# COMPARATIVE ANALYSIS BETWEEN CLASSIFICATION ALGORITHMS AND DATA SETS (1: N & N: 1) THROUGH WEKA

**G. D. K. Kishore[1], Dr. M. Babu Reddy[2]**
*Research Scholar, Computer Science, Krishna University, India[1]*
*Assistant Professor, Computer Science, Krishna University, India[2]*
*kishore.galla1@gmail.com[1], babureddy@yahoo.com[2]*

---

***Abstract: Data mining is a stage in the knowledge disclosure process comprising of data mining algorithms that used to discovers patterns in the data. Data Mining likewise can be characterize as an analytic procedure proposed to consider a large data in scan for reliable patterns and deliberate connections amongst factors and after that to agree the discoveries by applying on new subsets of data by the distinguished patterns. Classification [14] is the mainly usually attached data mining system, which utilizes an plan of pre-classified cases to construct a model that can order the number of inhabitants in records on the loose. In classification strategies a model is manufactured in view of preparing data and connected to test data. WEKA is an open source data mining apparatus which incorporates usage of data mining algorithms. In this paper explain about two different tasks Compare between different Datasets with single algorithm and Comparison between different classification Algorithms with Single dataset through different factors.***

*Keywords: Data mining, classification algorithms, datasets, MAE, RMSE, RAE, RRSE.*

----------∴∴∴∴----------

## I INTRODUCTION

Data mining [1][20][21][22] is a method of quickly developing interdisciplinary field, which consolidates database management, statistics on datasets, machine learning algorithms and related regions going for removing helpful knowledge from vast accumulations of data. The data mining process comprises of three essential stages: investigation, build model or pattern definition, and validate/confirmation. In a perfect world, if the idea of accessible data permits, it is commonly rehashed iteratively until a "robust/standard" model is distinguished. In any case, in business tradition the alternatives to approve the model at the phase of examination are commonly restricted and, in this way, the underlying outcomes frequently have the status of heuristics that could impact the choice procedure.

Data mining should be possible with vast number of algorithms and strategies which include regression analysis, classification techniques, clustering techniques, and association rules, artificial intelligence, neural networks, and so forth. Basically classification [14] and clustering algorithms also known as supervised and unsupervised classification. Supervised learning means Supervision: The training data (observations, measurements, etc.) are accompanied by labels indicating the class of the observations, that new data is classified based on the training set. Unsupervised learning means the class labels of training data is unknown given a set of measurements, observations, etc. with the aim of establishing the existence of classes or clusters in the data. WEKA[26] incorporates usage of different classification algorithms [14] like "*Bayes Net classifier", "NAIVE BAYES"* [25], Meta-Ada Boost- M1, Attribute Selected Classifier, Iterative classifier optimizer, Multiclass Classifier, Randamizable Filtered classifier, Decision Table[24] using single dataset from UCI dataset repository [2].

Naive Bayes and Bayes net algorithms [4] were successfully utilized for apparatus condition checking too [3]. Naive Bayes classification algorithm [6] is a probabilistic classifier and utilizations statistical method for every classification. Bayes Net model represents probabilistic connections among an arrangement of random factors graphically. It shows the quantitative quality of the associations between factors, enabling probabilistic convictions about them to be refreshed consequently as new data that ends up noticeably accessible. This is a "*Coordinated Acyclic Graph"* (DAG) G that determines a combined likelihood conveyance, where the different nodes of graph stand for random variable and circular segment stand for relationship between variables [7]. A decision table is a straightforward structure to use "divide and conquer method to separate an composite decision making process into an accumulation of more straightforward decisions, subsequently giving an effectively interpretable arrangement [8][9][10][16]. The decision table is simple to understand and

we can get after a tree structure effortlessly to notice how the decision is ended. This is a predictive demonstrating method consumed as a module of classification, clustering and prediction assignments [8], [16].

