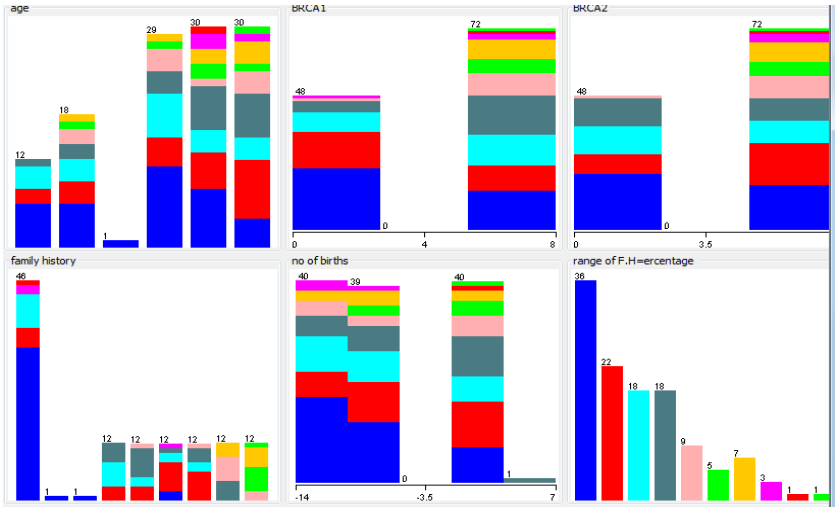
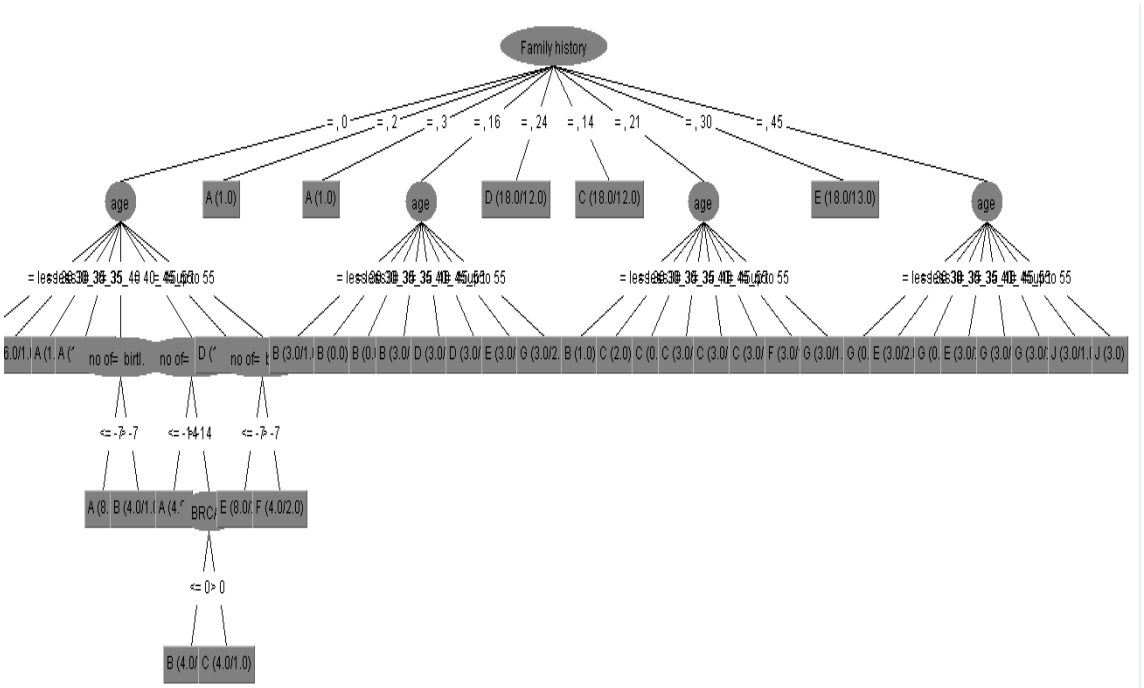


