WEKA Tool With Different Data Mining Technique

1 *Monika Lambsonge   2*Girish Talmale 3*Poonam Prasad

1*.2* G.H Raisoni College Of Engineering, 3* Senior Scientist CSIR-NEERI
Computer Science And Engineering   Analytical Instrumentation Department
Nagpur   Nagpur

Abstract:
Now a day's huge data is a very important issue today. Everyday users suffer major problems like "How to analyze large data or how to predict a large amount of data?" here WEKA tool gives a solution to these types of problems, we can easily solve the big data issue. Weka is an extreme engineering technology that helps to analyze big data and sort out the solution by using different algorithms. analysis of big data plays an important role for a better solution.

Index term: Data mining techniques, Weka tools, Weka Explore interface, classification, clustering
I. Introduction

In this article, a vast data is collected daily. analyze that large data and abstract meaningful information from that data. peoples perform many types of work in the world like industrial, banking, medical, business in the company, shop, mall, etc. so they have a lot of data but nobody have a time to sort important data manually there is a need to extract our important data in easy way. We find out the effective solution that is WEKA tool. We can easily sort our important data daily and make our work easy that helps to predict large data and give a better solution by using different data mining algorithms.
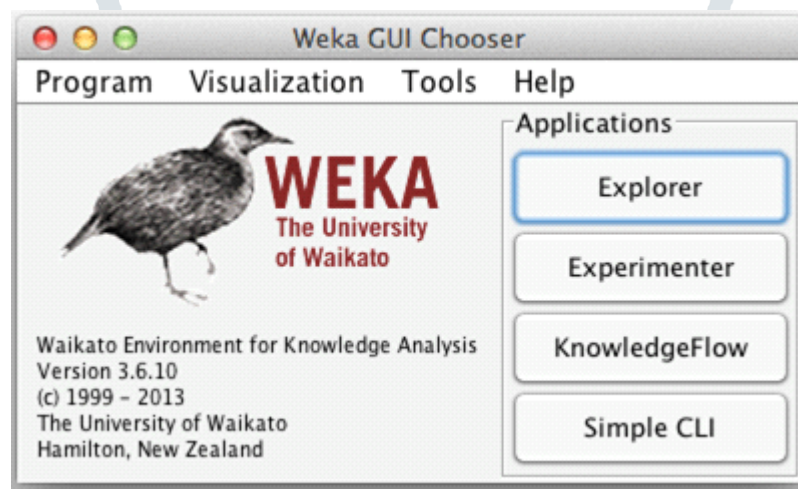
II. About Weka:

Fig. Weka GUI Chooser

Waikato Environment for Knowledge Analysis shortly denoted as WEKA developed in 1993 from the Waikato University in New Zeland, it is free open-source software fully implemented in JAVA, easily available on the internet. WEKA is mainly created to analyze several amount of big databases like industrial, banking, medical, school, colleges, etc. as well as support several data mining tasks like pre-processing, clustering, classification, regression, visualization. WEKA helps to support big data with the help of different algorithms and find out the differences between two similar variables. Graphical method support to imagine the future selection tasks.

WEKA is a open source tool developed by the Waikato university from New Zealand. This is a free software under the GNU license [6].

- WEKA is fully created in the JAVA and easily runs on a new platform.
- A extensive compilation of data pre-processing and unloading techniques.
- Ease to understand graphical user interface model.

WEKA support to collect several data like clustering, classification and feature selection algorithms.

