COMPARISON OF STATISTIC PREDICTION RESULTS IN WEKA EXPLORER INTERFACE AND EXPERIMENTER ENVIRONMENT INTERFACE ON DATASET

[1]Thamer Khalil Esmeel, [2]Roslina Abd Hamid, [3]Rahmah Mokhtar
[1,2,3] Faculty of Computing. Universiti Malaysia Pahang. Malaysia.

Abstract: With the increased interest into data mining as an important tool for data processing and analysis, the researchers are concerned into data mining for real decision making, data mining helps in the organizational decision making, inaccurate information can mislead decision-makers and cause costly errors. With more data collected for analytical purposes. Techniques data mining through Weka Explorer interface and experimental environment interface into determining the prediction and accuracy using different algorithm ratings to know the performance of best. Study confirm is to categorize data and help users mining useful data and easily identify an appropriate algorithm for an accurate predictive model, to access the best-performing algorithms, minimize errors and minimum time to build models through the Explorer interface and Experimental Environment Interface to get accurate.

IndexTerms **- Data mining techniques, Weka tools, Weka Explore interface, Weka Experimenter Environment interface.**

I. INTRODUCTION

Data mining involves searching for certain patterns and facts about the structure of data within large complex datasets [1]. Data mining can discover important relationships which can improve health, business processes, and many other specializations, mining patterns of hidden and strategic knowledge from big datasets that are stored electronically, and the challenges faced by many organizations. Data mining scope is harmless mining of implicit data that was unknown before and which may be useful from data warehouses. Machine learning uses statistical techniques and visualization to discover information and present it easily understood by humans. Data mining methods in Weka tools through Explorer Interfaces and Knowledge Flow Interfaces and Experimenter Interfaces, to validate their approach, using the dengue dataset within 108 cases, but Weka tools used 99 rows, 18 attributes for determining disease prediction and accuracy of their use. Classifications from different algorithms for better performance and data classification, help users mining useful information from data and easy diagnostic an algorithm suitable for its accurate predictive model, from the results, conclude that Naive Bayes, J48 is the best algorithms in performing for rated accuracy because the maximum of Detective accuracy = 100% With 99 correctly categorized cases, maximum ROC = 1, it meant at least absolute error and took minimal time to build this model through the results of Explorer Interfaces and Knowledge Flow Interfaces, the model has been by used [2]. The approach is to perform thirteen workbooks using cross-checking (N-fold) available in the Weka tool for machine learning and obtained an accuracy of better than 96.28% to a detection of malware, which was more than the upper detection accuracy (95.9%), in these upper five classifiers (FT, LMT, J48, Random forest, and NBT), the approach got the accuracy of detection 97.95% by used (Random forest), the approach has been used by Sharma & Sahay, [3].

The polynomial logistic regression is best rated with the highest accuracy in each binning level for both SLECR and SLDC followed by a decision tree, the ability to accurately predict rejections with low ranges to predict the lead time period, usually the performance of the works used such as (K-nearest neighbors, multinomial logistic regression, decision tree, support vector machine) are better when bidirectional categorical variables and Naive Bayes work best when categorical variables are converted into ordinal values, and the results will greatly benefit different parties in the supply chain by providing improved Vision and Vision Predictability, the approach has been used by Hathikal, [4]. The Boosted tree and Random forests gave logical results for analysis, Random Forests generated 100 trees, and Boosted Tree generated 200 trees to give one result on based Bagging learning, each algorithm ranked the variables on based the most important, the best four ranked variables had been selected for further analysis bagging technique used to confirm variables on based the most popular instances, final variables were selected from (data mining, Machine Productivity, Pigment Fastness, and Pile Weight), multiple regression method was applied to predict the equation on based the textile quality score, Before applying linear regression, many algorithms such as artificial (neural network and multivariate adaptive regression) were applied to predict the equation, but these algorithms did not yield good, the approach has been used by Saad, [5].

