International Research Journal in Global Engineering and Sciences. (IRJGES)
Vol. 1, No. 1, May, 2016 | ISSN : 2456-172X

Effective Customer Patterns Analysis Using Open Source Weka Data Mining Tool

[1]N.Suresh Kumar, [2]Dr. M. Thangamani
[1]Research Scholar, Anna University, Chennai
[2]Assistant Professor, Kongu Engineering College, India

Abstract: Data mining playing vital information in extracting useful information from large amount of data set. Apriori algorithm generate useful rule by finding frequent itemset from huge data set. In this paper can apply the Apriori Algorithm to generate rules for the given data set (bank) using Waikato Environment for Knowledge Analysis tool. Bank dataset is taken from UCI machine learning repository. These articles explore and visualize the apriori technique in data mining concept and analysis the customer good patterns to take the decision to give loan for customers.

Keywords: Data mining, Apriori algorithm, Waikato Environment

1. INTRODUCTION

The data mining represents mining the knowledge from large data. Topics such as knowledge discovery, query language, decision tree induction, classification and prediction, cluster analysis, and how to mine the Web are functions of data mining. Manual analyses are time consuming in the real world. In this situation, Waikato Environment for Knowledge Analysis tool (WEKA) can use for automating the task.

Weka is a collection of machine learning algorithms for data mining tasks. Classification was performed using WEKA in data mining research. WEKA is a data mining workbench that allows comparison between many different machine learning algorithms. In addition, it also has functionality for feature selection, data pre-processing and data visualization [1]. The algorithms can either be applied directly to a dataset or called from Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules and visualization. Well-suited for developing new machine learning schemes. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well-suited for developing new machine learning schemes.

2. RELATED WORK

The more associations between accident factors and accident severity were illustrated when applying Apriori algorithm [2]. The predictive Apriori algorithm could derive more number of rules that could be useful when studying the effect of each individual factor to accident severity. These results can help the decision makers in the traffic accident department to take actions based on various hidden patterns from the data. The swarm based techniques to extract association rules for student performance prediction as a multi-objective classification problem is analysis by [3]. In this algorithm takes a low convergence time and it used a few number of parameters. Honeybee Colony Optimization and Particle Swarm Optimization are the

two used metaheuristics to extract association rules. These are used in this investigation and WEKA, Rapidminer and KEEL tools are used for comparing the technique. Various type of analysis is carried out using association rules [4-6] in data mining through WEKA environments. The Weka tool is used for disease prediction. Data mining is a well known technique used by health organizations for classification of diseases such as dengue, diabetes and cancer in bioinformatics research [7].

3. EXPERIMENTS DESIGN

Implementation of Association Rule Mining is carried out in bank datasets using Weka tool.

3.1 Dataset description

Association rule works only with nominal type and the data values are discrete in nature.
Number of Instances: 600
Number of Attributes: 12

3.2 Attributes description

Table.1 shows the list of attributes in bridge dataset. It also represents the data type for each attributes. Fig.1 and Fig.2 show the front panel of the Weka and location of the bank-data.csv file respectively. Bank datasets attributes are viewed by viewer in the WEKA explorer panel. It is illustrated in Fig. 3.

Table.1 List of attributes

Attribute	Description	Data type
id	a unique identification number	Nominal
age	age of customer in years	**Numeric**
sex	MALE / FEMALE	Nominal
region	inner_city/rural/suburban/town	Nominal
income	income of customer	**Numeric**
married	is the customer married (YES/NO)	Nominal
children	number of children	**Numeric**
car	does the customer own a car (YES/NO)	Nominal
save_acct	does the customer have a saving account (YES/NO)	Nominal
current_acct	does the customer have a current account (YES/NO)	Nominal
mortgage	does the customer have a mortgage (YES/NO)	Nominal
pep	did the customer buy a PEP (Personal Equity Plan) after the last mailing (YES/NO)	Nominal