WEKA Features
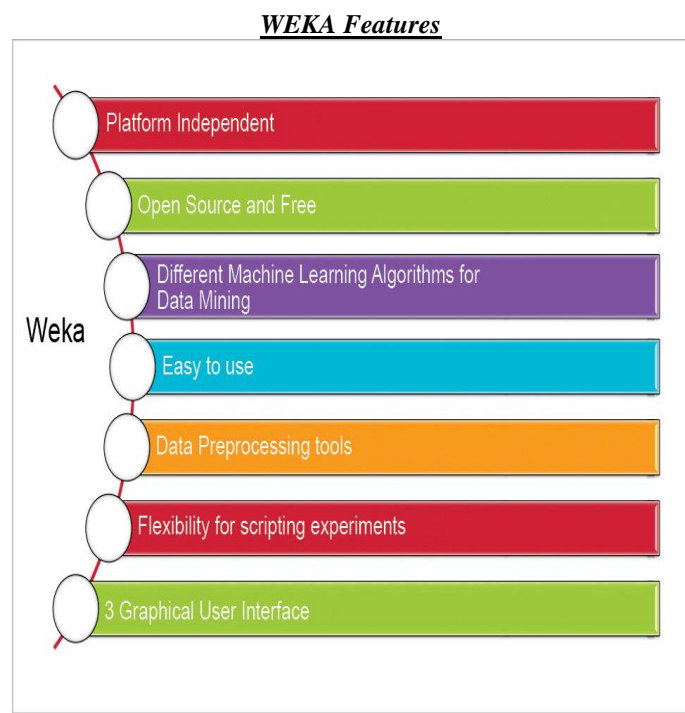
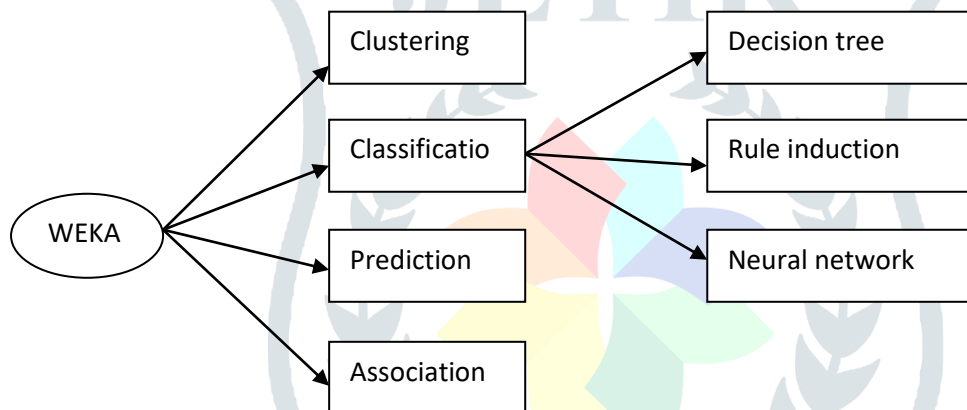## III. Classification method of weka [1]:

Fig.1. Classification method

1. Clustering [7]:

Clustering is a technique of segmentation a set of data that builds a valuable cluster of objects which have similar characteristics. Clustering defines classes and puts object in each class. For example, in the prediction of obnoxious gases by using clustering we can sort out that the list of obnoxious gas which has the same risk factor.

2. Classification [1][8]:

Classification is the method to identify the objects in a group to target classes. in the classification method mainly used mathematical techniques such as linear programming, decision trees, statistics, and neural network. by using that we can learn how to classify the group of objects.

- **A decision tree :**

In a decision tree, there are some nodes like internal, branches, leaf. The internal node is used to refer a test on attribute[15]. Branch node is used to refer the result of a test & leaf node is used to hold a class label. The upmost node is a root node in the tree.

- **Rule induction :**

Rule induction is a data mining technique of deducing if-then rules from a data set. These symbolic decision rules explain an inherent relationship between the attributes & class labels in the data set. But, in general, these rules hold & are deduced from real-life experience based on our everyday observation.

- **Neural Network:**

Neural networks and data mining. An artificial neural network, often just called a neural network is a mathematical model inspired by biological neural networks [14]. A neural network consists of an interconnected group of artificial neurons and it processes information using a connectionist approach to computation.

3. Prediction [1]:
Prediction is the method to identify the relations between two variables.

4. Association [1]:
Association is the method to find out the relation between different modules that are present in the same database.
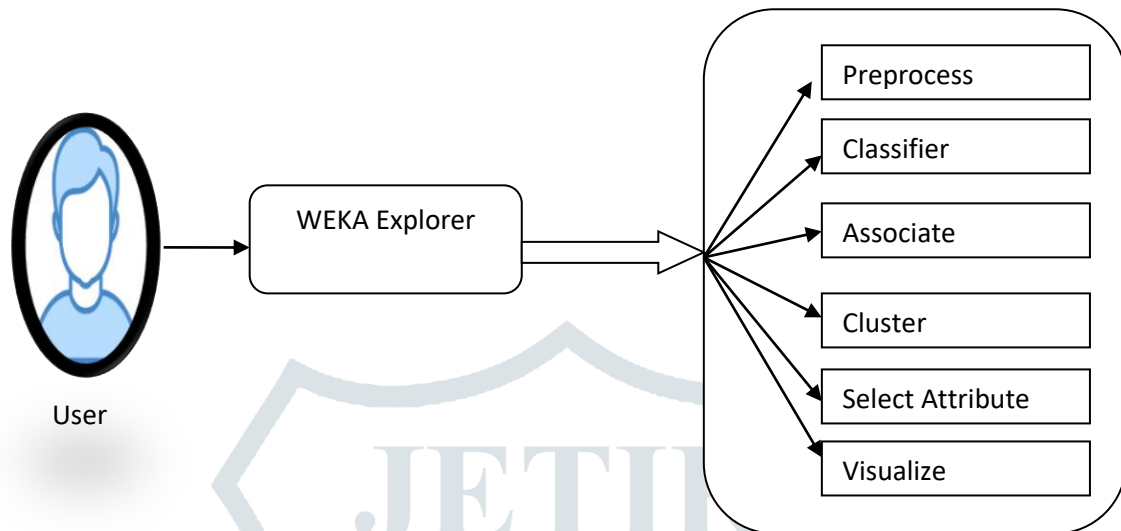
IV. USER INTERFACE