There are different data mining techniques which are usually employed in Weka tools as enumerated below:

- *Naive Bayes algorithm:* is a selective classifier calculates the probability set by calculating the combination and succession of values in the dataset. It is assumed all that variables that contribute to the classification are independent of others. The Bayes naive workbook is on based Bayes theory and total probability theory, the algorithm has been used by Bhagyashree et al.,[6].
- *J48 algorithm:* Optimal enforcement of C4.5 is called. The output by J48 is (Decision Tree). A (Decision Tree) is the same structure tree that contains different nodes, such as the root node, middle node, and the leaf node, each node in the tree has a resolution lead to decision leads to a result. A (Decision Tree) divides the data set entry area into reciprocal spaces, where each region contains a label, value, or procedure to describe or clarify its data points. The partitioning criterion in (Decision Tree) is used to calculate which attribute is best for dividing that part tree of training data that reaches a particular, the algorithm has been used by Kiranmai & Laxmi, [7]

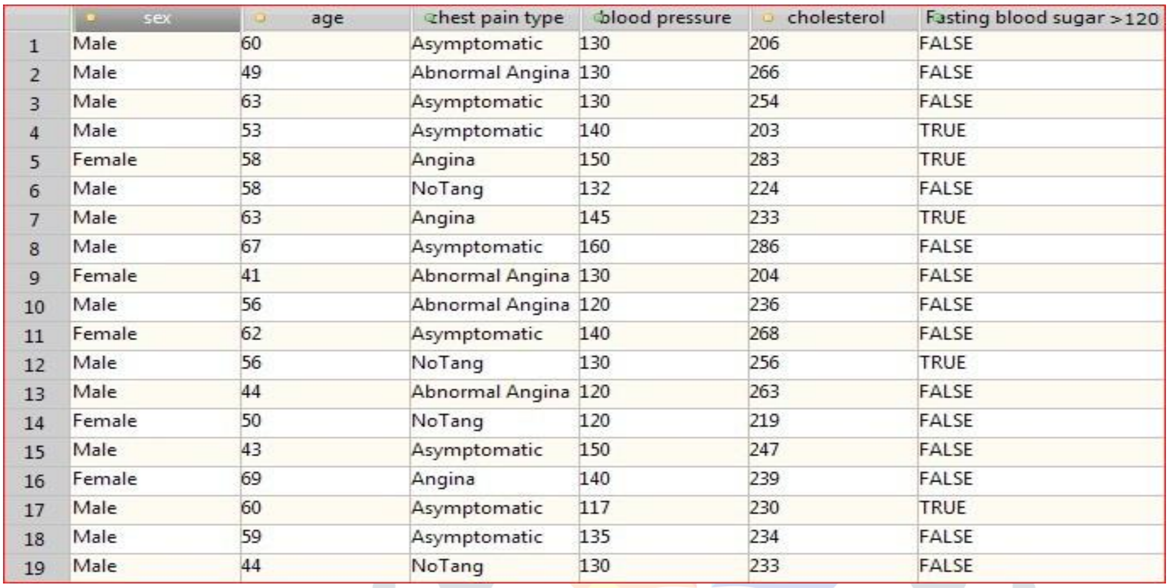
- *Multilayer Perceptron algorithm (Neural network):* A single-layer perceptron can only classify linear separable problems. For inseparable problems, it is necessary to use more layers. The multilayered network (forward feed) contains more than hidden layer whose nerve cells are called hidden neurons, the algorithm has been used by Amin & Habib, [8]

The data mining scope is a major recognition of the gathering of the data big amounts and stores it easily across computer systems [9]. The propose has been in the comparison of statistic prediction result by Weka Explorer and Experimenter Environment interface by data mining in Weka tools on the dataset.

II. METHODS

There are three main phases in this paper, which are associated with the three research objectives, respectively consists of the analysis phase, techniques phase, and results evaluating phase and for every phase has research objective. In this study, the work on data collection to chosen suitable the dataset and analyses it, where the dataset used one of the heart patients files downloaded from one the website in the formula Attribute-Relation File Format (ARFF) And changes its format to Comma-Separated values (CSV) to use it in the research experience shows as Table 3.1

