Breast Cancer Prediction Using Naive Bayes Classifier

VIGNESHWARAN G (210701307) , JAGATHRATCHAHAN V (210701701)

Rajalakshmi Engineering College (REC)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

Abstract: In this paper we present a prtd.cti ve model to identify the t)pe of breast cancer as benign or malignant. For this purpose. we develq>ed OtW own naive bayes classifier v.hich helps oncologist in diagnosing the cancer type with in no time and then helps oncologist in decision making in treatment method for the same purpose, we have taken dataser from uci ml repository which consislS of 699 valid instances and 10 attributes on the basis of which we will find out the type of cancer one is suffaing from. We have used our own algorilhm1oclean the data by providing the missing 1uple a vaJid vaJuc based on 1hc nearby anribmc value. unlike wcka which skips 1he in• slances wi1h missing tuples. After a series of procedures to cleanse the data. we applied machine learning algo rithm: na'ive bi.yes. usingjava net beans interfa~ to predict the type ofbrta.'it cancer. In this study, we compire lhe 4 machine learning algori1hm~:- smo. bayes network, naive bayes, j•48 decision to the same data. Aflcr comp.1rt son with wcka. it has been found that our implcmcntalion of the machine learning algorithm naive baycs on java ~tbtans interface pm!ict bcuer and provicrs better accuracy.

K~yword: UCI ML repository. WEKA. Narve Bayes. JAVA Net beans. machine learning

1. Introduction

Breast cancer is a malignant tumor that starts in cells of the breast. A malignant tumor is a group of cancer cells that spread into distant areas of the body [1]. Breast Cancer, one of the commonest malignancies, is a major cause of death among women in developed countries like UK, USA and in developing countries like India[2]. With the growth of developing countries grows the risk of suffering from diseases like breast cancer among its people[3]. An analysis has shown that survival rate is 88% after 5 years of diagnosis and 80% after 10 years of diagnosis .Therefore it is necessary to detect breast cancer at earliest stage possible[4].

The data provided by UCI repository[5] is quite helpful in identifying the attributes that count in investigating the type of breast cancer one is suffering from. The attributes we have taken into account are:

1. Sample code number Id-number
2. Clump thickness 1-10
3. Uniformity of cell size 1-10
4. Uniformity of cell shape 1-10
5. Marginal Adhesion 1-10
6. Single Epithelial cell size 1-10
7. Bare Nuclei 1-10
8. Bland Chromatin 1-10
9. Normal Nucleoli 1-10
10. Mitoses 1-10
11. Class (2 for benign, 4 for malignant)

After identifying the attributes we have to apply a machine learning algorithm to accurately predict the breast cancer type. So, we implemented Naïve Bayes Algorithm and compared our results with results of the tool WEKA [6]. On comparison we found that our model predicted more accurately.

2. Background Study

The Naive Bayes Classifier technique is based on the so-called Bayesian theorem and is particularly suited when the dimensionality of the inputs is high. Despite its simplicity, Naive Bayes can often outperform more sophisticated classification methods [7].

To demonstrate the concept of Naïve Bayes Classification, consider the example displayed in the illustration above. As indicated, the objects can be classified as either GREEN or RED. Our task is to classify new cases as they arrive, i.e., decide to which class label they belong, based on the currently exiting objects.

I. Introduction

Breast canccr is a malignant tumor tha starts in cells of the breast. A malignant tumor is a group of cancer cells that spread into distant areas of the body [I]. Breast Cancer, one of the com• monest malignaocies. is a major cause of death among women in developed countries like UK. USA and in developing countries like Jndi:a{2). Wilh the growth of developing countries grows lhc risk of suffering from diseases like breast cancer among its people[3). An analysis has shown that survival rate is 88% after 5 years of diagnosis and 80% after IO years of diagnosis .Thcrcfcre it is ncccs~ to detect breast cancer al earliest stage possibk:[4].

The data provided by UC! reposito,y(SJ is quite helpful in identifySlg the attriburs that count in investigating the type of bn:ast cancer one is suffering from The auributcs we have ttken into account are:

I. Sample code J'Rlmbcr Id-number
2.Clumpthickness 1-10
3. Uniformity of cell sire 1-10
4. Uniformity **of** cell sh'l)e 1-10
5. Marginal Adhesion 1-10
6. Single Epithelial cell size 1-10
7. Bare Nuclei 1-10
8. Blam Chromatin 1-10
9. Normal Nuclooli 1-10
10. **Mi.,ses 1-10**
11. Class (2 tor benign. 4 fer mali1']ant) After identifying the attributes we have to ap ply a machine learning algorithm to accurately predict the breast cancer type. So. we imple· mcntcd Naive Bayes Algorithm and compared our results with results of the tool WEKA (6). On comparison we found that our model pre• dieted more accurately.