Yoav Freund and Robert Schapire were introduced a machine learning algorithm, that is known as Ada Boost M1 (or) Adaptive Boosting. This can be developed as a part of concurrence with various dissimilar sorts of learning algorithms to improve their execution. The reschedule of the other learning algorithms ('feeble students') is joined into a weighted aggregate that represents the last yield of the supported classifier. Ada Boost [12] is adaptive as in consequent feeble students are changed for those occurrences misclassified by previous classifiers. Ada Boost is touchy to loud data and anomalies. Attribute selection classifier evaluated by independent and learning algorithm. The independent method evaluated the attributes and subsets of attributes through filters but learning algorithms are evaluated through subset of attributes through Wrappers. In the event that two items are connected, gathering something around one protest can help deductions concerning the others. We label this approach iterative classification [13].

Deductions prepared with lofty trust in beginning iterations are nourished once more interested in the data to fortify deductions about correlated questions in consequent iterations. Multi class classification means each training point fit in to one of n different classes. To predict the new data point through new data function related to which class. Class for running a subjective classifier on top of data that have been gone throughout a discretionary filter. Similar to the classifier, the structure of the filter is constructing only in light of the training data and test occurrences will be prepared by the filter without changing their structure.

Commencing to writing only can recognize to facilitate numerous classification algorithms contain and to utilized for characterizing the issues in thrusts and others turning individuals. So as to state firmly that a specific algorithm is well again to contrasted with different algorithms for relative investigation should be finished. Henceforth, this paper predominantly manages the exhibitions of different algorithms in different perspectives at the same different datasets with single algorithm.

## II BREAST CANCER DATASET

Breast cancer [5][15][27] is a standout amongst the most well-known cancers among ladies on the planet. Early identification of breast cancer is basic in lessening their existence misfortunes. Data mining is the way toward examining huge data and exactness it into helpful information disclosure along with the part of data mining come nears to developing quickly particularly classification systems are exceptionally viable approach to characterizing the data, which is basic in managerial method for medical experts. This examination at hands the distinctive data mining classifiers going on the database records of breast cancer; next to utilizing classification exactness with and with no include determination systems**.** Breast cancer dataset contains 287 rows or instances and 10 features or attributes (age, menopause, tumor-size, inv-nodes, node-caps, breast quad, deg-malig, breast, irradiat, class) and 287 instances are classified into 2 categories, the classification is done which algorithm is best based on some factors those are "*CCI" (correctly classified instances), "ICI" (incorrectly classified instances), "MAE" (mean absolute error), "RMSE" (Root mean square error), "RAE" (Relative absolute error), RRSE (Root Relative squared error).*

***Table 1 Comparison between different classification Algorithms with Single dataset through different factors***

| SNo. | Algorithm | Factors | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | CCI (instances & %) | ICI (instances & %) | Kappa Statistics | MAE | RMSE | RAE (%) | RRSE (%) | Time taken (Sec) |
| 1 | BayesNet classifier | 208 (72.47%) | 79 (27.52%) | 0.313 | 0.217 | 0.3716 | 76.774 | 99.096 | 0.03 |
| 2 | Naïve Bayes | 208 (72.43%) | 79 (27.52%) | 0.313 | 0.216 | 0.3693 | 76.244 | 98.461 | 0.02 |
| 3 | Meta-Ada Boost M1 | 200 (69.68%) | 85 (30.27%) | 0.246 | 0.297 | 0.3557 | 104.925 | 100.184 | 0.04 |
| 4 | Attribute Selected classifier | 208 (72.47%) | 79 (27.52%) | 0.313 | 0.217 | 0.3716 | 76.774 | 99.096 | 0.03 |
| 5 | Iterative classifier optimizer | 207 (72.12%) | 80 (27.87%) | 0.241 | 0.246 | 0.3605 | 87.041 | 96.128 | 0.74 |
| 6 | Multiclass Classifier | 196 (68.29%) | 91 (31.7%) | 0.177 | 0.248 | 0.3802 | 87.677 | 101.37 | 0.16 |
| 7 | Randamizable Filtered classifier | 191 (66.55%) | 96 (33.44%) | 0.22 | 0.227 | 0.4642 | 80.043 | 123.768 | 0.02 |
| 8 | Decision Table | 210 (73.17%) | 77 (26.82%) | 0.269 | 0.264 | 0.3551 | 93.391 | 94.689 | 0.12 |

