Competitive analysis of classification algorithms for Breast Cancer

Saima Asghar
Department of Computer Scince
Riphah International University
Vehari, Pakistan
zoniaangel666@gmail.com

Samreen Saleem
Department of Computer Scince
Riphah International University
Lahore, Pakistan
samreensaleem125@gmail.com

Muhammad Usman
Department of Computer Scince
Riphah International University
Burewala, Pakistan
Ranamuhammadusman919gmail.com

Abstract— **Data mining is a region of computer programming with a colossal approaching, which is the route toward finding or removing information from enormous data set or datasets. There are a wide scope of districts under Data Mining and one of them is Classification or the controlled learning. Request similarly can be executed through different procedures or computations. We have driven the relationship between's three figurings with help of WEKA (The Waikato Environment for Knowledge Analysis), which is an open source programming. It contains unmistakable sort's data mining computations. This paper explains discussion of Decision tree, Bayesian Network and K-Nearest Neighbor figurings calculated relapse. Here, for taking a gander at the result, we have used as limits the adequately requested models, incorrectly gathered events, time taken, kappa estimation, relative incomparable error, and root relative squared goof.**

Keywords— Classification algorithms, Breast cancer, Weka

I. INTRODUCTION

The subsequent driving reason for death among ladies is bosom malignant growth and it comes straightforwardly after cellular breakdown in the lungs [1]. The wellbeing and clinical area is more needing information mining today. At the point when certain information mining strategies utilized, important data can be extricated from huge information base and that can help to clinical professional to take choice, and improve wellbeing administrations. There are a couple of contentions that can uphold the utilization of information digging in wellbeing area for bosom malignant growth like early recognition, early evasion, and sign based drug, amending emergency clinic information blunders [2].WEKA is an amazing asset as it contains managed learning too solo learning techniques. It contains Classification, Clustering, Association Mining, Feature Selection, Data Visualization, and so forth The primary purpose for utilizing WEKA is encourages analysts like us to actualize and think about information mining methods effectively on genuine or manufactured information. It is additionally appropriate for growing new machine learning methods. WEKA goes under the open source programming gave under GNU General Public License .Bosom disease is a harmful or generous tumor, inside bosom, wherein cellsdivide and develop without control [3, 4]. Researchers have attempted to know the specific purpose for bosom malignancy, as there are a couple of danger factors which increment the like hood of a lady creating bosom disease. Age, hereditary danger and family ancestry are whatever components being considered for bosom malignancy [3]. Therapies of bosom malignancy are partitioned into two sorts, nearby and efficient. Medical procedure and radiation are neighborhood kind of therapies while chemotherapy and chemical treatment are instances of methodical treatments. For getting best outcomes, the two therapies are utilized together in various varieties according to the patient and illness power. [3]

II. ALGORITHMS USED

A. *Decision tree*

Decision tree is amazing arrangement calculation in information mining. There are a few well known choice calculations, for example, Quinlan's ID3, C4.5, C5, and CART [8]. A choice tree is a stream outlining like structure, where each inside hub indicates a test on a trait, each branch speaks to a result of the test, and each leaf hub holds a class mark [5]. This strategy isolates perception into branches to build tree on reiteration premise. As a rule, tree classifiers perform arrangement in two phases: tree-developing and tree-pruning. The tree-developing is top down methodology. In this stage, the tree is part in a recursive way called recursive parceling. It is finished.

B. *Naïve Bayes*

Bayesian classifier is a factual classifier just as a directed Learning strategy. It will foresee class participation probabilities. It gives helpful recognition to comprehension and surveying many learning calculations. It ascertains express probabilities for theory and it is vigorous to commotion in input information. At the point when Bayesian classifier is applied to huge database, it shows high exactness and speed. Innocent Bayesian classifiers expect that the impact of a quality incentive on a given class is autonomous of the estimations of different characteristics. This supposition that is called class-contingent autonomy. For our testing, we have utilized her Naïve Bayes characterization.