2. Background Study

The Naive Bayes Classifier technique is based on the so•callc:d Bayesian theorem and is par• ticularly suited when the dimensiooality of the inputs is high. Despite its simplicity. Naive Bayes can often outperform more sophisticated classification methods (7].

To demonstrate the rnncept of NaYve Bayes Classification. consider the example displayed in the illustr..:ion above. As indicated. the ob jects can be classified as either GREEN or RED. Our task is to classify new cases as they arrive, i.e .. decide to which class label they be· long, based on lhe currently exiting objects.

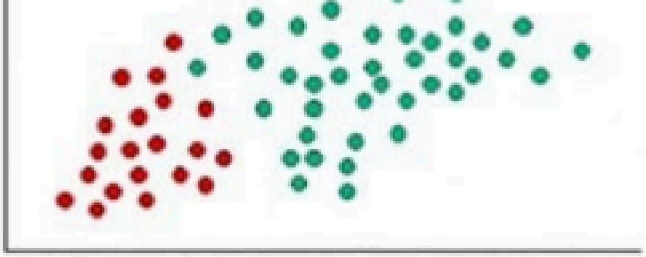

Fig. 1. Objects for Clasification

Since there are twice as many GREEN objects as RED, it is reasonable to believe that a new case (which hasn't been observed yet) is twice as likely to have membership GREEN rather than RED. In the Bayesian analysis, this belief is known as the prior probability. Prior probabilities are based on previous experience, in this case the percentage of GREEN and RED objects, and often used to predict outcomes before they actually happen.

Thus, we can write:

Probability for GREEN α $\frac{\textit{Number of GREEN objects}}{\textit{Total number of objects}}$

Probability for RED α $\frac{\textit{Number of RED objects}}{\textit{Total number of objects}}$

Since there is a total of 60 objects, 40 of which are GREEN and 20 RED, our prior probabilities for class membership are:

Probability for GREEN α $\frac{40}{60}$

Probability for RED α $\frac{20}{60}$

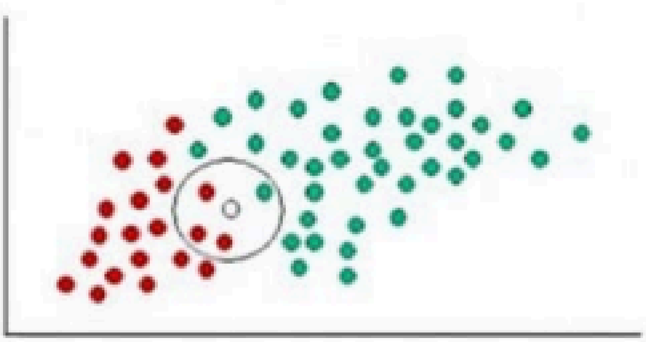

Fig. 2. Classification of newly arrived object

Having formulated our prior probability, we are now ready to classify a new object (WHITE circle). Since the objects are well clustered, it is reasonable to assume that the more GREEN (or RED) objects in the vicinity of X, the more likely that the new cases belong to that particular color. To measure this likelihood, we draw a circle around X which encompasses a number (to be chosen a priori) of points irrespective of their class labels. Then we calculate the number of points in the circle belonging to each class label. From this we calculate the likelihood:

Likelihood of X given RED α $\frac{\textit{Number of RED in the vicinity of X}}{\textit{Total number of RED cases}}$

From the illustration above, it is clear that Likelihood of X given GREEN is smaller than Likelihood of X given RED, since the circle encompasses 1 GREEN object and 3 RED ones. Thus:

Probability of X given GREEN α $\frac{1}{40}$

Probability of X given RED α $\frac{3}{20}$

Although the prior probabilities indicate that X may belong to GREEN (given that there are twice as many GREEN compared to RED) the likelihood indicates otherwise; that the class membership of X is RED (given that there are more RED objects in the vicinity of X than GREEN). In the Bayesian analysis, the final classification is produced by combining both sources of information, i.e., the prior and the likelihood, to form a posterior probability using the so-called Bayes' rule (named after Rev. Thomas Bayes 1702-1761).