Fig.2. Weka Explorer

When we open the WEKA tool, there are five applications in WEKA GUI chooser. Our main work in explorer, because explorer is a user interface, we can upload our dataset and explore it by using different algorithms. The selection of algorithm apply according to our dataset for getting the better result.

1. PREPROCESS
In pre-processor, there has a facility for upload data from a database, comma-separated values (CSV) file, etc. and pre-processing this data using filtering algorithms use for filtering the database according to their specific attribute and instances.

- **FILTER**

  There are many filter present in Weka used for clean up the unwanted dataset like attributes and instances which can be easily removed by the use of filter.
  In filter two different packages i.e supervised and unsupervised.

- **SUPERVISED**
  Contains a supervised filter that is the filter that takes class distribution into account must utilize for eg. the weka.filters.supervised filter interface.

  Attribute:-column-wise filtering.
  Instances: - row-wise filtering.

- **UNSUPERVISED**
  Contains unsupervised filters that are filter work without taking any class distribution method into account, must utilize the weka. filters.unsupervised filter interface.

  Attribute:-filter column-wise.
  Instances:-filter row-wise.

2. CLASSIFIER
In classifier, there has a facility to apply classification and regression algorithms for eg. predictive model, visualize erroneous prediction, receiver operating characteristic curves, visualization etc.

3. ASSOCIATE
In associate, there has a facility to provide association rule learners to identify the interrelationship between the attribute in data.

4. CLUSTER
In a cluster there has a facility of clustering technique for eg. simple K-means.

5. SELECT ATTRIBUTE
In select attribute, there has a facility to provide the most predictive attribute.

6. VISUALIZE
In visualize there has a facility to show scatter plot matrix, and analyze it using several selection operators.
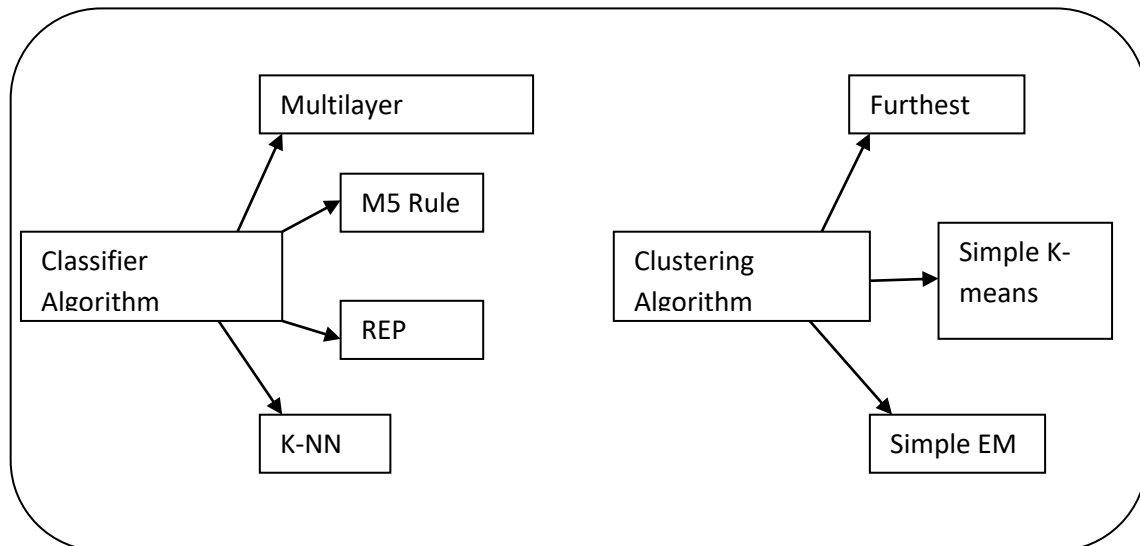
V. CLASSIFIRE AND CLUSTERING Technique

Fig.3. Classifier and clustering Algorithm

1. **Classifier Algorithm:**
   Classifier algorithm there are four different algorithm techniques.