Fig.1 Weka GUI Chooser

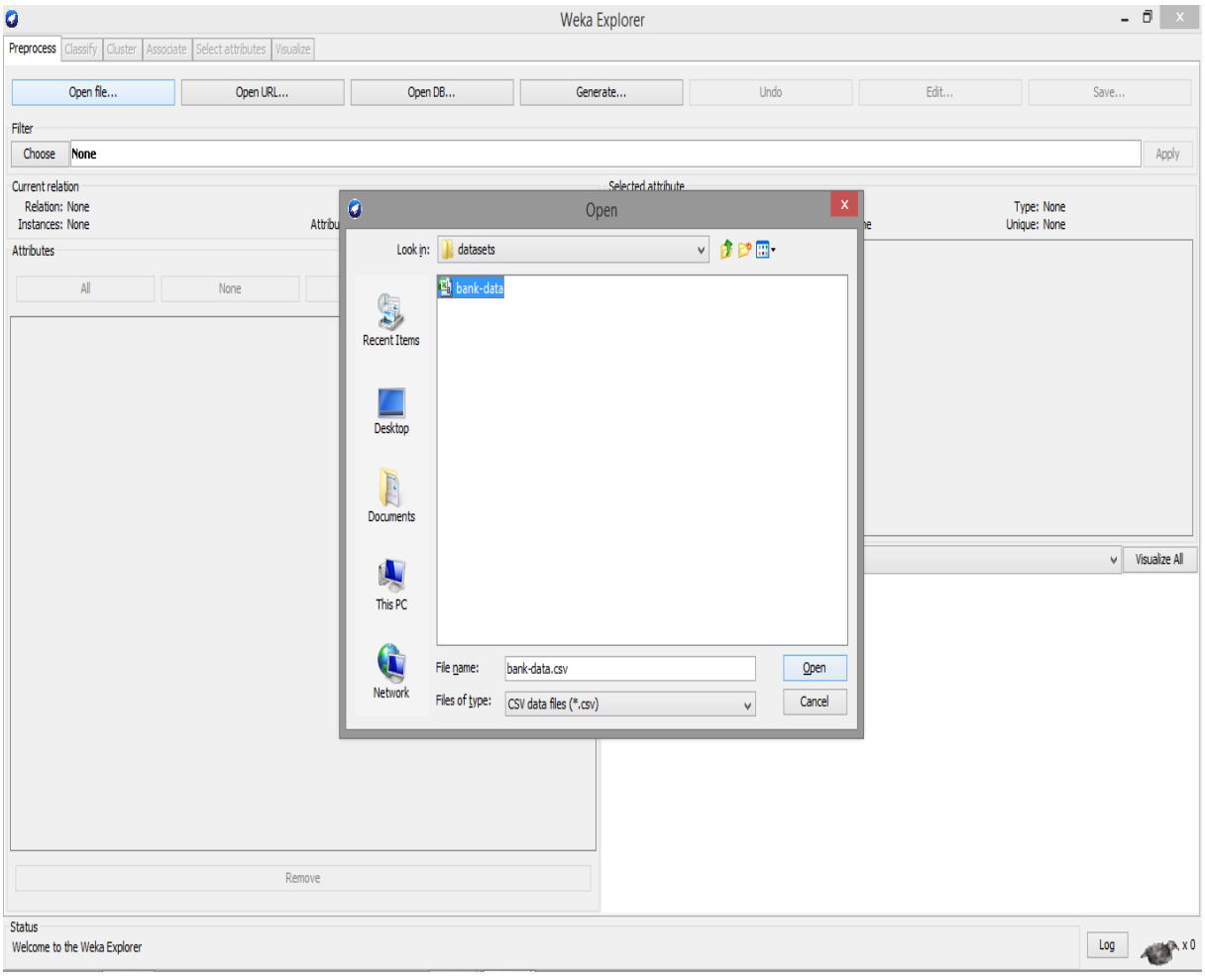


Fig.2 bank-data.csv file

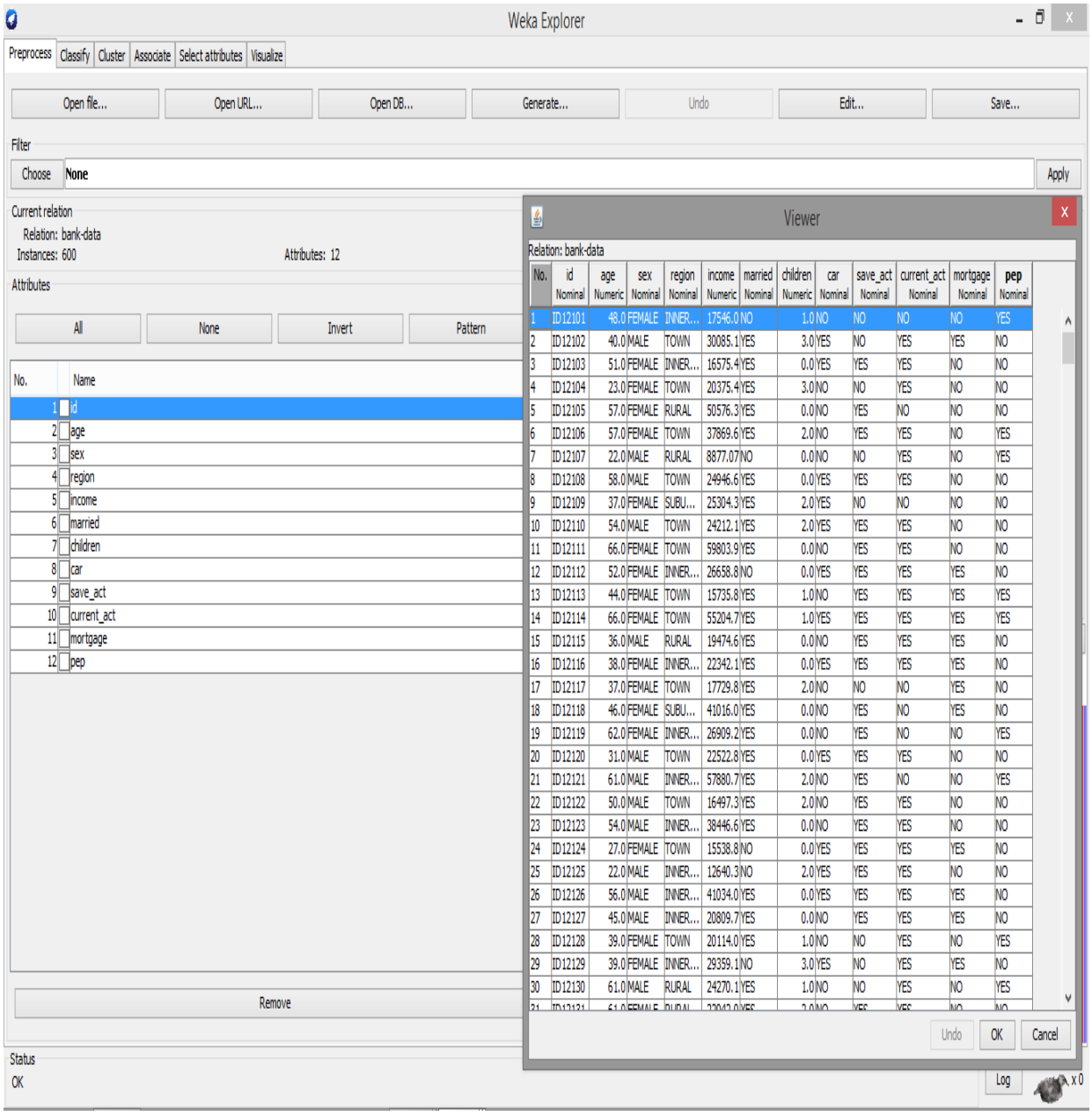


Fig.3 Weka Database Viewer and front panel

4. IMPLEMENTATION STEPS

Since Apriori algorithm works with only nominal data, the data set is preprocessed. Save the intermediate files after each step. The preprocessing WEKA is shown in Fig.4 and Fig.5. The Fig.6 represents the pure data after preprocessing.

The following preprocessing methods are applied:

- **Removing the attribute:**

- Remove the attribute id, since it uniquely identifies the tuples. It is done by selecting the remove attribute filter.
- Remove the attribute location, since it does not play a vital role in generating the rules.

