Prediction of Recurrence of Breast Cancer

Harshitha E,Megha M Goutham,Nagamallesh M P,Chaithanya C,Divya C D
UG Students , Dept of CSE,Vidyavardhaka College of Engineering,Mysuru
Assistant Professor, Dept of CSE, Vidyavardhaka College of Engineering,Mysuru
Affiliated to Visveswariah Techonological University,Belgaum

I. INTRODUCTION

People care deeply about their health and want to be in charge of their health and healthcare. Life is more hectic than has ever been; the medicine that is practiced today is not only based on years of practice but on the latest discoveries as well. Tools that can help us manage and better keep track of our health such as Google Health are reasons and facts that make people more powerful when it comes to healthcare knowledge and management. The traditional healthcare system is also becoming one that embraces the Internet and the electronic world. Researches and studies show that Decision support—the ability to capture and use quality medical data for decisions in the workflow of healthcare are helpful in obtaining treatments tailored to specific health needs.Breast cancer is the most common type of cancer and it is the second leading cause of cancer death among women. It is not just a woman's disease. It is quite possible for men to get breast cancer, although it occurs less frequently in men than in women. After increasing at an alarming rate for more than 20 years, breast cancer incidence rates in women began decreasing in 2000 and dropped by about 7% from 2002 to 2003. But statistics have shown that nearly 1.7 million new cases had been diagnosed in 2012 which is second most common cancer overall. This represents about 12% of all new cancer cases and 25% of all cancers in women. Early detection and prediction is considered to be the best way to fight against this deadly disease. Most importantly, predicting the recurrence of cancer has become a real-world medical problem. Recurrent breast cancer is a cancer that comes back in the same or opposite breast or chest wall after a period of time when the cancer could not be detected. Recently, Data mining has become a popular and efficient tool for knowledge discovering and extracting hidden patterns from large datasets. It involves the use of sophisticated data manipulation tools to discover previously unknown, valid patterns and relationships in large dataset. Data sets having less attributes and higher instances can provide good result than the result we have got using this data set where it has higher attributes and less instances. Too much attributes can miss guide a classifier from gaining its maximum result, which gave us the idea of feature selection method. Feature selection algorithm gave us upper ranked attributes as well as better result than the result we got without feature selection algorithm.

II. COMPARISION OF CLASSIFICATION ALGORITHMS

SR.NO	ALGORITHMS	FEATURES	LIMITATIONS
1	C4.5 algorithm	• Build models can be easily interpreted. • Easy to implement. • Can use both discrete and continuous values. • Deals with noise.	• Small variation in data can lead to different decision trees. • Does not work very well on a small training data set. • Overfitting.
2	ID3 algorithm	• It produces the more accuracy result than C4.5 algorithm. • Detection rate is increased and space consumption is reduced.	• Requires large searching time. • Sometimes it may generate very long rules which are very hard to prune. • Requires large amount of memory to store tree.
3	K-Nearest Neighbor	• Classes need not be linearly separable. • Zero cost of learning process. • Sometimes it is robust with regard to noisy training data. • Well suited for multimodal classes.	• Time to find the nearest neighbours in a large training data set can be excessive. • It is sensitive to noisy or irrelevant attributes. • Performance of algorithm depends on the number of dimensions used.
4	Naive Bayes algorithm	• Simple to implement. • Great computational efficiency and classification rate. • It predicts accurate results for most of the classification and prediction problems.	• The precision of algorithm decreases if the amount of data is less. • For obtaining good results it requires a very large number of records.
5	Support vector machine algorithm	• High accuracy. • Work well even if data is not linearly separable in the base	• Speed and size requirement both in training and testing is more.

		feature space.	• High complexity and extensive memory requirements for classification in many ways.
6	Artificial Neural Network algorithm	• It is easy to use, with few parameters to adjust. • A neural network learns and reprogramming is not needed. • Easy to implement. • Applicable to a wide range of problems in real life.	• Requires high processing time if neural network is large. • Difficult to know how many neurons and layers are necessary. • Learning can be slow.

II. METHODOLOGY

This system is a medical oriented application which analyses the set of attributes using Data Mining. It makes use of Feature Selection technique for extracting the more important data for prediction. Our work predominantly focuses on detecting life threatening diseases like Breast Cancer using Classification algorithms. Proposed system is automation for breast cancer disease prediction using classification technique "Naive Bayes".