The value of Kappa [18] is defined as $K = \frac{p_o - p_e}{1 - p_e}$

$$MAE = \frac{1}{n}\sum_{j=1}^{n}|y_j - \hat{y}_j| \quad RMSE = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \hat{y}_j)^2} \quad RAE = \frac{\sum_{i=1}^{N}|\hat{\theta}_i - \theta_i|}{\sum_{i=1}^{N}|\bar{\theta} - \theta_i|}$$

$$RRSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{\theta}_i - \theta_i)^2}{\sum_{i=1}^{N}(\bar{\theta} - \theta_i)^2}}$$

Above table explain about the different algorithms applied on single dataset (breast cancer dataset) identify the dataset is suitable to this breast cancer dataset. After building a number of different regression models, there is an

abundance of criteria by which they can be assessed and compared.

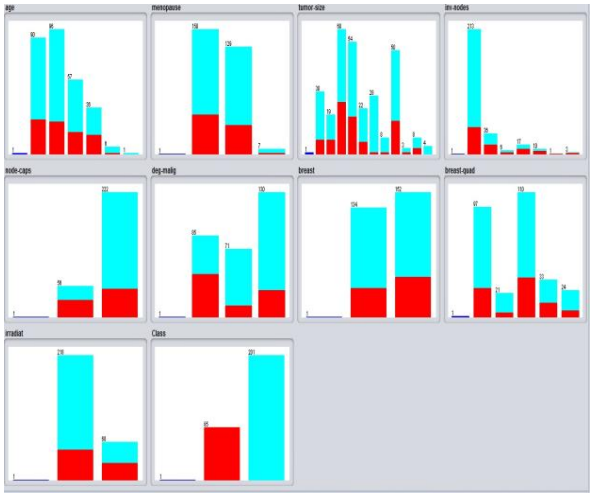

**Figure 1: Breast cancer dataset 10 attributes measurements**

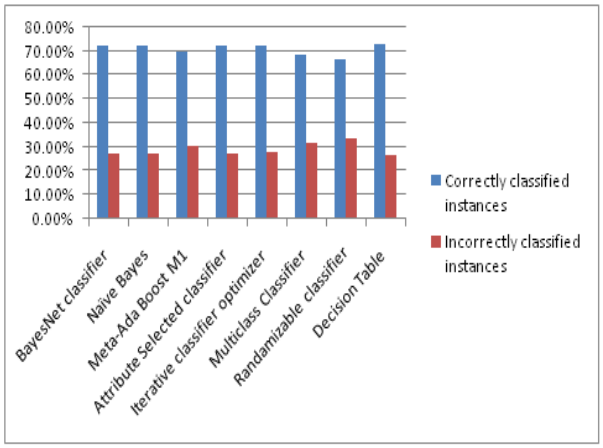


**Figure 2: Comparison between classification algorithms through CCI and ICI factors**

Above figure 1) shows the number of instances correctly classified. If more number of instances are classified correctly then we identify particular algorithm is suitable to this dataset, this is the one factor for observation.