Table 3.1 Dataset File Format (CSV)



	sex	age	chest pain type	blood pressure	cholesterol	Fasting blood sugar >120
1	Male	60	Asymptomatic	130	206	FALSE
2	Male	49	Abnormal Angina	130	266	FALSE
3	Male	63	Asymptomatic	130	254	FALSE
4	Male	53	Asymptomatic	140	203	TRUE
5	Female	58	Angina	150	283	TRUE
6	Male	58	NoTang	132	224	FALSE
7	Male	63	Angina	145	233	TRUE
8	Male	67	Asymptomatic	160	286	FALSE
9	Female	41	Abnormal Angina	130	204	FALSE
10	Male	56	Abnormal Angina	120	236	FALSE
11	Female	62	Asymptomatic	140	268	FALSE
12	Male	56	NoTang	130	256	TRUE
13	Male	44	Abnormal Angina	120	263	FALSE
14	Female	50	NoTang	120	219	FALSE
15	Male	43	Asymptomatic	150	247	FALSE
16	Female	69	Angina	140	239	FALSE
17	Male	60	Asymptomatic	117	230	TRUE
18	Male	59	Asymptomatic	135	234	FALSE
19	Male	44	NoTang	130	233	FALSE

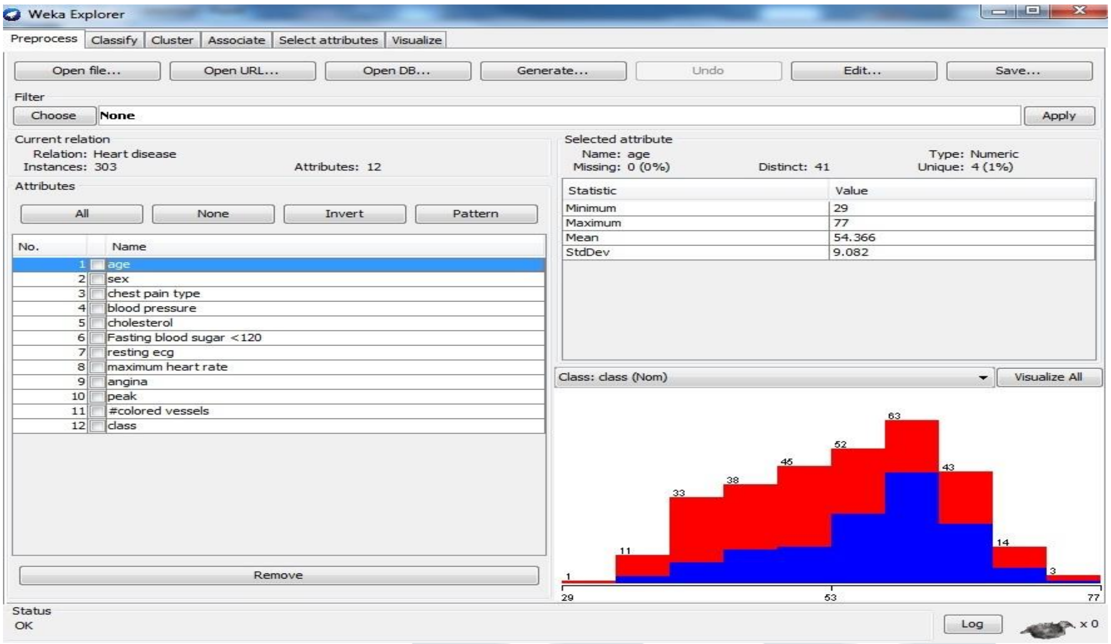
Moreover, the comparison of prediction result by Weka Explorer and Experimenter Environment interface by data mining in weka tools on the dataset, and knowledge discovery and prediction by Weka tools algorithms which are (J48, Naive Bayes, and Neural Network) were presented. This was achieved by applying data mining in Weka tools on the dataset and get the results:

- Step1: Start.
- Step2: Input the dataset.
- Step3: Choose the Weka Explorer interface.
- Step4: Go to Pre-process panel to choose the dataset.
- Step5: Go to Classification panel to apply three algorithms in Weka tools (J48, Naive Bayes, and Neural Network) on the dataset.
- Step6: Save the first prediction.
- Step7: Choose the Weka Experiment Environment interface.
- Step8: Choose a dataset.
- Step9: Go to the Run panel to choose three algorithms in Weka tools (J48, Naive Bayes, and Neural Network) to apply on the dataset.
- Step10: Save the second prediction.
- Step11: Compare the first prediction with the second prediction to get the most accurate result of statistical prediction.
- Step12: End.