Fig (3) graphical representation

```
(1_10) (20_29) (11_19) (30_39) (40_49) (60_69) (50_59) (11_20) (21_30) (70_73)
```

The figure shows that the first layer attribute is the family history, which means that it has priority to other factors such as age and the no of births. The values related to the F.H represents the ability of woman to develop cancer as a percentage while the value in the nodes represents the range of future possibility to develop cancer. Thus, F.H gives either a direct rule output (gives a prediction without the need to other factors) for example if the family history is 24 the range is (30_39) percentage, or a second rule output (the output is the combination of two factors). As an example if family history 16 and the age is, (30_30) percentage the future range is (11_19) percentage the graphical representation of the dataset shown in fig (4) with the indication of each color.

By using the second dataset group in the Same algorithm this data set group2 includes the same factors in dataset 1 take in mind that the age factor in this group the age above55 year so that a different classifier tree had been produced as shown in fig(4)



Fig(4) classification tree of data set 2

We can see from the tree above that the decision-making process may differ from earlier classifier tree shown in fig(2). When it was taking the experimental samples of age above the 55 so that the classification process depends on family History as a first layer attributes, and took into account other characteristics such as age and no of births. as a second rule attributes and also the BRCA1 in stages Category at other levels (stage classification in the second level has been relying on the age and in the third level of some of the cases have been relying on the number of births and then was relying on BRCA1)

For example if F.H is 45, age 35 then the predictive 30
If F.H is 0, age 45, no of births 2, and BRCA1 is 4 then the predictive range 4.1

The graphical representation of dataset2 shown in fig (5)

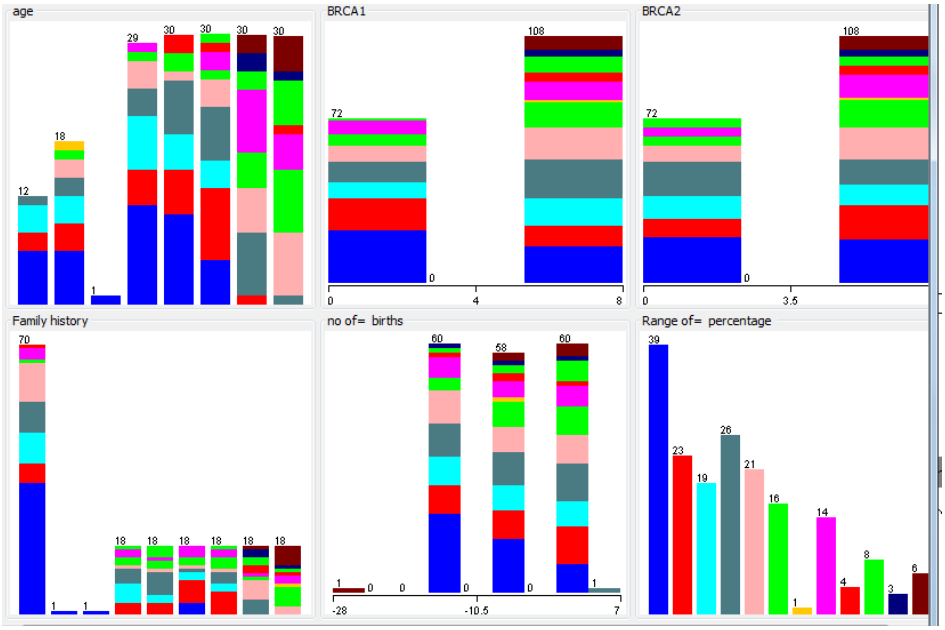


Fig (5) graphical representation

The classification of a\two data set groups is done, now we are going to test the efficiency of this predictive system by applying incomplete data that only one or two attributes such as age (40_45) and PRDCA1 is 15 only this data enter to the system and the output classification tree is shown in fig(6). The id3 algorithm will classify the attributes in an intelligent method predict the future range of cancer development according to the available information.