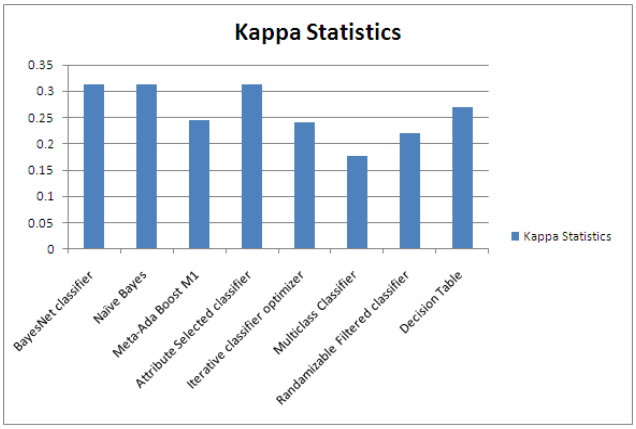


**Figure 3: Comparison between classification algorithms through Kappa statistics**

At the point while two binary variables are endeavors by two people to quantify a similar object, you can utilize Cohen's Kappa [17] (regularly basically called Kappa) as a compute of understanding between the two individuals. Kappa measures the values of data esteems in the fundamental transverse of the table and afterward changes these qualities for the measure of assertion that might be required because of chance alone.

Now is one conceivable translation of Kappa [17][18].

- Poor conformity = Less than 0.20
- Fair conformity = 0.20 to 0.40
- Moderate conformity = 0.40 to 0.60
- Good conformity = 0.60 to 0.80
- Very good conformity = 0.80 to 1.00

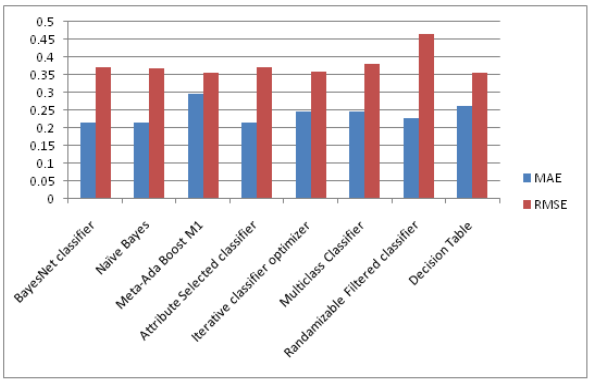


**Figure 4: Comparison between classification algorithms through MAE and RMSE**

*"Mean Absolute Error" (MAE) [19] and "Root mean squared error (RMSE")* [19] are two of the most well-known metrics used to measure accuracy for ceaseless variables. Not sure in the event that we envisioning it but rather we think there used to be a period when there were significantly more distributed MAE results. It appears that distributions we go over now generally utilize either RMSE or some version of R-squared.

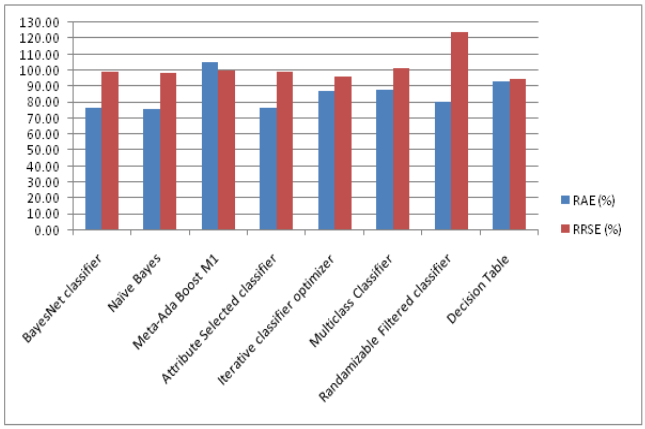


**Figure 5: Comparison between classification algorithms through RAE and RRSE**

ISSN (Online) 2456-3293

RAE [19] and RMSE [19] is a popular formula to measure the error rate of a regression display. However, it must be compared between models whose errors are measured in similar units.