III. RESULTS

3.1Weka Explorer Interface

The study comprised of non-financial companies listed at KSE-100 Index and 30 actively traded companies is selected on the bases of market capitalization. And 2015 is taken as the base year for KSE-100 index. The explorer interface has several panels like preprocess, classify, cluster, associate, select attribute and visualize. But in this study, the focus is on the Preprocess Panel and Classification Panel as shown in Figure 3.1.



Filter: None

Current relation
Relation: Heart disease
Instances: 303 Attributes: 12

Selected attribute
Name: age Type: Numeric
Missing: 0 (0%) Distinct: 41 Unique: 4 (1%)

Statistic	Value
Minimum	29
Maximum	77
Mean	54.366
StdDev	9.082

Attributes

No.	Name
1	age
2	sex
3	chest pain type
4	blood pressure
5	cholesterol
6	Fasting blood sugar <120
7	resting ecg
8	maximum heart rate
9	angina
10	peak
11	#colored vessels
12	class

Class: class (Nom)

Status
OK

Figure 3.1 Screenshot view of Heart Disease Dataset in Weka Explorer interface

JETIR

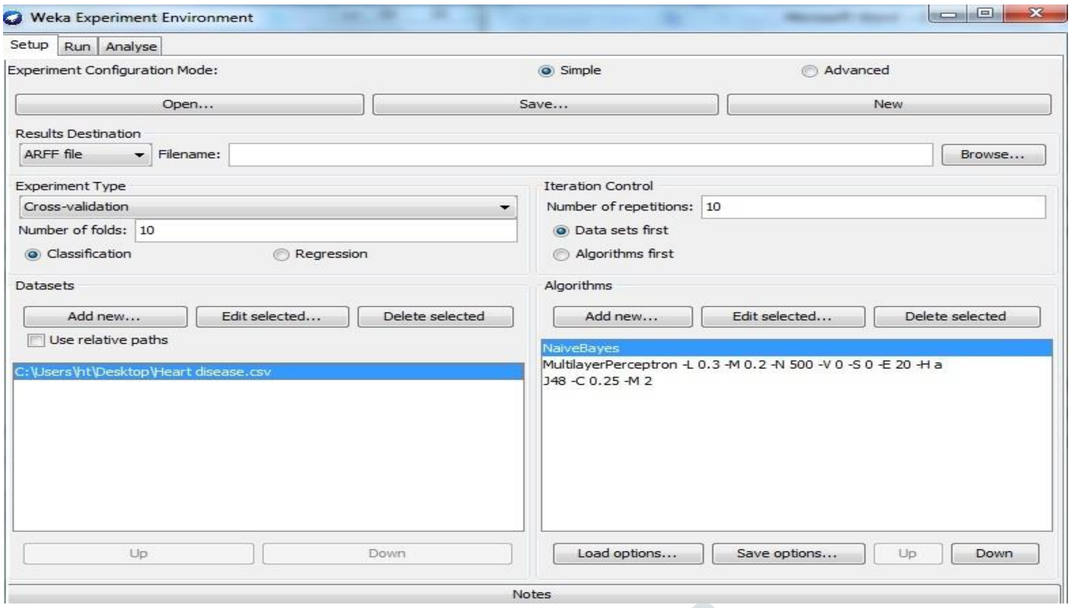
- *Naive Bayes algorithm:* After running this algorithm by used cross-validation 10 folds, achieved a classification accuracy of 82.1782% for 249 correctly classified instances from the total number 303 instances, mean absolute error rates achieved is 0.2117, time is taken for building model is 0.05 seconds and ROC area is 0.887.
- *J48 algorithm:* After running this algorithm by used cross-validation 10 folds, achieved a classification accuracy of 75.9076% for 230 correctly classified instances from the total number 303 instances, mean absolute error rates achieved is 0.2831, time is taken for building model is 0.02 seconds and ROC area is 0.747.
- *MultilayerPerceptron (Neural Network) algorithm:* After running this algorithm by used cross-validation 10 folds, achieved a classification accuracy of 81.8482% for 248 correctly classified instances from the total number 303 instances, mean absolute error rates achieved is 0.1966, time is taken for building model is 1.89 seconds and ROC area is 0.876.