1.1) MULTILAYER PERCEPTRON
Multilayer perceptron [10] is classified as feed forward ANN[15] leave out for the input node. MLP solve a non-linear optimization problem as well as use a back propagation algorithm method which belongs to supervised learning.
1.2) M5 RULES
M5 is a plain multilayer perceptron model which belongs to the artificial neural network.
1.3) REP TREE
REP tree algorithm is a fast decision tree which is based on the C4.5 algorithm [12][10]. It can able to produce classification and regression algorithms however random feature method generates multiple classifiers as well as random decision tree method generate combines bagging.
1.4) K-NN
K nearest neighbors (K-NN) [8] is a simple algorithm. K is used to denote the number of nearest neighbors. this algorithm help to calculate the distance between the query - instances [13] based on Kth minimum distance.
1. **Clustering algorithm** :
In clustering there are three different clustering techniques.
2.1) FARTHEST FIRST
Farthest first is a fast and greedy algorithm based on clustering. In this algorithm, k points are firstly select as a center and second point select as a greedy for selecting the farthest first center and remaining points are cluster whose center is closest. Clustering is the method of a group of similar objects.
2.2) SIMPLE K-MEANS
Simple k-means is the method of clustering [11]. Which is used in data mining techniques. Here K is the number of clustering points which is defined center[15] however this algorithm uses the Euclidean distance to measure the distance between instances and clusters [10].
2.3) SIMPLE EM
Expectation-Maximization is an iterative aspect method to find MLE or MAP evaluation of data distribution when data is partially missing or hidden. The EM iteration alternates between the performance of one expectation E-step as well as one maximization M-step [9].

IV. Literature survey on WEKA tool:

The author Kulwinder Kaur et al [1] studied the data mining technique with WEKA tool. in this paper author claim an data mining technique has affectively used been in many fields including large data sets such as banking, medical, business, etc. in this paper author represents the several data mining technique in WEKA such as clustering, classification, association, prediction.

The author R.Ramya et al [2] studied the plant disease detection by the used of WEKA tool. in this paper author research on agricultural field. they have represents several different algorithms such as k-means, decision tree, multilayer perceptron, Naive Bayes, SUM, RFA for predicting disease using WEKA tool.

The author Sarangam Kodati et al [3] analysis the heart disease by using the data mining methods in WEKA and orange tool. In this paper, the author represents various data mining methods such as SVM, NB, RF, KNN algorithm for analysis of the heart disease patients.

The author Rausheen Bal et al [4] comparatively studied about metadata with the classifiers algorithm by used the techniques of WEKA tool. In this article, the author represents different classifiers algorithm technique.

The author Dr.Anil Sharma et al [5] having comparatively study about data mining techniques, tools and parameters. In this paper author represents different data mining methods to the analysis of a huge database. These tools greatly support in business database. In this paper authors did experiment with two different database on WEKA tool.

Table 1. Studied research paper on WEKA

Sr.no	Title name	Author name	country	year	references	Method used
1.	The Further Development of Weka Base on Positive and Negative Association Rules	Y. Shen, J. Liu and J. Shen	China	2010	https://ieeexplore.ieee.org/document/5523114	positive and negative association rules
2.	Web-Weka Meets D3.js in Web Based Medical Data Mining	B. Liu, Y. Peng, Y. Zou, J. Wang and T. Jiang	China	2015	https://ieeexplore.ieee.org/document/7469109	D3, Visualisation
3.	Predicting Burn Patient Survivability Using Decision Tree In WEKA Environment	B. M. Patil, D. Toshniwal and R. C. Joshi	India	2009	https://ieeexplore.ieee.org/document/4809213	machine learning algorithm c4.5
4.	A Comparable Study Employing WEKA Clustering/Classification Algorithms for Web Page Classification	I. Charalampopoulos and I. Anagnostopoulos	Greece	2011	https://ieeexplore.ieee.org/document/6065094	Classification
5.	Comparative analysis of XLMiner and WEKA for pattern classification	C. V. Subbulakshmi, S. N. Deepa and N. Malathi	India	2012	https://ieeexplore.ieee.org/document/6320821	classification
6.	Selection of most relevant input parameters using WEKA for artificial neural network based concrete compressive strength prediction model	S. Saad, M. Ishtiyaque and H. Malik	India	2016	https://ieeexplore.ieee.org/document/8077368	Artificial neural network
7.	Review analyzer analysis of product reviews on WEKA classifiers	A. A. Kshirsagar and P. A. Deshkar	India	2015	https://ieeexplore.ieee.org/document/7193034	Classify
8.	Comparing the partitional and density based clustering algorithms by using WEKA tool	G. Jenitha and V. Vennila	India	2014	https://ieeexplore.ieee.org/document/6966310	Various clustering algorithm