Posterior Probability of X being GREEN α

Prior probability of GREEN × Likelihood of X given GREEN= $\frac{4}{6} \times \frac{1}{40} = \frac{1}{60}$

Posterior Probability of X being RED α

Prior probability of RED × Likelihood of X given RED

$= \frac{2}{6} \times \frac{3}{20} = \frac{1}{20}$

Finally, we classify X as RED since its class membership achieves the largest posterior probability.

3. Methodology

In this paper, we have implemented naïve Bayes algorithm to predict cancer type by using JAVA Netbeans interface and then compared the result with the other algorithm using WEKA.

To carry out this whole operation we have firstly cleansed the data through data mining techniques [9] and then applied Naïve Bayes algorithm to classify the breast cancer type as benign or malignant[10].

The dataset that we have used in our study is from UCI ML repository and it consists of 699 instances and 10 attributes. It has positive samples and negative samples and every sample has the 10 attributes defined for them.

Fig. I. Objects for Clasification

Since there are twice as many GREEN objects as RED, it is reasonable lo believe that a new case (which hasn't been observed yet) is twice as likely to have membersh~ GREEN rather than RED. In the Bayesi.11 analysis, this belief is known as the prior probabilily. Prior prob• abilities are based on previous experience, in this case the percentage of GREEN and RED objects. and often used to predict outcomes be· fore they actually happen.

Thus, we can write:

Probability for GREEN a ~:7a7ru~11~b~:e:;0:~!:::' Probability for RED a

;;r:::::,::::;:~~:~, Since there is a total of 60 objects, 40 of which are GREEN and 20 RED, our prior prd>abilities for class membership are:

Probability for GREEN a*

Probability for RED a ¥.

Fig. 2. Classification of newly arrived object

Having formulated our prior prob.bility, we arc now ready IO classify a new object (WHITE circle). Since the objects arc well clusterOO. it is reasonable to assume that the more GREEN' (or RED) objects in the vicinity of X, the m::,re likely that the new cases belong to that particu
lar color. To measure this likelihood, we draw a circle around X which encompas,es a number (to be chosen a priori) of points irrespective of their claS'i labels, Then we cakulak! the number of points in the circle belonging lo each class label. From thi'i we calrulatc the likelihood:

Likelihood of X given GREEN *a~"Mk••fCllfil:/111.,1wn,,:w,,., .1*

rfl lll" .... •/Clfff/1/c H•

Likelil1ood of X given RED a *"'•"'""~•r•uo1.1ow-1.uy•I* r

70,t•ln.,,IM'••f•t.11te HJ

From the illustration above, it is clear that Like lihood ofX given GREEN" is smaller than Like lihood of X gh~n RED, since the ciocle encom passes I GREEN d>jcct ard 3 RED ones. Thus: Probability ofX given GREEN a-}.

Probability ofX given RED a-!;;

Although the prior probabilities indicate that X may belong to GREEN (given th:i there are twice as many GREEN compared to RED) the likelihood indicates otherwise; that the class membership of X is RED (given that there arc more RED objects in lhe vicinity of X than GREEN). In the Bayesian analysis. the final classification is produced by combining bolh sources of infonmtion, i.e., the prior and the likelihood. to fonn a posterior prob:t>ility using the so-called Bayes' rule (named after Rev. Thomas Bayes 1102-1761).

Postcriu- Prob:bility or X being GREEN a Prior probability of GREEN x Likelihood of X given GREEN=!x~ =1o

Posterx:Jr Problnility d X being RED a Prior probability of RED x Likelihood of X given RID

=*Jx.!.* =.!.

6 20 20

Finally, we cbssify X as RED since its class membership achieves the largest posterior probability .

3. Methodology

In this paper. we have implemented naive Bayes ali,'Orithm to prcdi:t cancer t)pc by using JAVA Nctbeans i'lterface and then compared the result with the other algorithm using WEKA.

To carry out this whole operatiln we have firstly cleansed the dat.l thrwgh datl mining techniques (91 and then appticd Naive Bayes algorithm to classify lhc breast cancer type as benign *a* malignan1f 10).

The dataset that we have used in our study is from UCI ML repository and it consists of 699 instances and 10 attributes. It has positive sam ples and negative samples and every s.111plc has the JO anributc:s defined for lhcm.