Table 3.1 Statistic prediction results in Weka tools Explorer

Algorithms	correctly classified accuracy	mean absolute error	Time is taken to build this model	ROC area
Naive Bayes	82.1782%	0.2117	0.05	0.887
J48	75.9076%	0.2831	0.02	0.747
MultilayerPerceptron (Neural Network)	81.8482%	0.1966	1.89	0.876

3.2 Weka Experimenter Environment interface:

The Experiment Environment allows us to performs statistical tests on the different performance measures to allow us to conclude the experiment. For example, which algorithm evaluated in the experiment had the best performance, and what is the rank of algorithms by performance. This is useful to know algorithms that performed the best on the problem as shown in Figure 3.2. Shows from Figure 3.3 and Table 3.2 that the Naïve Bayes delivered the highest accuracy of 82.32%, Neural Network came in the second also the accuracy of 77.74% and J48 came in third also with the accuracy of 76.79% for respectively.



Experiment Configuration Mode: ● Simple ○ Advanced

Results Destination

ARFF file Filename:

Experiment Type

Cross-validation

Number of folds: 10

● Classification ○ Regression

Iteration Control

Number of repetitions: 10

● Data sets first

○ Algorithms first

Datasets

☐ Use relative paths

C:\Users\ht\Desktop\Heart disease.csv

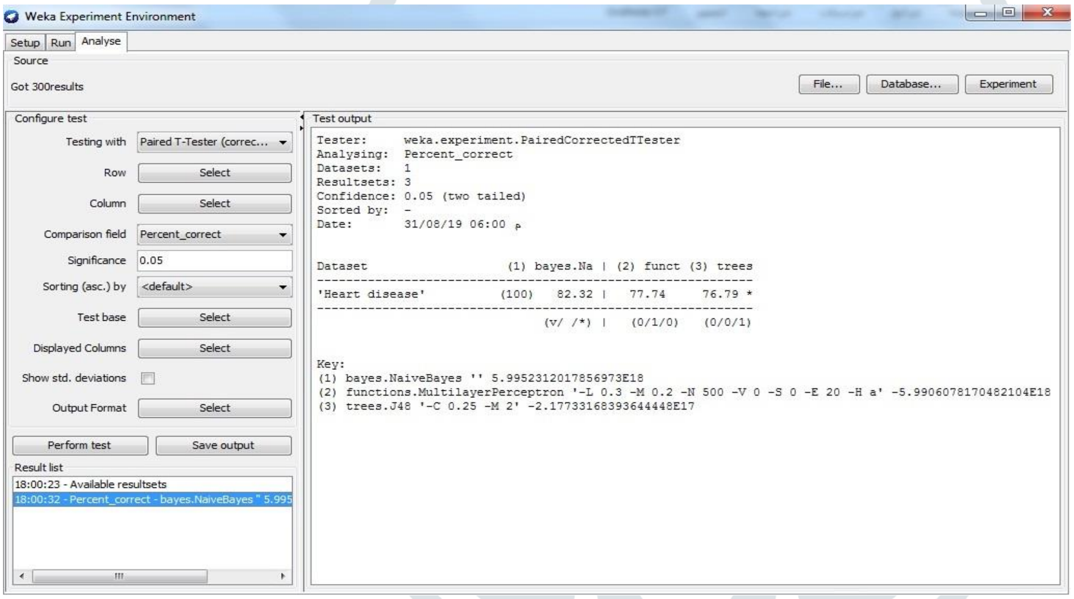
Algorithms

NaiveBayes

MultilayerPerceptron -L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a

J48 -C 0.25 -M 2

Figure 3.2 Screenshot view of Heart Disease Dataset in Weka Experimenter Environment Interface



Source

Got 300results

Configure test

Testing with: Paired T-Tester (correc...