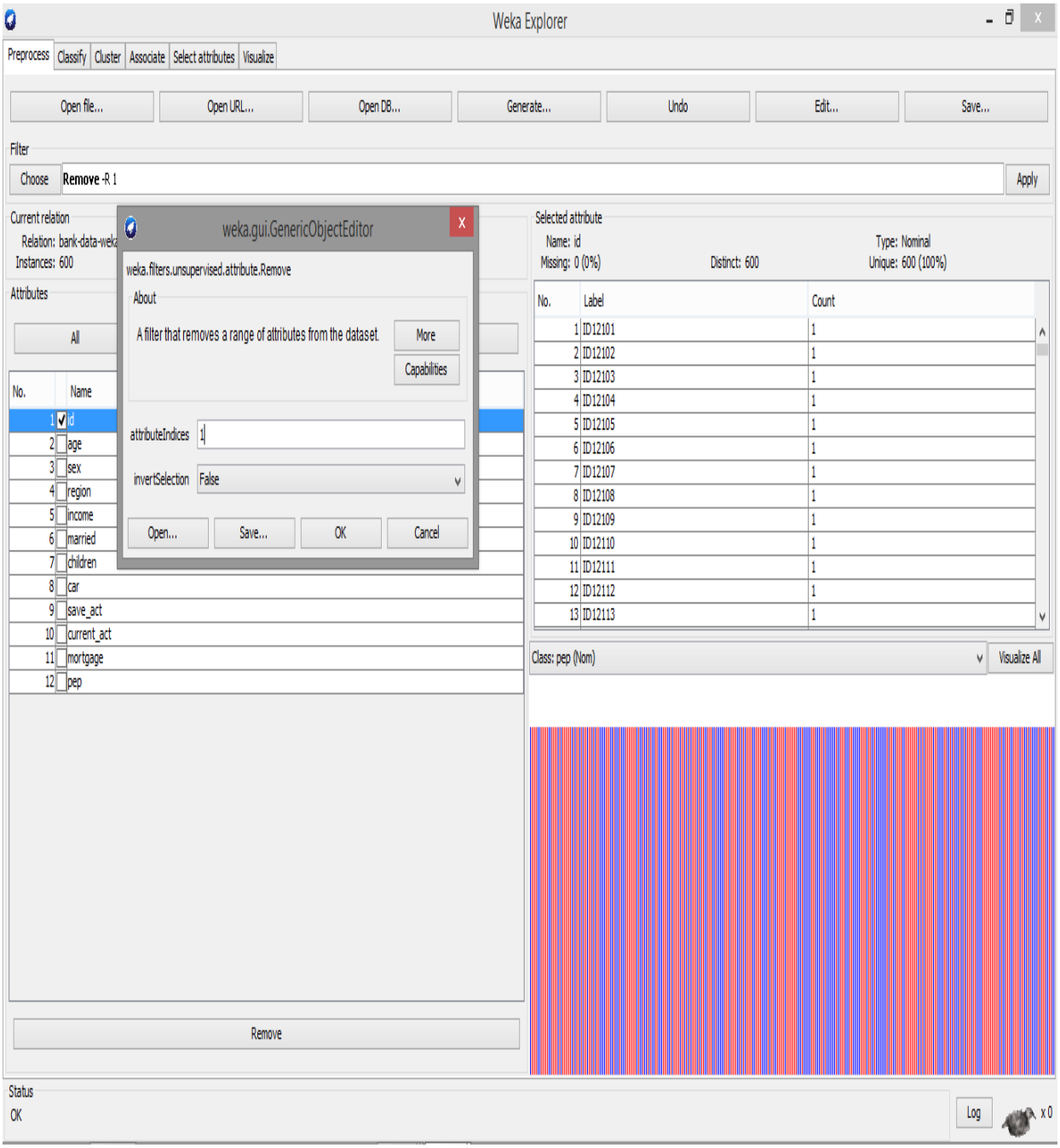


Fig.4 Preprocessing Weka

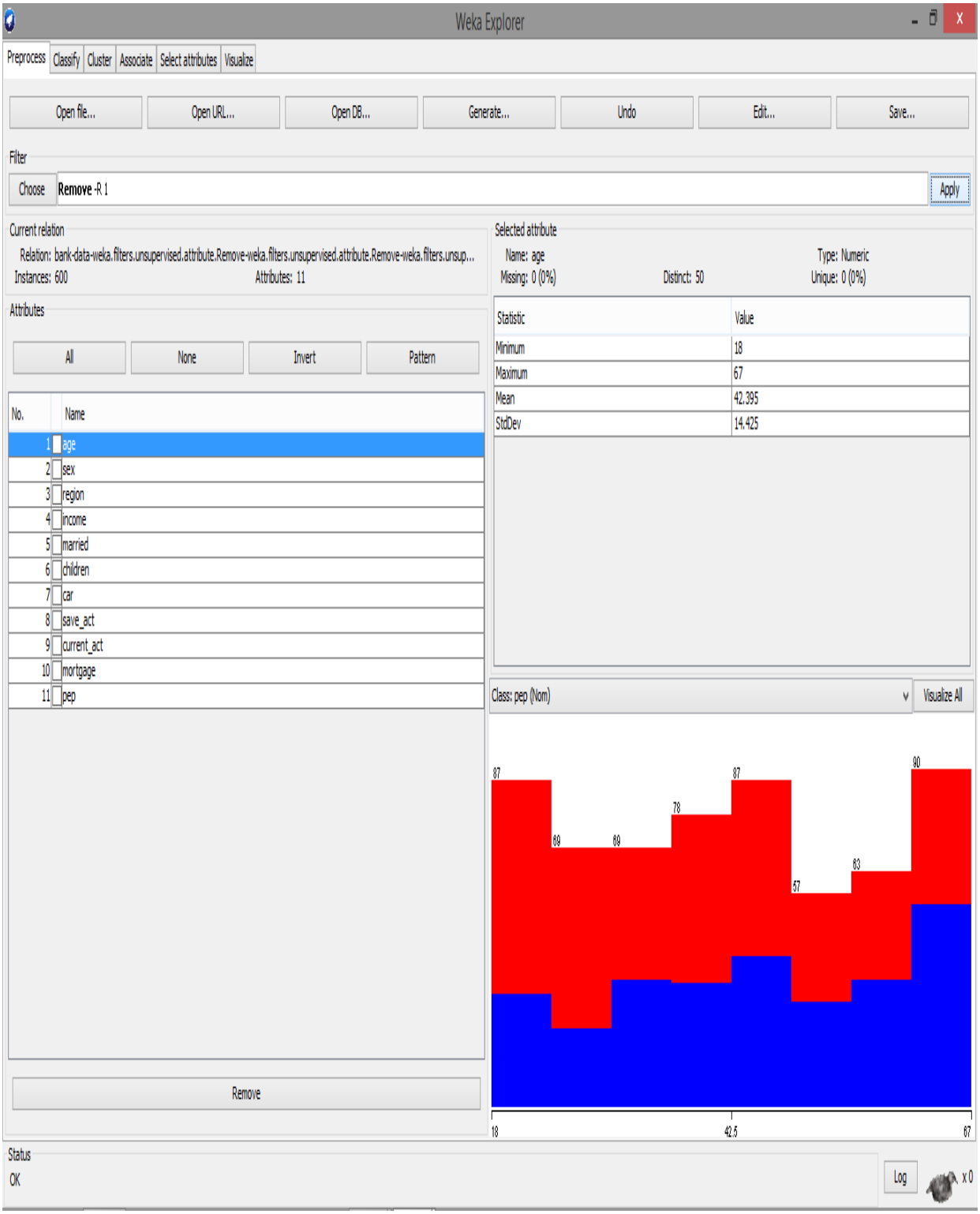


Fig.5 Unwanted attribute removing in Preprocessing Weka

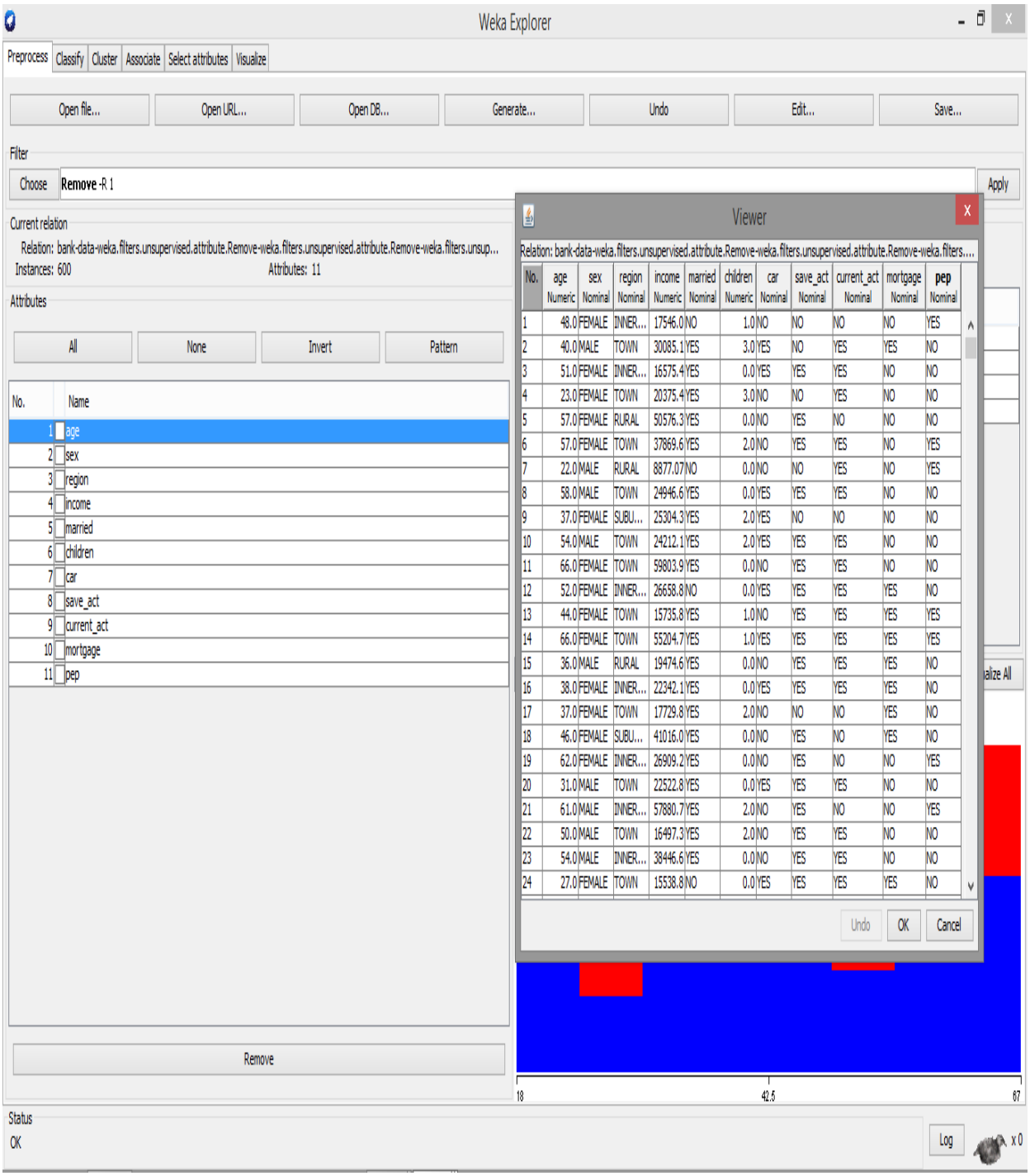


Fig.6 After preprocessing

Discretization: Association rule mining can be applied on categorical data, so the three numeric attributes erected, length and lanes in the data set are discretized and it shown in Fig.7. The Fig.8 represents the how to modify the normalized value for discretization.