9.	C. A. Fowler and R. J. Hammel	Converting PCAPs into Weka mineable data	USA	2014	https://ieeexplore.ieee.org/document/6888681	Several different algorithm
10.	A. Koliopoulos, P. Yiapanis, F. Tekiner, G. Nenadic and J. Keane	A Parallel Distributed Weka Framework for Big Data Mining Using Spark	USA	2016	https://ieeexplore.ieee.org/document/7207196	Weka with spark
11.	M. Ramzan	Comparing and evaluating the performance of WEKA classifiers on critical diseases	India	2016	https://ieeexplore.ieee.org/document/7975309	Data mining classification algorithm
12.	L. Dan, L. Lihua and Z. Zhaoxin	Research of Text Categorization on WEKA	China	2013	https://ieeexplore.ieee.org/document/6455773	SVM classification method
13.	V. Mhetre and M. Nagar	Classification based data mining algorithms to predict slow, average and fast learners in educational system using WEKA	India	2017	https://ieeexplore.ieee.org/document/8282735	Classification ,prediction
14.	R. Duriqi, V. Raca and B. Cico	Comparative analysis of classification algorithms on three different datasets using WEKA	Montenegro	2016	https://ieeexplore.ieee.org/document/7525775	Classification algorithm
15.	S. More and R. Kalkundri	Evaluation of deceptive mails using filtering & WEKA	India	2015	https://ieeexplore.ieee.org/document/7193262	Classification algorithm
16.	B. Ziolko, S. Manandhar, R. C. Wilson and M. Ziolko	Logitboost weka classifier speech segmentation	Germany	2008	https://ieeexplore.ieee.org/document/4607680	Classifier
17.	N. Bhargava, S. Dayma, A. Kumar and P. Singh	An approach for classification using simple CART algorithm in WEKA	India	2017	https://ieeexplore.ieee.org/document/7855983	CART algorithm
18.	J. H. Hayes, W. Li and M. Rahimi	Weka meets TraceLab: Toward convenient classification: Machine learning for requirements engineering problems: A position paper	Sweden	2014	https://ieeexplore.ieee.org/document/6894850	Machine learning algorithm
19	S. Katla, D. Xu, Y. Wu, Q. Pan and X. Wu	DPWeka: Achieving Differential Privacy in WEKA	USA	2017	https://ieeexplore.ieee.org/document/8166630	Data mining with DPweka
20	J. Khalfallah and J. Ben Hadj Slama	A Comparative Study of the Various Clustering Algorithms in E-Learning Systems Using Weka Tools	Tunisia	2018	https://ieeexplore.ieee.org/document/8726188	Clustering with E-learning
21	S. Jain, R. Gupta and R. K. Dwivedi	Generating Patterns from Pizza Ontology using Protégé and Weka Tool	India	2018	https://ieeexplore.ieee.org/document/8746935	Protégé and weka tool
22	F. Rodriguez-Teja and E. Grampin	WekaTIE, a traffic classification plugin integrating TIE and Weka	Cyprus	2014	https://ieeexplore.ieee.org/document/6906428	Different classification algorithm

VII. CONCLUSION:

In this article, we studied the different datasets that are analyzed in WEKA through the use of different data mining techniques and algorithms. WEKA is easily handled all types of huge data like medical, banking, business, etc. that help to predict huge data and sort important information related to your dataset. WEKA included various algorithms which is important to analyze large dataset.