C. *K-Nearest Neighbor*

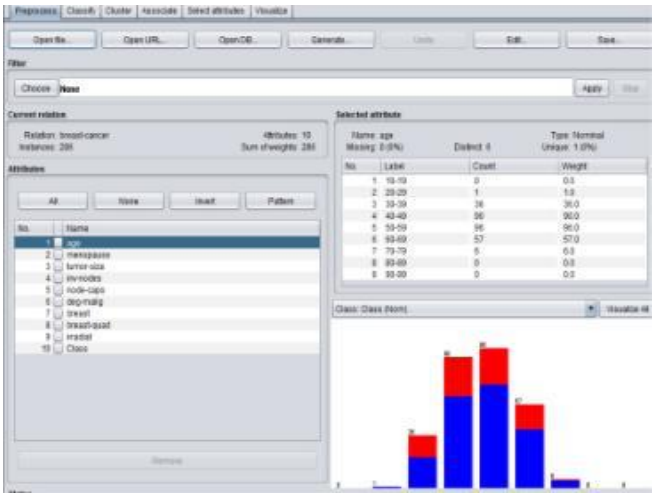
K-Nearest Neighbor classifier is otherwise called a separation based classifier. Closest neighbor classifiers depend on similarity learning. So implies, it is contrasting given test tuples and preparing tuples that are like it. The obscure tuple is appointed to most basic class among its K-closest neighbors. At the point when K = 1, the obscure tuple is doled out the class of the preparation tuple that is nearest in the example space. It is additionally utilized for numeric forecast, that is, it returns genuine worth expectation for obscure tuple [5]. This technique is likewise called the lethargic student strategy.

III. RESEARCH METHODOLOGY

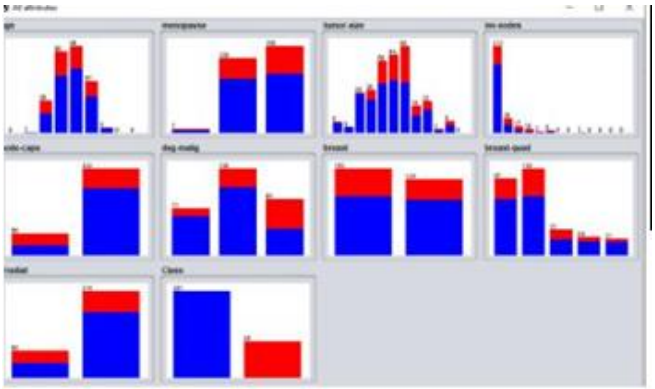
A. Dataset

Data contain 10 attribute and 286 instance

B. Data preprocessing



C. Data visualization



IV. WEKA

WEKA is uninhibitedly accessible device valuable for information mining. This device was created by University of Waikato, New Zealand. In this device numerous information mining calculations were actualized utilizing Java language. As indicated by created yield, calculations are gathered into various classes, for example, tree based, rule based, work based and so on. WEKA device is extremely helpful and easy to utilize and furthermore it is freeware subsequently utilized for present investigation. Here we utilized distinctive grouping calculations for heart expires and think about their exhibition based on schedule and precision [9]. Present investigation utilizes distinctive order calculation in particular J48, NaiveBayes, JRip, SMO, IBk. All the grouping calculations utilizes 'Use preparing set' strategy as a test mode and full preparing set as a classifier model in light of the fact that in this method dataset is partitioned into 10 sections. Each part is utilized as test dataset while other nine sections are utilized as preparing dataset. In each part and test set, execution is measure for various case based setup.