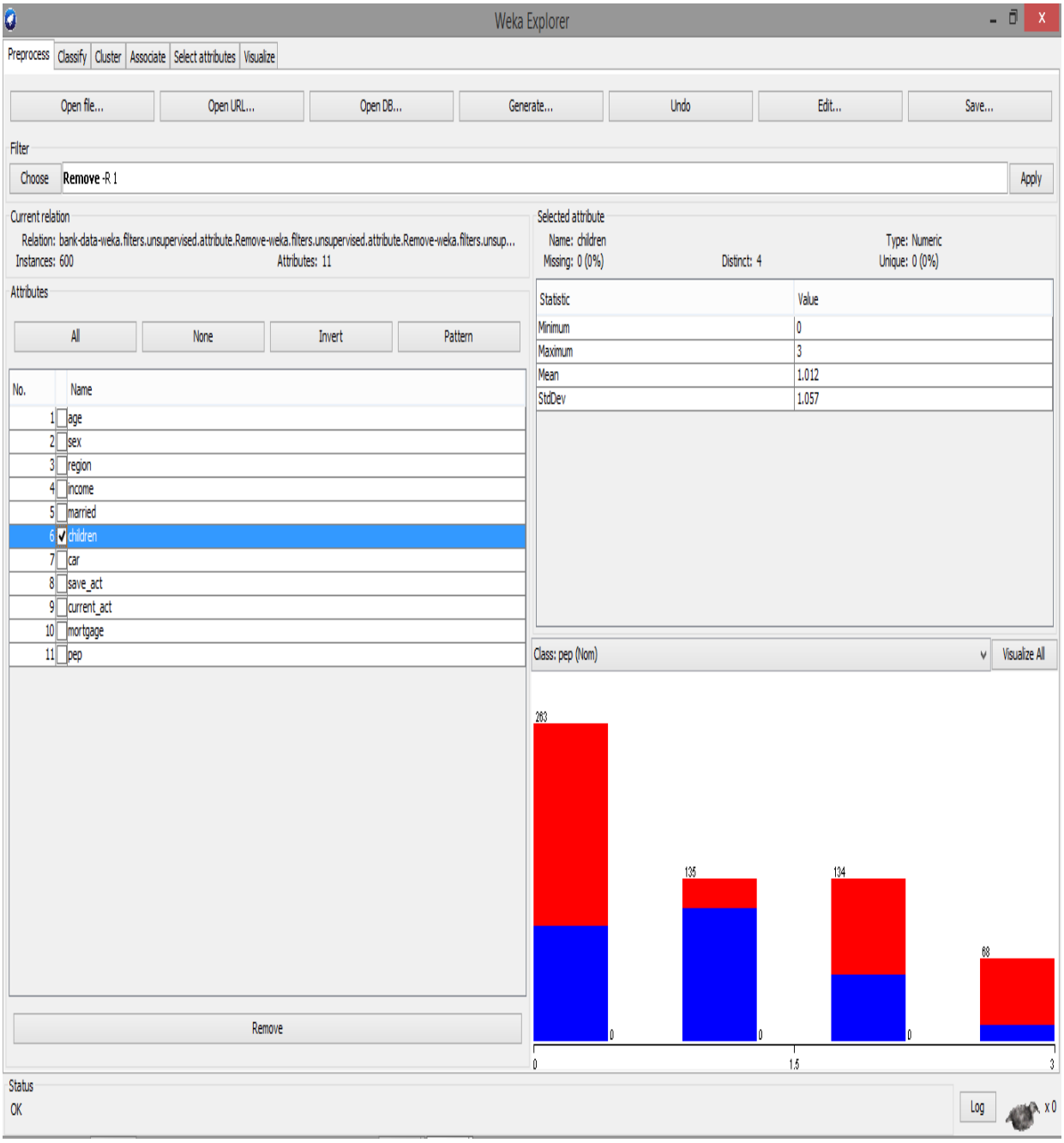


Fig.7 Discretization in customer datasets

The attribute children has only the values 0, 1, 2 and 3, so it is discretized by just removing the keyword numeric from the input file and replacing it with set of discrete values.

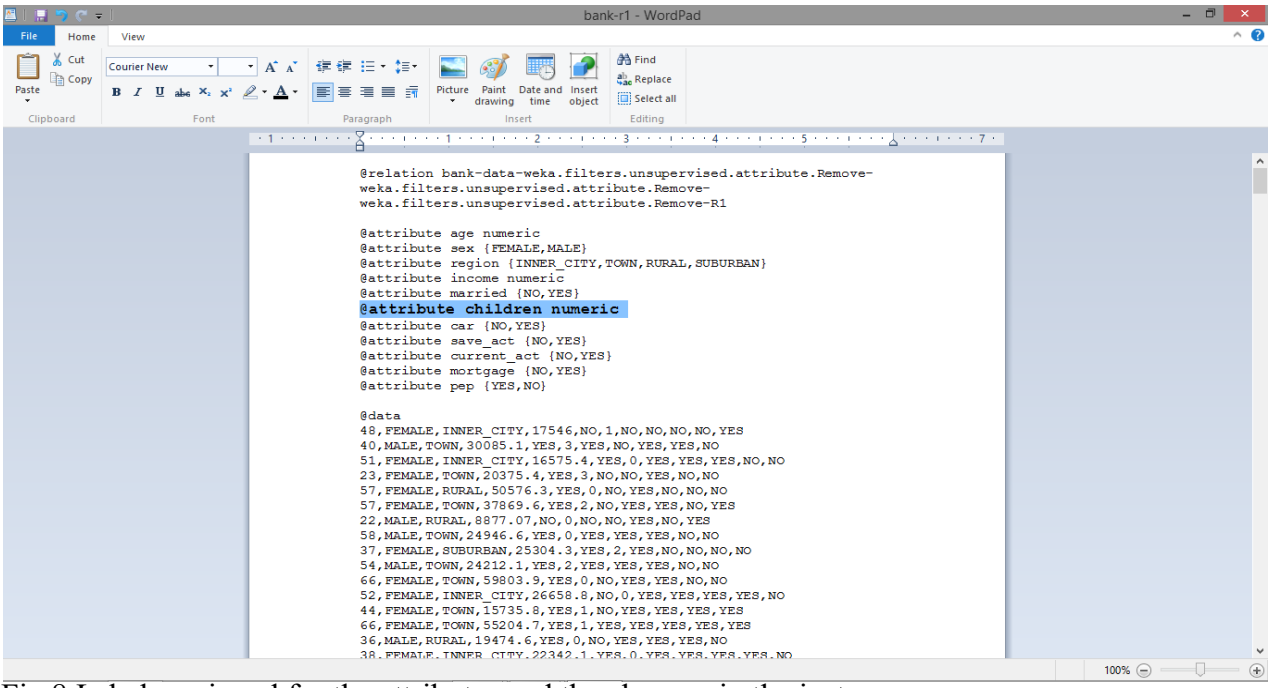


Fig.8 Labels assigned for the attributes and the changes in the instances.