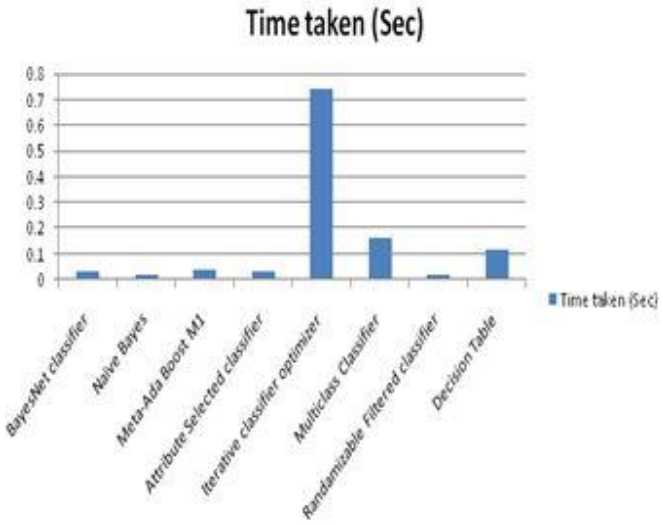


**Figure 6: Comparison between classification algorithms through Time in Seconds**

Naive Bayes and Randamizable Filtered classifier had low time complexity compare to other algorithms but Randamizable Filtered classifier contains less correctly classified instances. Based on kappa statistics and CCI factor Attribute selected classifier is a best algorithm is suitable for breast cancer dataset.

***Table 2 Comparison between different Datasets with single algorithm through different factors***

| SNo. | Dataset Name | Factors | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Data type | No. Instances | No. Attributes | CCI (instances & %) | ICI (instances & %) | Kappa Statistics | MAE | RMSE | RAE (%) | RRSE (%) | Time taken (Sec) |
| 1 | S/W Engineering Team Assessment & prediction in education setting | Sequential, Timeseries | 74 | 116 | 50 (67.567%) | 24 (32.432%) | 0.0472 | 0.3341 | 0.4928 | 89.69 | 114.67 | 0.16 |
| 2 | Diabetes | Multivariate | 768 | 9 | 576 (75%) | 192 (25%) | 0.4218 | 0.324 | 0.424 | 71.46 | 89.13 | 0.14 |
| 3 | Yeast | Multivariate | 1483 | 1 | 0 (0%) | 1483 (100%) | 0.0014 | 0.0014 | 0.026 | 100.03 | 100.037 | 0.05 |
| 4 | Iris | Multivariate | 149 | 1 | 3 (2.01%) | 146 (97.98%) | 0 | 0.013 | 0.083 | 100.33 | 10.37 | 0.001 |
| 5 | Weather | Multivariate | 14 | 5 | 6 (42.85%) | 8 (57.14%) | -0.365 | 0.45 | 0.55 | 95.99 | 113.27 | 0.02 |
| 6 | Supermarket | Multivariate | 4627 | 217 | 2948 (63.71%) | 1679 (36.28%) | 0 | 0.46 | 0.48 | 99.99 | 100 | 0.46 |
| 7 | Vehicle | Multivariate | 847 | 19 | 347 (40.96%) | 500 (59.03%) | 0.2118 | 0.2505 | 0.41 | 83.42 | 107.62 | 0.14 |
| 8 | Heart-h | Multivariate | 295 | 14 | 235 (79.66%) | 60 (20.33%) | 0.544 | 0.2063 | 0.3309 | 66.24 | 84.06 | 0.08 |

Above table explain about the different datasets applying on single algorithm is Attribute selected classifier through different factors. Each dataset contain different instances and attributes and also different data types. We contains different types of datasets are Sequential, Time series, Multivariate and different count of instances and attributes.