The Naïve Bayes technique depends on the famous Bayesian approach following a simple, clear and fast classifier [11]. A naive Bayes classifier is a simple proabilitic classifier based on applying Bayes' theorem with strong (naive) independence assumptions. A more descriptive term for the underlying probability model would be "independent feature model". A naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature, given the class variable.

The different parameters that are computed are

Accuracy =(TP+TN)/(TP+FP+TN+TP)--- (1)

Sensitivity = TP/(TP+FP)---------- (2)

Selectivity = (TP+FP)/(TP+FP+TN+TP)---- (3)

Specificity = TN/(TN+FP)-------- --- (4)

Missed Alarm Rate = FN/(TP+FN)----- (5)

False Alarm Rate = FP/(TP+FP)------- (6)

From the confusion matrix to analyze the performance criterion for the classifiers in detecting breast cancer, accuracy, precision (for multiclass dataset), sensitivity and specificity have been computed to give a deeper insight of the automatic diagnosis [12]. Accuracy is the percentage of predictions that are correct. The precision is the measure of accuracy provided that a specific class has been predicted. The sensitivity is the measure of the ability of a prediction model to select instances of a certain class from a data set. The specificity corresponds to the true negative rate which is commonly used in two class problems. Accuracy, precision, sensitivity and specificity are calculated using the equations given above, where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives and FN is the number of false negatives[13].

4. Experimental Study

In this paper, accuracy of our own implemented naïve Bayes is compared with accuracy of four different algorithms on WEKA. Here, our goal is to have high accuracy, besides high precision and recall metrics.

5. RESULT AND DISCUSSION

Table 1 reflects the result that we are getting from our implemented algorithm. As reflected in the table our implemented algorithm pro- implemented algorithm provides better results when compared with other existing algorithm of WEKA (i.e. SMO, J-48 Decision etc.)

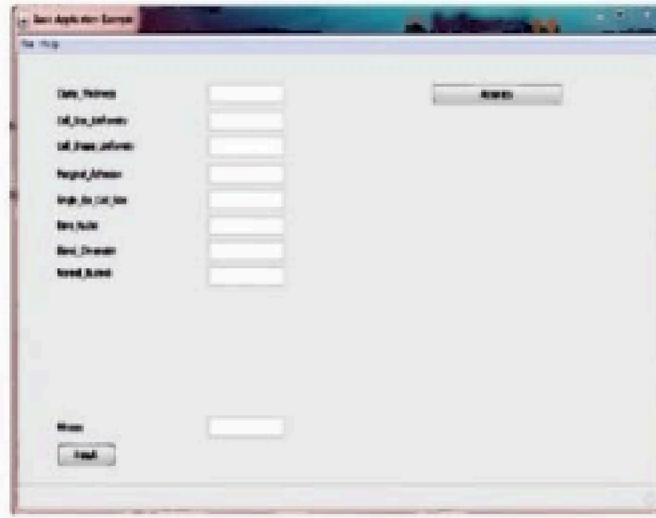

Fig.3. Application Interface

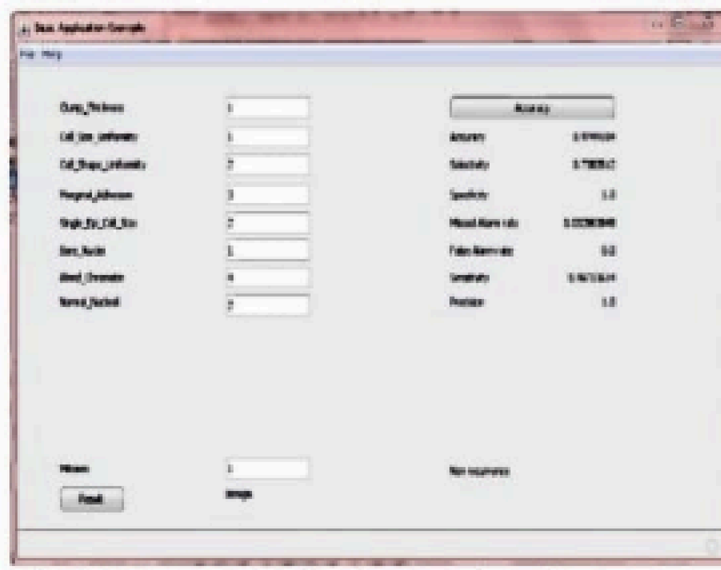

Fig.4. Integrated Application Interface showing result

Table 1. Result From Study

Accuracy	0.975
Precision	1.0
Sensitivity	0.967
Selectivity	0.738
Specificity	1.0
Missed Alarm Rate	0.0328
False Alarm Rate	0.0