Comparison field: Percent_correct

Significance: 0.05

Sorting (asc.) by: <default>

Show std. deviations ☐

Result list

18:00:23 - Available resultsets

18:00:32 - Percent_correct - bayes.NaiveBayes '' 5.995

Test output

```
Tester:     weka.experiment.PairedCorrectedTTester
Analysing:  Percent_correct
Datasets:   1
Resultsets: 3
Confidence: 0.05 (two tailed)
Sorted by:  -
Date:       31/08/19 06:00 م


Dataset                   (1) bayes.Na | (2) funct (3) trees
------------------------------------------------------------
'Heart disease'          (100)   82.32 |   77.74     76.79 *
------------------------------------------------------------
                               (v/ /*) |   (0/1/0)   (0/0/1)


Key:
(1) bayes.NaiveBayes '' 5.995231201785697 3E18
(2) functions.MultilayerPerceptron '-L 0.3 -M 0.2 -N 500 -V 0 -S 0 -E 20 -H a' -5.9906078170482104E18
(3) trees.J48 '-C 0.25 -M 2' -2.1773316839364444 8E17
```

Figure 3.3 Screenshot view of Heart Disease Dataset in Weka Experimenter Environment Interface

Table 3.2 Statistic prediction results in Weka Experimenter Environment

Dataset	Naive Bayes	J48	MultilayerPerceptron (Neural Network)
Heart Disease	82.32%	76.79%	77.74%

3.3 Comparison of Statistic Prediction Results

Table 3.3 below shows the comparison of Weka explorer interface and experimental environment interface

Table 3.3 Comparison of Statistic prediction Results in Weka interface

Interface	Naive Bayes	J48	MultilayerPerceptron (Neural Network)
Explorer	82.1782%	75.9076%	81.8482%
Experimenter Environment	82.32%	76.79%	77.74%

IV. DISCUSSIONS

In this study, the steps of data mining techniques in Weka tools on the dataset were carried out and explained it. Data mining technique in Weka tools on the dataset by three algorithms (Naive Bayes, J48, Neural Network) was succeeded in the best prediction statistic results.

V. CONCLUSION

There are three different phases in the methodology they are so-called phase one, phase two and phase three. Phase one is linked with the analysis phase, phase two is linked with the combination techniques phase, and phase three is linked with the results evaluation phase. Phase one focuses on the analysis phase which involves conducting the literature review and exploring the existing research works related to these techniques, the analysis in the first phase provides commonly used techniques as listed in this thesis. In this phase, the suitable dataset and determining the suitable techniques and understanding framework in the research phase were determining. Phase two emphasized on the proposed techniques (data mining techniques in Weka tools). Phase three focused on evaluating comparison the results of predicting in data mining in Weka tools by Weka interface, which was satisfying.

VI. Acknowledgement

We wish to acknowledge the Ministry of Higher Education and Scientific Research, University of Mosul, College of Arts, Iraq. and members of the staff of the Faculty of Computing (Fkom), who helped me in many ways.

References

[1] Muhammad, A. M. 2016. Advances in Clustering based on Inter-Cluster Mapping. Western Sydney University (Australia). 69(7):45-67

[2] Shakil, K. A., Anis, S., & Alam, M. 2015. Dengue disease prediction using weka data mining tool. arXiv preprint arXiv:1502.05167.

[3] Sharma, A., and Sahay, S. K. 2016. An effective approach for classification of advanced malware with high accuracy. arXiv preprint arXiv:1606.06897.

[4] Hathikal, S. 2018. Prediction of Ocean Import Shipment Lead Time for Freight Forwarder Using Machine Learning Techniques. ProQuest Dissertations and Theses, 74-74.

[5] Saad, H. 2018. An Integrated Framework of Data Mining and Process Mining to Characterize Quality and Production Processes. State University of New York at Binghamton. 123-168

[6] Bhagyashree, S. I. R., Nagaraj, K., Prince, M., Fall, C. H., and Krishna, M. 2018. Diagnosis of Dementia by Machine learning methods in Epidemiological studies: a pilot exploratory study from south India. Social psychiatry and psychiatric epidemiology, 53(1): 77-86.

[7] Kiranmai, S. A., and Laxmi, A. J. 2018. Data mining for classification of power quality problems using WEKA and the effect of attributes on classification accuracy. Protection and Control of Modern Power Systems, 3(1), 29

[8] Amin, M. N., & Habib, M. A. 2015. Comparison of different classification techniques using WEKA for hematological data. American Journal of Engineering Research, 4(3):55-61.

[9] Abbas, O., Mustafa, M. E., and Ibrahim, S. B. 2015. The Role of Data Mining in Information Security. International Journal of Computer (IJC), 17(1):1-20.

.