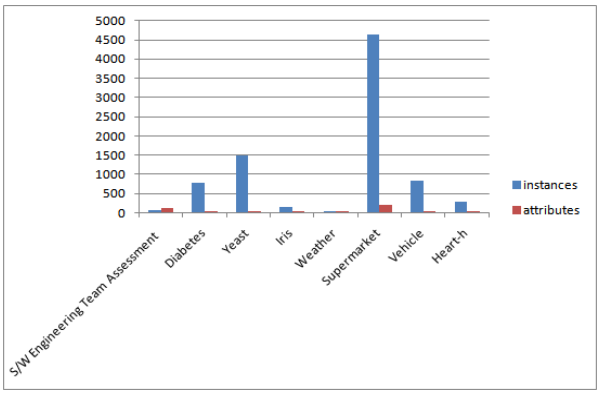


**Figure 7: Comparison between different datasets through No. of instances and attributes.**

Here simply show the number of instances in each dataset. We take number of datasets (S/W Engineering Team Assessment & prediction in education setting, Diabetes, Yeast, Iris, Weather, Supermarket, Vehicle, Heart-h) , and each dataset had some instances and attributes, here supermarket dataset had the more instances and attributes compared to other datasets.

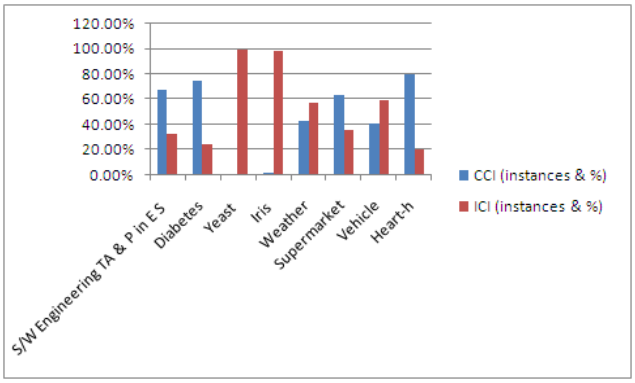


**Figure 8: Comparison between different datasets through CCI and ICI factors**

Heart-h [23] dataset had contain more correctly classified instances (CCI=77.66%) compared to other datasets especially super market dataset contains more instances but CCI is only 63.71%. but heart-h contains 235 instances and 14 attributes only.

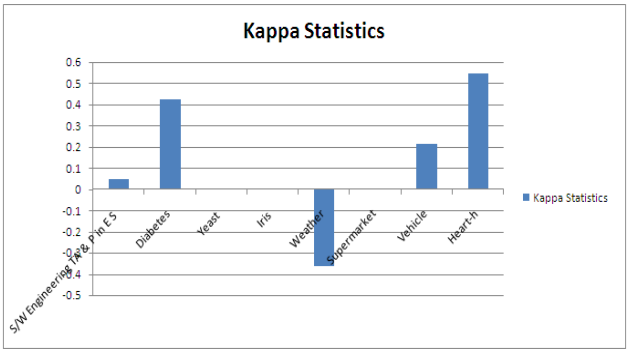


**Figure 9: Comparison between different datasets through Kappa Statistics**

Heart-h dataset contains more kappa statistics compared to other datasets. Some algorithms had –ve kappa statistics. Kappa statistics value is more that means it contains more reliable.