The input file with the above changes is shown Fig.9

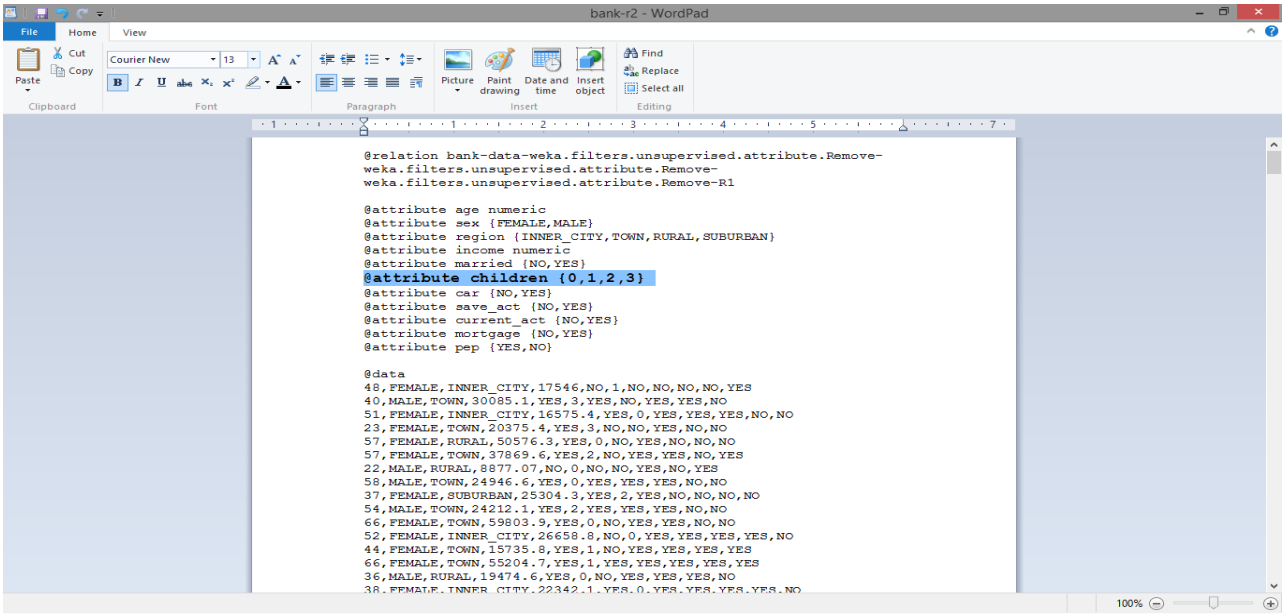


Fig. 9 Automatic generated coding