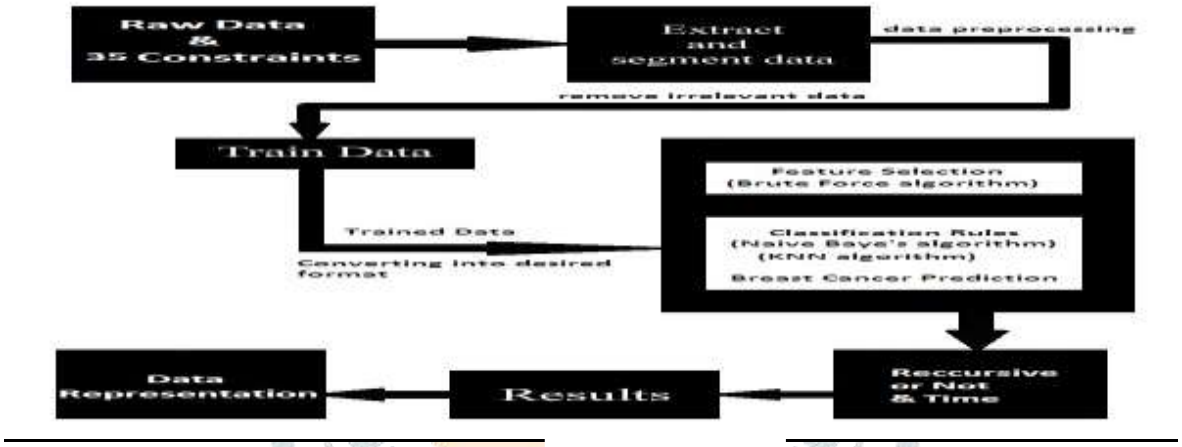

Fig 3.1 Block diagram of proposed system

In the proposed system, the raw data and the constraints of the patient are collected and stored in the server from which the constraint values are extracted and segmented using data processing method which removes the irrelevant data such as patient personal information. This data is called the train data which is given as input to the feature selection algorithm. The most important features are selected from the inputted constraints using Brute Force algorithm. The output of the feature selection algorithm is given as input to the classification algorithm. Naive Baye's algorithm is implemented to predict whether the Breast Cancer is recurrent or non-recurrent. The Figure 3.1 shows the block diagram of the proposed system

.

III. RESULTS AND DISCUSSION

Using prediction model to classify recurrent or non-recurrent cases of breast cancer is a research that is statistical in nature. Still this work can be linked to bio-medical evidences. In this project the dataset is used for predicting the recurring or non-recurring nature of the disease. This might help oncologists to differentiate a good prognosis (non-recurrent) from a bad one (recurrent) and can treat the patient more effectively.

Training Dataset

PatientName	Diabetest	Weight	Height
Anil	0	55	160
Ajay	5	55	155
Akash	5	60	165
Kumar	3	55	160
Punith	2	55	160

Now retrieve the distinct values from each feature
Diabetest (0,3,2,5)
Weight(55,60)
Height(140,155,160,165)
Calculate gain
Gain(Diabetes) - 2 [number of occurrences of 5]
Gain(weight) - 4 [number of occurrences of 55]
Gain(Height) - 3 [number of occurrences of 160]

Now calculate Model Score
Diabetes [Model Score=2.0 * 2/2 =2.0]
Weigth [Model Score=2.0*4/2=4.0]
Height [Model Score=2.0*3/2=3.0]

Sample Taken

Attributes (Constraints) – S1,S2,S3 [m=3]
Subject (Disease) – Recursive, Not Recursive[p=1/2=0.5]

Training Dataset

Patient Name	S1(X,Y,Z)	S2 (A,B,C)	S3 (P,Q,R)	Disease (subject)
Anil	X	A	P	Recursive
Ajay	X	B	Q	Recursive
Arun	Y	B	P	Not Recursive
Kumar	Z	A	R	Recursive
Naveen	Z	C	R	Not Recursive

New Patient data – Akash Constraints (S1 -X,S2-A,S3-R) Disease – Recursive/ Not Recursive
P=[n_c + (m*p)]/(n+m)

Recursive	**Non Recursive**
X P=[n_c + (m*p)]/(n+m) n=2, n_c=2,m=3,p=0.5 p=[2+(3*0.5)]/(2+3) p=0.7	X P=[n_c + (m*p)]/(n+m) n=2, n_c=0,m=3,p=0.5 p=[0+(3*0.5)]/(2+3) p=0.3
A P=[n_c + (m*p)]/(n+m) n=2, n_c=2,m=3,p=0.5 p=[2+(3*0.5)]/(2+3) p=0.7	A P=[n_c + (m*p)]/(n+m) n=2, n_c=2,m=3,p=0.5 p=[2+(3*0.5)]/(2+3) p=0.3
R P=[n_c + (m*p)]/(n+m) n=2, n_c=1,m=3,p=0.5 p=[1+(3*0.5)]/(2+3) p=0.5	R P=[n_c + (m*p)]/(n+m) n=2, n_c=1,m=3,p=0.5 p=[1+(3*0.5)]/(2+3) p=0.5

Recursive – 0.7 * 0.7 * 0.5 * 0.5 (p) =0.1225 **Not Recursive** – 0.3 * 0.3 * 0.5 * 0.5 (p) =0.0225

Since **Recursive** > **Not Recursive.** So this new patient is classified to **Recursive.**