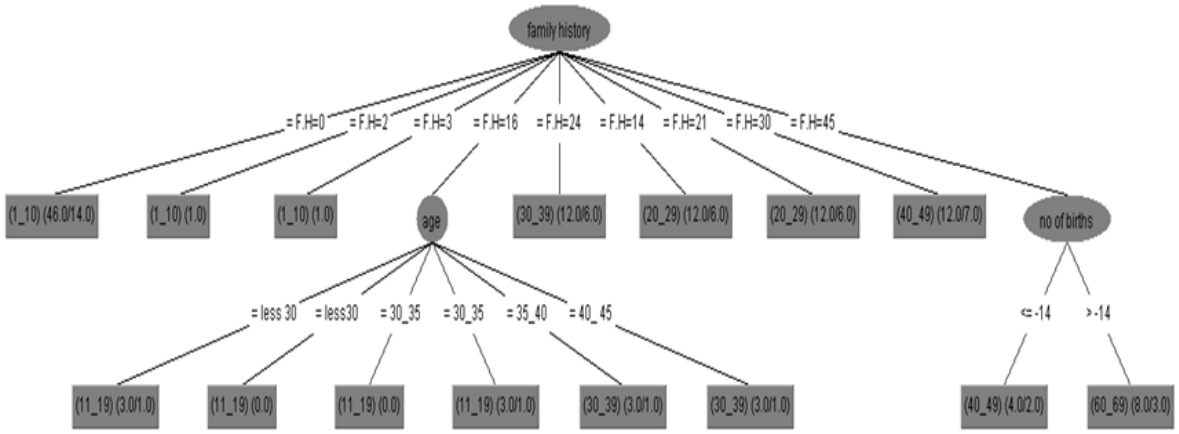


Fig (6) accuracy test of predictive system

IV. CONCLUSION

From the ID3 predictive system and the two datasets we conclude that the family history is the most affecting risk factor (attribute) that that helps in a future prediction probability to develop breast cancer , this is obvious from tree classification and this system is efficient to predict future cancer development

REFERENCES

[1]. MARY CIANFROCCA, LORI J. GOLDSTEIN. Prognostic and Predictive Factors in Early-Stage Breast cancer. Fox Chase Cancer Center, Philadelphia, Pennsylvania, USA, *the Oncologist* 2004;9:606-616.

[2]. Robert Gramling1*, Timothy L Lash2, Kenneth J Rothman[34] et.al Family history of later-onset breast cancer, breast healthy behavior and invasive breast cancer among postmenopausal women: a cohort study. Gramling et al. Breast Cancer Research 2010, 12:R82.

[3]. Azim HA Jr, Santoro L, Pavlidis N, Gelber S, Kroman N, Azim H, Peccatori FA. Safety of pregnancy following breast cancer diagnosis: a meta-analysis of 14 studies. *Eur JCancer*. 2011 Jan;47(1):74-83. Epub 2010 Oct 11.

[4]. S. Syed Shajahaan[1], S. Shanthi[2], V. ManoChitra[3]. Application of Data Mining Techniques to Model Breast Cancer data. International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 11, November 2013.

[5]. Shweta Kharya. USING DATA MINING TECHNIQUES FOR DIAGNOSIS AND PROGNOSIS OF CANCER DISEASE. International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012.

[6]. AbdelghaniBellaachia, ErhanGuven. Predicting Breast Cancer Survivability Using Data Mining Techniques. The George Washington University Washington DC 20052

[7]. Sung-Hyuk Cha Charles Tappert . A Genetic Algorithm for Constructing Compact Binary Decision Trees. JOURNAL OF PATTERN RECOGNITION RESEARCH 1 (2009) 1-13.

[8]. A Classifying Text with ID3 and C4.5.

[9]. June 2001. March 2007 <http://www.ddj.com/184410304/>.