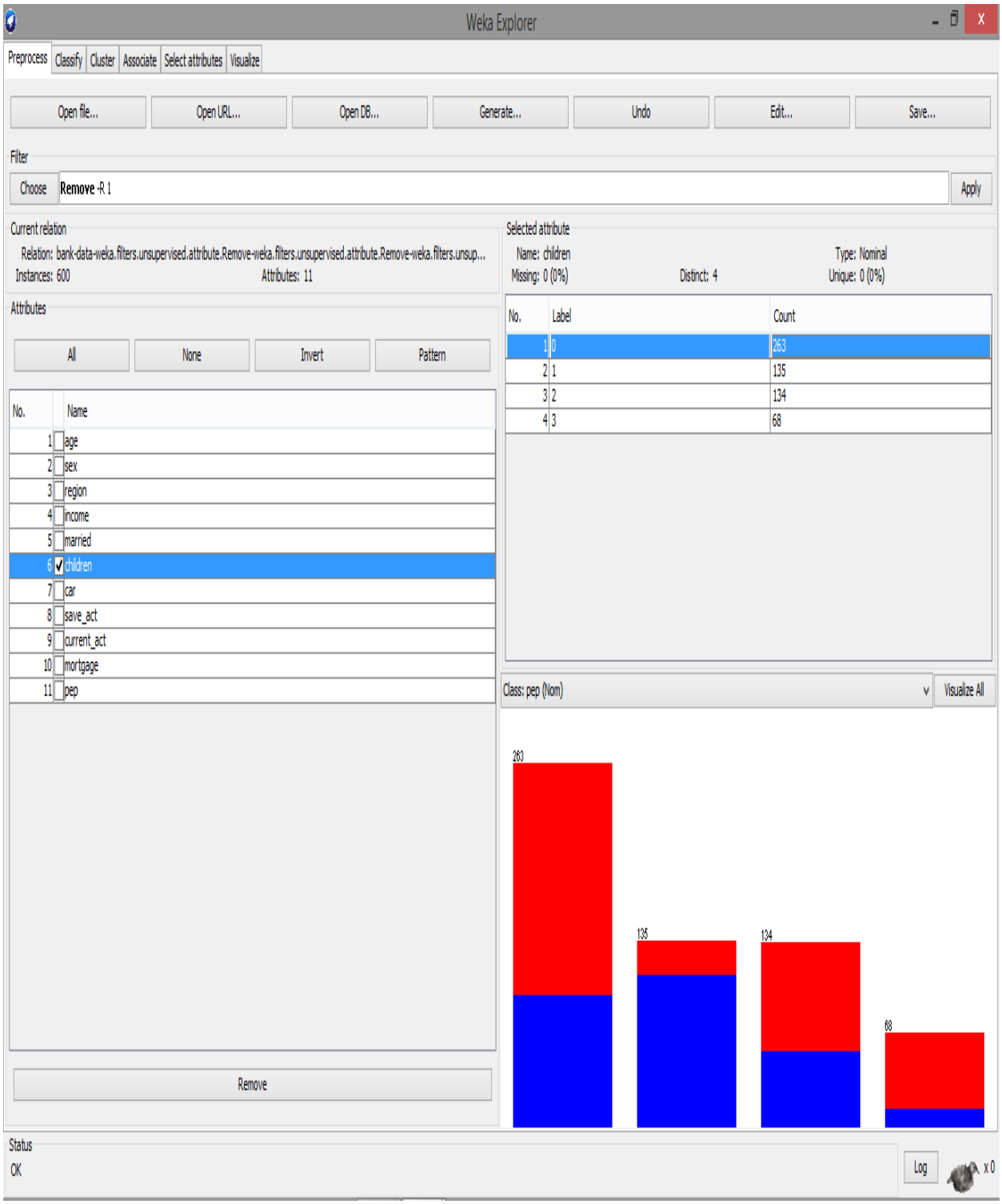


Fig.10 Attribute children details

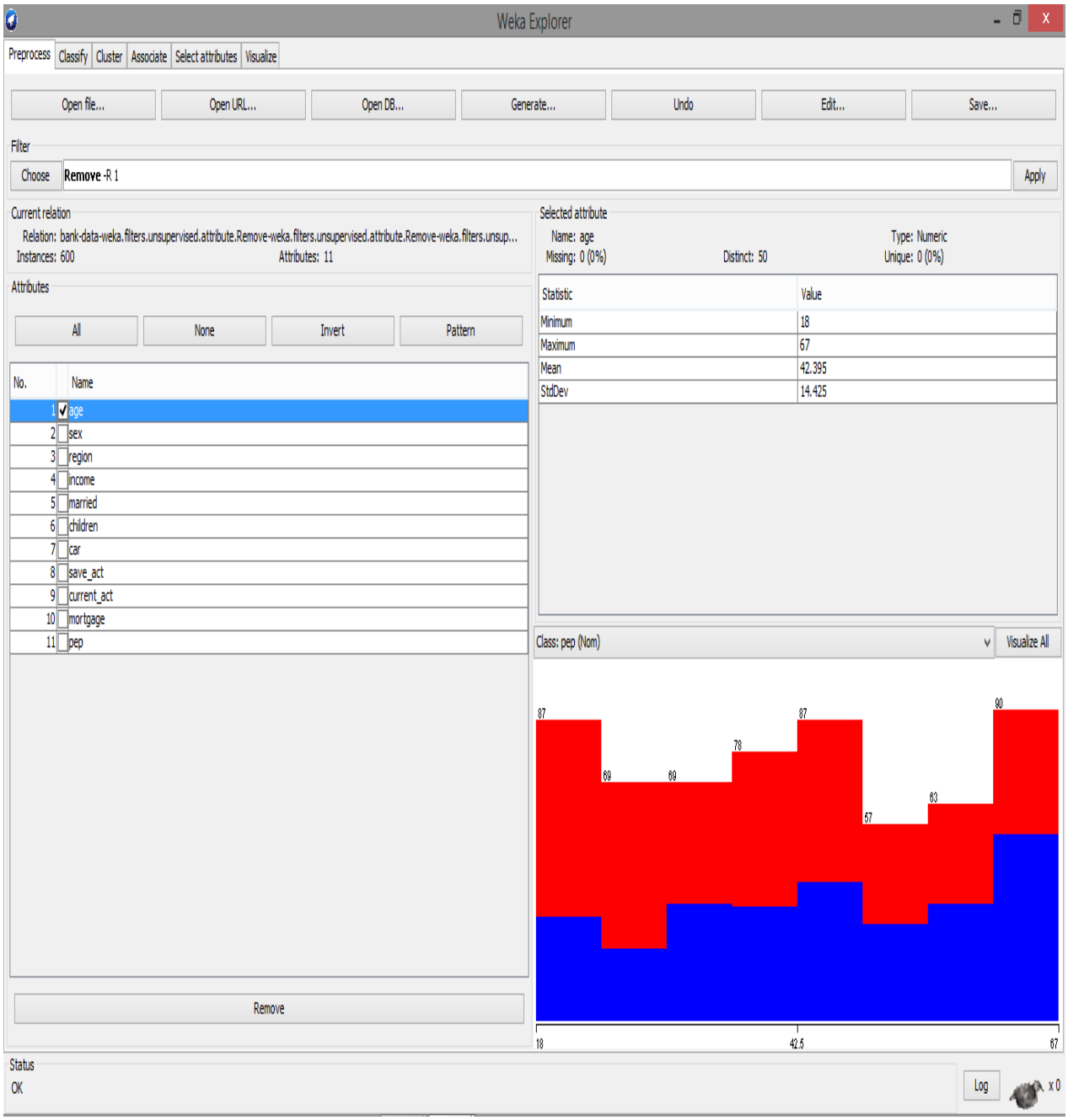
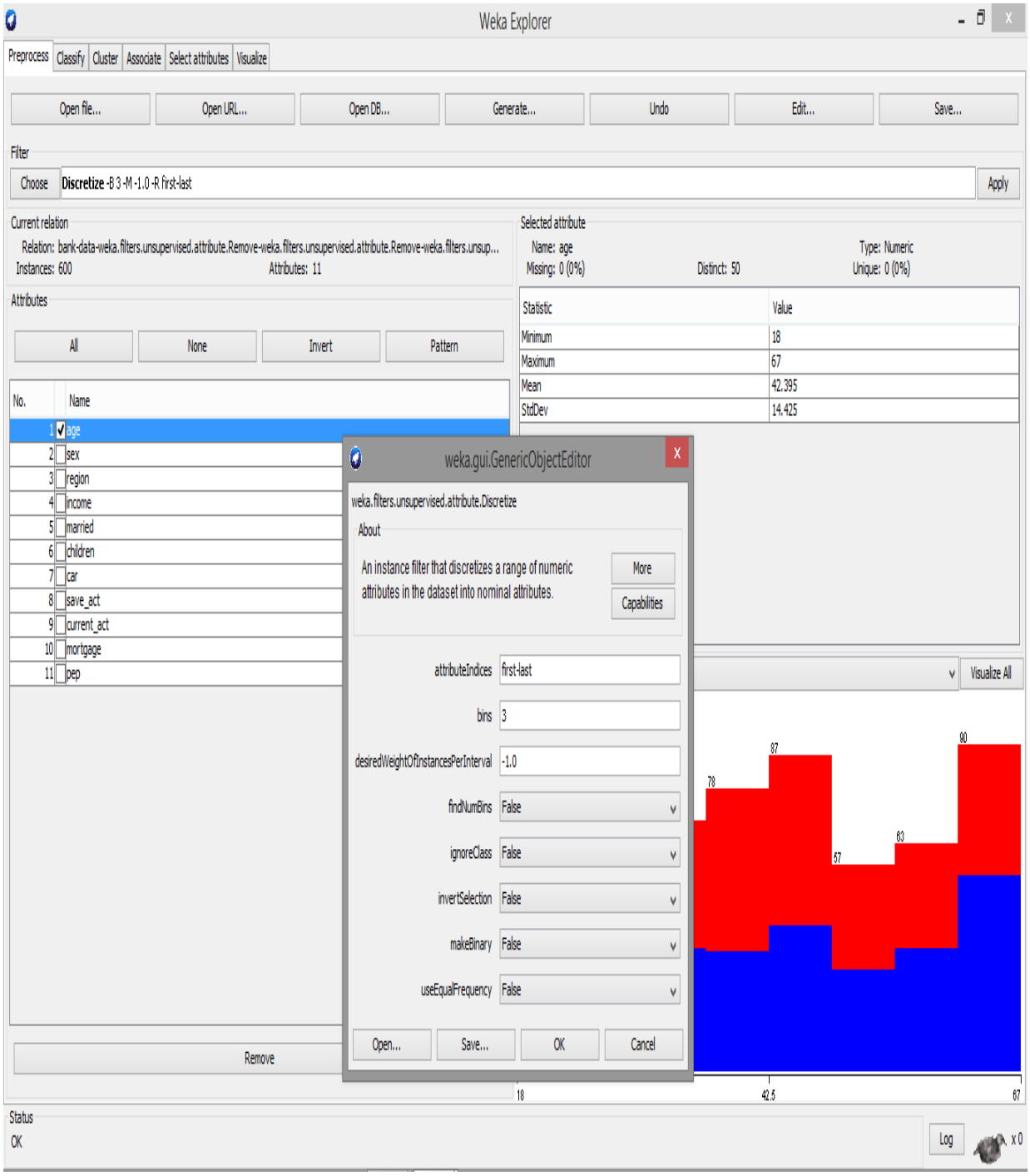