The Naive Bayes technique depends on the fa mous Bayesian awroach following a simple.

clear and fast classifier l 11 i A naive Bayes classifier is a simple proabilitic classifier based

on applying Bayes' theorem with strong (naive) indcpcrdcnce assumptions. A more de;criptive tenn for the underlying probability model would be "independent feature model'". A naive Bayes classifier assumes that the presence (or absence) of a particular feature of a class is un related to the presence (or abscnre) of any other feature, given the class v-J.riable.

The dilfcrent parameters that are computal are
Accuracy =(TP+ TN)/(TP+FP+ TN+ TP)--- (I)
Sensitivity= TP/(TP+FP)---------- (2)
Selectivity= (TP+FP)l(TP+FP+ TN+TP)---- (3)
Specificity =TN/(TN+FP)------- --- (4)
Missed Alann Rate= FN/(TP+R'I)----- (5) False Alann Rate= FP/(TP+FP)--· (6)

From the confusion matrix to analyze the per formance criterion for the classifiers in detect *ing* brea~t cancer. accuracy, precision (for mul ticlass dataset). sensitivity and specificiy have been computed to give a deeper insight of the automatic diagnosis [12J. Accuracy is the per centage of predictions that are correct The pre cision is the measure of accuracy provided that a specific class has been predicted. The sensi tivity is the mca~ure of the ability of a predic tion model to select instances of a certain class from a data set. The specificity corresponds to the true negative rnte which is commonly used in two cl~ problem;. Accuracy. prectiion. sen sitivity and spccifrity arc calculated using the equations given above, where TP is the number of true positives, TN is the number of true nega tives, FP is the number of false positives and FN is lhe number of false negatives[l3).

4- **Experimental Study**

In this paper. accurocy of our own implanented na'ive Bayes is compared with accuracy of four differmt algorithms on WEKA. Here. our goal is to have high accuracy, besides high precision and recall metrics.

5- **RESUU' AND** DISCUSSION

Table I reflects the result that we arc gelling from our implemented algorithm. As reflected in the table our implerrcntcd algorithm pro vides better results than WEKA in tenns of *Ac* curacy. Sensitivity. and Sp:cificity. Also. the implenrntcd algorilhm provides better results when compared with ott-cr existing alg<rithm of WEKA (i.e. SP...10, J-48 Decision etc.)

Fig.3. Application Interface

F1g.4. Integrated Application Interface show ing result

Table I Result From Studv

Accuracy	0.975
Precision	1.0
Sensitivity	0.967
Sclec1ivity	0.738
Specificity	1.0
MissedAlann	0.0328
FalseAlann	O.O

Accuracy	Precision	Sensitivity	Specificity
94.762%	0.948	0.961	0.935
96.19%	0.962	0.97	0.948
95.714%	0.957	0.962	0.935
93.80%	0.938	0.940	0.897

and Future Work

ented using machine learning in diagnosing cancer type into decision taking for cancer pa- se we have implemented Naïve ing JAVA Net beans interface. esults show that our approach d provides better accuracy in er type as benign and malig-

at results are better for our im- ayes algorithm as compared to ning algorithm – SMO, Bayes ves, J-48 Decision.

ply data cleaning algorithm to to include the records with ell on the basis of value of ich are not included in case of

clude experimenting other ma- rithm using JAVA Net beans e hybrid algorithm which is a ting two or more algorithms to model which can predict with le 2 reflects the result that we are A.

cer society. Breast cancer facts 05-06 (http://www.cancer. org)

Breast Cancer (http:// g /docroot/CRI/ ?dt=5)

[3] Breast Cancer (http://www.cancer.gov /cancertopics/types/breast)

[4] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques, 2nd Edition. San Fransisco: Morgan Kaufmann; 2005..

[5] Breast Cancer dataset. http://archive.ics.uci edu/ml /datasets /Breast+ Cancer +Wisconsin +%28Original.