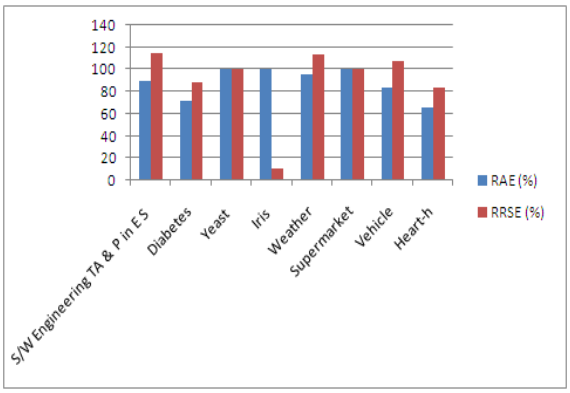


**Figure 10: Comparison between different datasets through RAE and RRSE**

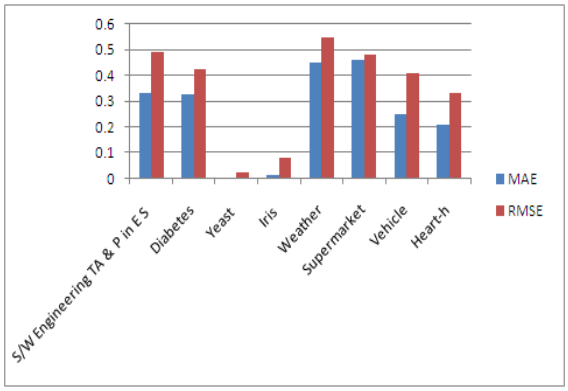


**Figure 11: Comparison between different datasets through MAE and RMSE**

We measure "Mean Absolute Error" (MAE) [19] and "Root Mean Squared Error" (RMSE) [19] on different datasets. Observation of above graph RMSE value is more to compare MAE to all the datasets Yeast (CCI=0%) and iris datasets contains very less attributes, so ignore that datasets. Compared to other datasets heart-h dataset contains maximum instances, attributes and less error rates.

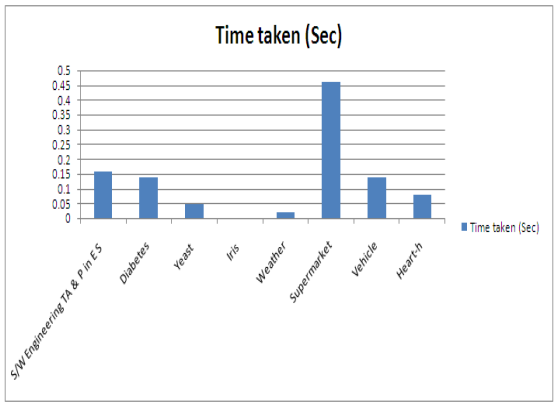


**Figure 12: Comparison between different datasets through Time in seconds**

We observe except iris and yeast datasets, weather dataset build time is less but instances are only 14 and attributes are 5, so we don't consider this dataset. Next one is heart-h got good classification with in less time compared to other datasets.

## III CONCLUSION

Super market dataset had more time complexity because it contains more instances and attributes. And another dataset is iris contains less time (0.001sec), it contains 149 instances and only one attribute and also correctly classified instances(CCI) rate is worst percentage (2.01%) so, we did not consider time complexity of iris. Another dataset is heart-h is best to compare other datasets because time complexity is less (0.08 sec) and CCI rate is high compared to ICI. And also kappa statistics also best (0.544). So finally we conclude attribute selected classifier is very suitable to heart-h dataset type is multivariate. So attribute selected classifier is very suitable to heart-h dataset.

## REFERENCES

[1] Daniel T. Larose, "Data Mining Methods and Models", John Wiley & Sons, INC Publication, Hoboken, New Jersey (2006).

[2] Xindog Wu, Vipin Kumar *et al.*, "Top 10 Algorithms in Data Mining", *Knowledge and Information Systems*, 14(1), 1-37 (2008).

[3] M. Elangovan, et al., Studies on Bayes classifier for condition monitoring of single point carbide tipped tool based on statistical and histogram features, Expert Systems with Applications 37 (2010) 2059– 2065.

[4] V. Muralidharan, et al. A comparative study of Naïve Bayes classifier and Bayes net classifier for fault diagnosis of mono block centrifugal pump using wavelet analysis, Applied Soft Computing (2012).

[5] Ahamed Lebbe Sayeth Saabith et al., Comparative Study On Different Classification Techniques For Breast Cancer Dataset, International Journal of Computer Science and Mobile Computing, October (2014).

[6] Amit Gupta, Azeem Mohammad et al., A Comparative Study of Classification Algorithms using Data Mining: Crime and Accidents in Denver City the USA, International Journal of Advanced Computer Science and Applications (2016).