V. RESULT

A. Table.1 NaïveBayes

Serial no	Algorithms
1	bayesNet
2	NaïveBayes

B. Table2. Lazy

Serial no	algoritms
1	IBK
2	KStar
3	LWL

C. Tables3.Trees

Serial no	algoritms
1	J48
2	LMT
3	RANDOMFOREST
4	RANDOMTR

D. Table 4. Functions

Serial no	algorithm
1	logistic
2	multilayerPerceptron
3	sgd
4	sgdext
5	smo
6	Simple logistc

Table 5 Accuracy measure for NaïveBayes classifier

Algorithm	Correctly Classified Instance %	Incorrectly Instances %	Time of model	Mean absolute error	Relative absolute error	Root relative sequred error	Kappa Statistics
BayesNet	72	21	0.16	0.2919	78.7898	99.9047	0.2919
Naïve Bayes	71	28	0.05	0.3272	78.2086	99.1872	0.2857

Table6 Accuracy measure for Lazy classifier

Algorithm	Correctly Classified Instance %	Incorrectly Classified Instances %	Time to build model	Root absolute error	Relative absolute error	Root relative sequred error	Kappa Statistics
IBK	72	27	0.02	0.3257	77.8513	111.6114	0.2438
KStar	73	26	0	0.3354	80.1634	99.76	0.2864
LWL	72	270	0.02	0.3775	90.2189	96.6758	0.2721

Table 7 Accuracy measure for Decision tree classifier

Algorithm	Correctly Classified Instance	Incorrectly Classified Instances%	Time for model	Mean absolute error	Relative absolute error	Root relative sequred error	Kappa Statistics

J48	% 75	24	0.06	0.3676	87.8635	94.6093	0.2826s
LMT	75	24	1.64	0.3589	85.766	93.8755	0.3042
RANDOMFOREST	69	30	0.63	0.3227	89.0857	100.9171	0.1736
RANDOMTREE	66	33	0.03	0.3533	84.4448	124.6837	0.1855
REPTREE	70	29	0.03	0.3797	90.7409	101.7845	0.1601

Table 8 Accuracy measure for function classifier

Algorithm	Correctly Classified Instance %	Incorrectly Classified Instances%	Time for model	Mean absolute error	Relative absolute error	Root relative sequred error	Kappa Statistics
LOGISTIC	68	31	0.36	0.37	88.4196	101.3094	0.1979
MULTILAYERPERCEPTRON	64	35	9.08	0.3552	84.8811	118.654	0.1575
SGD	69	30	0.34	0.3007	71.8664	119.9713	0.2105
SGDTEXT	70	29	0.31	0.5452	71.0308	119.2717	0.2972
SMO	69	30	0.73	0.3042	72.7021	120.6668	0.1983
SIMPLE LOGISTIC	75	24	0.38	0.3589	85.766	93.8755	0.3042

VI. CONCLUSION

A few information mining arrangement strategies can be applied for the distinguishing proof and avoidance for bosommalignant growth among patients. In this paper, we have utilized classification technique on breast cancer dataset.

We have analyzed on various boundaries for forecast of malignant growth. Be that as it may, for predominant forecast, we center on precision and most reduced processing time. Our investigations separated all calculations dependent on most minimal registering time and exactness and we came up. After comparison the higher accuracy have 3 algorithms LMT,J48 and simple logistics 75%, 75% and 75% with time taken to build model 1.64, 0.06 and 0.38. So LMT, J48 and simple logistic have best accuracy.

REFERENCES

[1] U.S. Cancer Statistics Working Group. United States Cancer Statistics: 1999–2008 Incidence and Mortality Web-based Report. Atlanta (GA): Department of Health and Human Services, Centers for Disease Control

[2] V.Karthikeyani, I Parvin, K.Tajudin, I.Shahina Begam. Comparative of data mining classification algorithm in Diabetes disease prediction. International journal of computer application 2012.12.26-31

[3] Breast cancer Q&A/facts and statistics (http://www.kome-n.org/bci/bhealth/QA/q_and_a.asp)

Jerez-Aragone ´s JM, Gomez-Ruiz JA, Ramos-Jimenez G, Munoz- Perez J, Alba-Conejo E. A combined neural network and decision trees model for prognosis of breast cancer relapse. Artif Intell Med.