References:

[1] Kulwinder Kaur1 and693 Shivani Dhiman" Review of Data Mining with Weka Tool",
JCSE International journal of computer science and engineering,2016,india,volume 4, issue-8,S-ISSN 2347-2

[2] R.Ramya1, Dr.P.Kumar2, D.Mugilan3, M.Babykala4"" a review of different classification techniques in machine learning using weka for plant disease detection" International Research Journal of Engineering and Technology (IRJET), 2018, *Tamilnadu, India*, Volume: 05 Issue: 05, e-ISSN: 2395-0056, p-ISSN: 2395-0072

[3] Sarangam Kodati & Dr. R. Vivekanandam," Analysis of Heart Disease using in Data Mining Tools Orange and Weka", Global Journals,Volume 18, Issue 1, Version 1.0, Year 2018,India, ISSN: 0975-4172 & Print ISSN: 0975-4350

[4] 1Rausheen Bal, 2Sangeeta Sharma" Review on Meta Classification Algorithms using WEKA", *International Journal of Computer Trends and Technology (IJCTT) – Volume 35 Number 1 - May 2016,India,* ISSN: 2231-2803

[5] Dr. Anil Sharma, Balrajpreet Kaur,"a research review on comparative analysis of data mining tools, techniques and parameters", International Journal of Advanced Research in Computer Science, Volume 8, No. 7, July – August 2017,Punjab,India, ISSN No. 0976-5697

[6] Richa Agrawal SOIT, RGPV Bhopal M.P, Jitendra Agrawal SOIT, RGPV Bhopal M.P" Analysis of Clustering Algorithm of Weka Tool on Air Pollution Dataset", *International Journal of Computer Applications (0975 – 8887) Volume 168 – No.13, June 2017*

[7] Christina Jayakumaran , Komathy Karuppanan," Pattern Identification using Rough Set Clustering for Spatio-Temporal Dataset", Department of Computer Science and Engineering Easwari Engineering College,Chennai, India, 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 978-1-4673-6217-7/13/$31.00_c 2013 IEEE

[8] P. Soucy and G. W. Mineau, "A simple KNN algorithm for text categorization," *Proceedings 2001 IEEE International Conference on Data Mining*, San Jose, CA, USA, 2001, pp. 647-648doi: 10.1109/ICDM.2001.989592

[9] F. Rodriguez-Teja and E. Grampin, "WekaTIE, a traffic classification plugin integrating TIE and Weka," *2014 International Wireless Communications and Mobile Computing Conference (IWCMC)*, Nicosia, 2014, pp. 623-628.doi: 10.1109/IWCMC.2014.6906428

[10] S. Shanmuganathan, R. Ibrahim and S. H. B. Bakori, "A data mining approach to analysing airborne wood particulate concentration and atmospheric data," *Proceedings of the 6th International Conference on Information Technology and Multimedia*, Putrajaya, 2014, pp. 238-244.doi: 10.1109/ICIMU.2014.7066637

[11] Richa Agrawal SOIT, RGPV Bhopal M.P, Jitendra Agrawal SOIT, RGPV Bhopal M.P," Analysis of Clustering Algorithm of Weka Tool on Air Pollution Dataset" *International Journal of Computer Applications (0975 – 8887) Volume 168 – No.13, June 2017*

[12] B. M. Patil, D. Toshniwal and R. C. Joshi, "Predicting Burn Patient Survivability Using Decision Tree In WEKA Environment," *2009 IEEE International Advance Computing Conference*, Patiala, 2009, pp. 1353-1356. doi: 10.1109/IADCC.2009.4809213

[13] Sarangam Kodati & Dr. R. Vivekanandam, Sri Satya Sai University," Analysis of Heart Disease using in Data Mining Tools Orange and Weka" Publisher: Global Journals, Online ISSN: 0975-4172 & Print ISSN: 0975-4350

[14] Dr. Anil Sharma, Balrajpreet Kaur, Department of Computer Science and Applications, LPU, Phagwara 144411, Punjab," a research review on comparative analysis of data mining tools, techniques and parameters", ISSN No. 0976-5697

[15] R.Ramya1, Dr.P.Kumar2, D.Mugilan3, M.Babykala4 *Assistant Professor1, 3,4, Department of ECE, K.S.Rangasamy College of Technology Professor2, Department of ECE, K.S.Rangasamy College of Technology, Tiruchengode Tamilnadu, India,*
International Research Journal of Engineering and Technology (IRJET), e-ISSN: 2395-0056, p-ISSN: 2395-0072