[7] Bayes Nets Retrieved from http://www.bayesnets.com

[8] J.S R Jang (1993), "ANFIS Adaptive Network Based Fuzzy inference System". IEEE Transaction on Systems, Man and Cybernetics. Vol. 23, no3, pp 665-685.

[9] Quinlan, J. R. (1986), "Introduction of decision trees", Machine Learning, 1(1), pp 81-106. [17]. Quinlan, J. R. (1986), "Simplifying Decision Trees", the MIT Press.

[10] Quinlan, J. R. (1986), "Simplifying Decision Trees", the MIT Press, GIIRJ, Vol.2 (3), MARCH (2014).

[11] Hughes, G.F. (January 1968) "On the mean accuracy of statistical pattern recognizers" *IEEE Transactions on Information Theory* 14 (1):55–63 doi: 10 1109/TIT 1968 1054102.
[12] https://en.wikipedia.org/wiki/AdaBoost
[13] J. Neville and D. Jensen (2000) Iterative classification in relational data. Proceedings of the AAAI 2000 Workshop learning statistical Models from Relational Data. AAAI Press. pp. 42-49.
[14] G. D. K. Kishore, Research scholar, Krishna University, Dr. M. Babu Reddy, Assistant Professor, Krishna University "A literature survey on object classification techniques" International Journal of Advanced Technology in Engineering and Science (2017).
[15] Endo, A., Shibata, T., & Tanaka, H. (2008).Comparison of Seven Algorithms to Predict Breast Cancer Survival (< Special Issue> Contribution to 21 Century Intelligent Technologies and Bioinformatics). Biomedical fuzzy and human sciences: the official journal of the Biomedical Fuzzy Systems Association, 13(2), 11-16.
[16] S. Cben, C. F. N. Cowan, and P. M. Grant, "Orthogonal least squares learning algorithm for radial basis function networks," IEEE Trans. Neural Networks, vol. 2, no. 2, pp. 302-309, Mar. 1991.
[17] Ah Dormer, Neil Klar "The Statistical Analysis of Kappa Statistics in Multiple Samples" Elsevier Science Inc. Vol. 49, No. 9, pp. 1, 1996.
[18] Barlow W, Lai MY, Azen S. A comparison of methods for calculating a stratified kappa. Stat Med 1991; 10: 1465-1472.
[19] Ashish Kumar Dogra, Tanuj Wala A Comparative Study of Selected Classification Algorithms of Data Mining" International Journal of Computer Science and Mobile Computing, Vol.4 Issue.6, June- 2015, pg. 220-229.
[20]http://www.zentut.com/data-mining,"data-mining-techniques", retrieved on: 26 April 2015.
[21]http://www.saedsayad.com/zeror.htm, "Zero R", Retrieved on: 14 May 2015.
[22]http://www.doc.ic.ac.uk/~yg/course/ida2002/ida-2002-2.PPT, "Data Pre-processing", Retrieved on: 30 May 2015.
[23]Vijayaran S, Sudha "An Effective Classification Rule Technique for Heart Disease Prediction", International Journal of Engineering Associates (IJEA), ISSN: 2320-0804, Vol.1, Issue 4, 2013.
[24] Xindog Wu, Vipin Kumar et al., "Top 10 Algorithms in Data Mining", Knowledge and Information Systems, 14(1), 1-37 (2008).
[25]Kim Larsen, "Generalized Naive Bayes classifiers" SIGKDD Explorations 7(1), 76-81, (2005).
[26] Bharat Deshmukh, Ajay S. Patil etc," Comparison of Classification Algorithms using WEKA on Various Datasets", IJCSIT International Journal of Computer Science and Information Technology, Vol. 4, No. 2, December 2011, pp. 85-90.
[27] Ryan Potter, "Comparison of Classification Algorithms Applied to Breast Cancer Diagnosis and Prognosis", Wiley Expert Systems, 24(1), 17- 31, (2007).