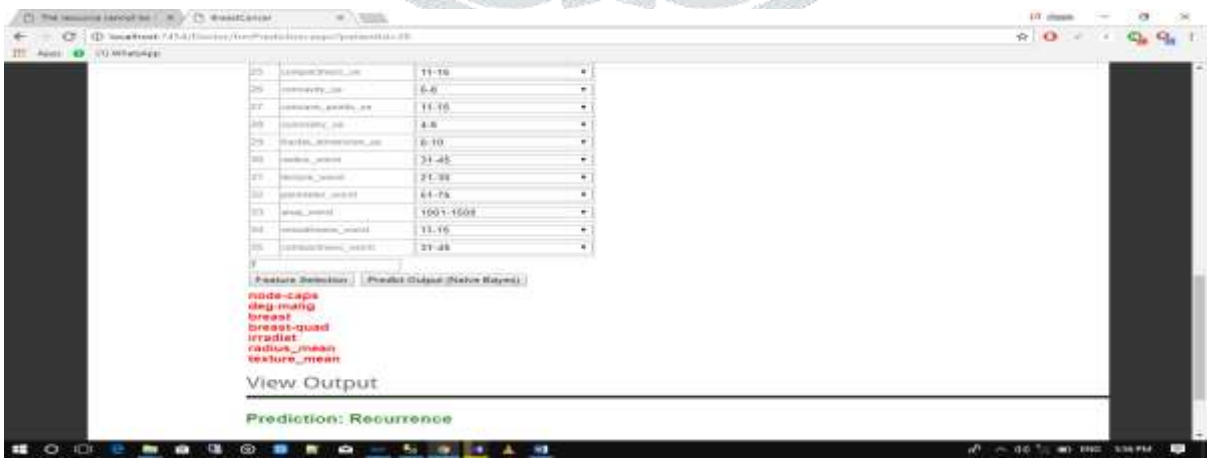

Figure 3.1 Predicting result- Recurrence

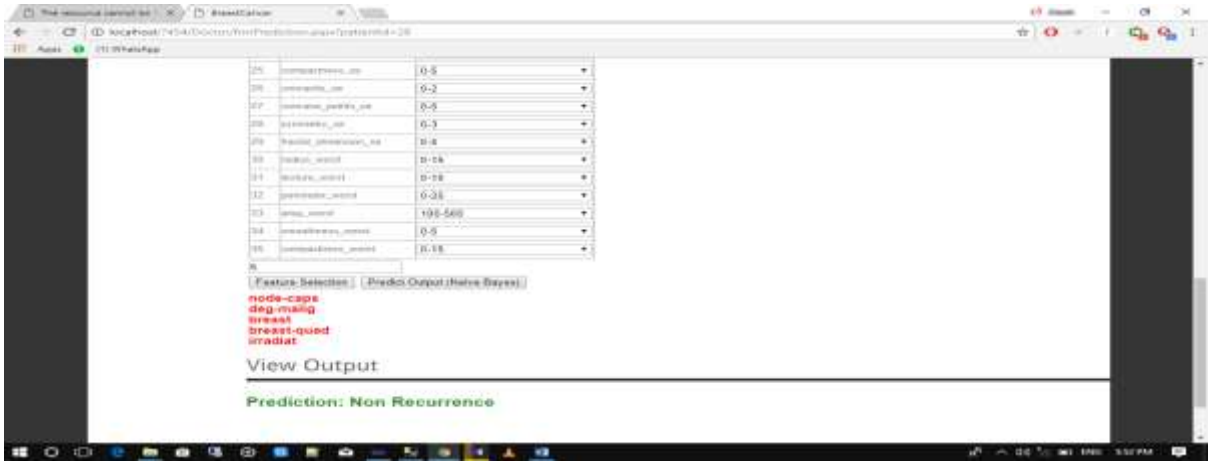

Figure 3.2 Predicting result- Non Recurrence

IV. FUTURE ENHANCEMENT

1. Time enhancement- In the proposed system, if the result of the prediction is recurrence then we can predict the time of recurrence.
2. Percentage of the result- If the result of the proposed system is recurrence or not, we can predict the percentage of accuracy.

REFERENCES

[1] Ahmed Iqbal Pritom, Md. Ahadur Rahman Munshi, Shahed Anzarus Sabab, Shiha Buzzaman Shihab. "Predicting Breast Cancer Recurrence using effective Classification and Feature Selection technique, 2016.

[2] Uma Ojha, Dr. Savitha Goel. "A Study on Prediction of Breast Cancer Recurrence using Data Mining technique, 2017.

[3] Joshi, Jahanvi, Rinal Doshi, and Jigar Patel. "Diagnosis of breast cancer using clustering data mining approach." International Journal of Computer Applications 101.10 (2014).

[4] Chaurasia, Vikas, and Saurabh Pal. "A novel approach for breast cancer detection using data mining techniques." International Journal of Innovative Research in Computer and Communication Engineering 2.1 (2014): 2456-2465.

[5] Liu, Huan, and Lei Yu. "Toward integrating feature selection algorithms for classification and clustering." IEEE Transactions on knowledge and data engineering 17.4 (2005): 491-502.