[6] Weka(http://weka. sourceforge.net /doc/ weka/ classifiers)

[7] Naïve Bayes Classifier.www.statsoft.com /textbook/naïve bayes –classifier

[8] V. Vapnik. The Nature of Statistical Learning Theory. NY: Springer-Verlag. 1995.

[9] Tang, Z., and MacLennan, J., Data Mining with Sql Server 2005. Wiley, 2005

[10] Delen, D., Walker, G., and Kadam, A., Predicting breast cancer survivability: a comparison of three data mining methods. Artif. Intell. Med. 34:113–127, 2005.

[11] Witten, I. H., and Frank, E., Data mining: practical machine learning tools and techniques. Morgan Kaufmann – Academic Press, America, p. 525, 2005.

[12] Alireza Osareh, Bita Shadgar, Machine Learning Techniques to diagnose Breast Cancer. IEEE, 2009

[13] Subbalakshmi G. , Ramesh K., Chinna Rao M., Decision support in Heart Prediction System using Naïve Bayes, IJCSE ,2010.

94.762%
0.948
Bayes
Network
96.19%
0.962 SMO
95.714%
0.957 J-48
Dec~ion
93.80%
0.938

6.

Table 2. Result from WEKA**Condusion and Future Work**

Al2orilhm Accuracv Precisi>n Naive Ba~sThis J>.l)Cr ift1)1emented usilg macl:ine learning technique is helpful in diagnosing cancer type into assisl oncologist in decision taking for cnnccr pa tient. N>r 1his purpose we have implemented Naive Bayes algorithm using JAVA Net beans interface. The experimental results show that our approach performs bellcr and pro,idcs bca.cr accuracy in predictl'lg the coocer type as benign illl.d m1lig nan1.

This study shows that rcsulls arc bctta for our im plemcnlal Narve Bayes algorithm as compared to other machine lcuming algorithm - SMO, Bayes Nc1work, Naive Bayes. J48 DccisK>n.

In this work, we upply data cleaning algorithm to clean the data and to include the records with missing data as well on the basis of value of nearby attribule, which arc not included in case of WEKA.

Future work will include experimenting olhcr ma chine ICUTiing algorithm using JAVA Ncl beans in1crfacc o' to make hybrid algoridlm which is a combinalion of existing two oc' more algorithms to create a predictive model which can predict with higher accura:y. Table 2 *ttfleas* the result lhal we arc getting t.tr tool WEKA.

RrfrrrnctS

(I) Amaican cancer society. Brca~n cancer facts and figures 2005-06 (hllp://www.cancer. org) [2) Ova-view: Breast Cancer (hup:// www.canca-.org /docrooL'CRl/ CRl_2_lx.asp?dt"6)

Sensitivitv	S.......,.ificitv
0.961	0.935
0.97	0.948
0.962	0.935
0.940	0.897

)3) Bn:ist Cancer (htq>://www.c:ncer.gov /canccropics/t)pes/breast)

(4) Ian H. Wiuen and Eibe Frank. Data Min• ing: Practical machine Caming tools .Ed techniques. 2nd Edition. San Fransisco: Morgan Kaufmann; 2005 ..

(SJ Brea.-« Canca-dataseL hup://archi\'C.ic5.uci edu/ml /datasets /Breas1+ Cancer +Wisconsin +%28Origilal.

[6] Vkka(http://wcb. sourceforgc.nct /doc/ weka/ cbssifiers)

(7] Naive Bayes Classificr.www.statsofl.com /tei\tbook/nai\'e bayes -classifia-

(8) V. Vapnik. The Nature of Statistical Learn• ing Theory. NY: Springcr-\1:rlag. 1995. (9) Tang.Z .. and Macl..ennan. J., Data Minilg with Sql Servcr2005. Wiley. *2005*

[IOJ Dek:n, D .. Walker. G .. and Kadam. A .. PredictTig breast cancer survivability: a comparison of three data mining methods. Anif. lntcll. Med. 34: 113-127. 2005.

(I I] Witten, I. H .. and Frank, E., Data mining: practical machine learning 1001s and tech niques. Morgan Kaufmam - Academic Press, America. p. *525,* 2005.

[12) Alireza Osareh. Bita Shadgar. Machine Leaming Techriqucs to diagnose Breast Cancer. IEEE. 2009

[13) Slilbalakshmi G .. Ramesh K., Chinna Rao M .. Decision suppon in Hean Predic tion System using Naive Bayes. IJCSE .20!0.