Fig.11 Attribute Age details

The other two attributes age and income have different (continuous) values for different instances so it is discretized by applying discretize filter in the WEKA tool

The age and income ranges are divided into three categories (arbitrary) and the attributes are discretized. Fig.12 and Fig.13 shows parameter modification in age and income attributes.

Fig.12 Discretization in Age attributes



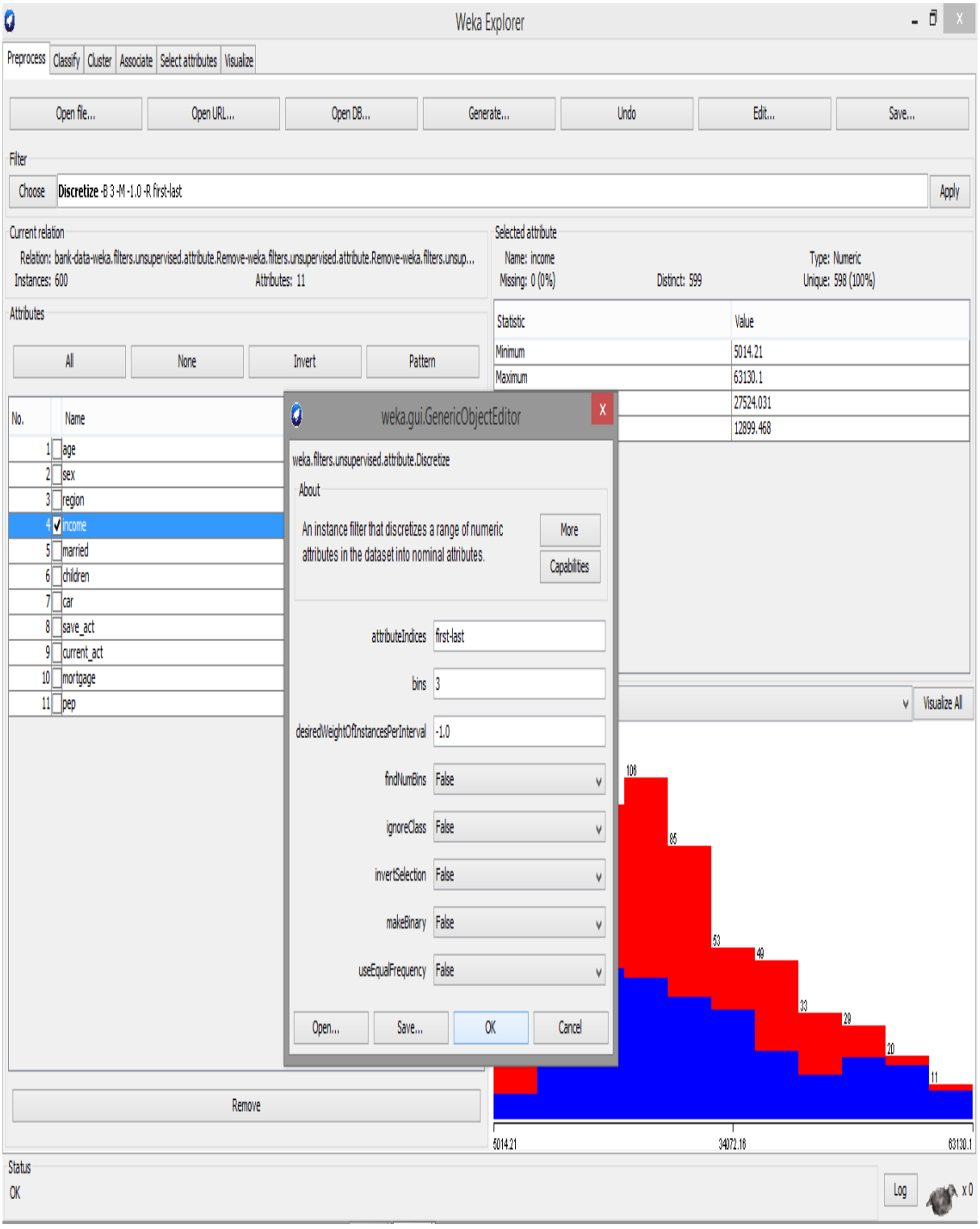
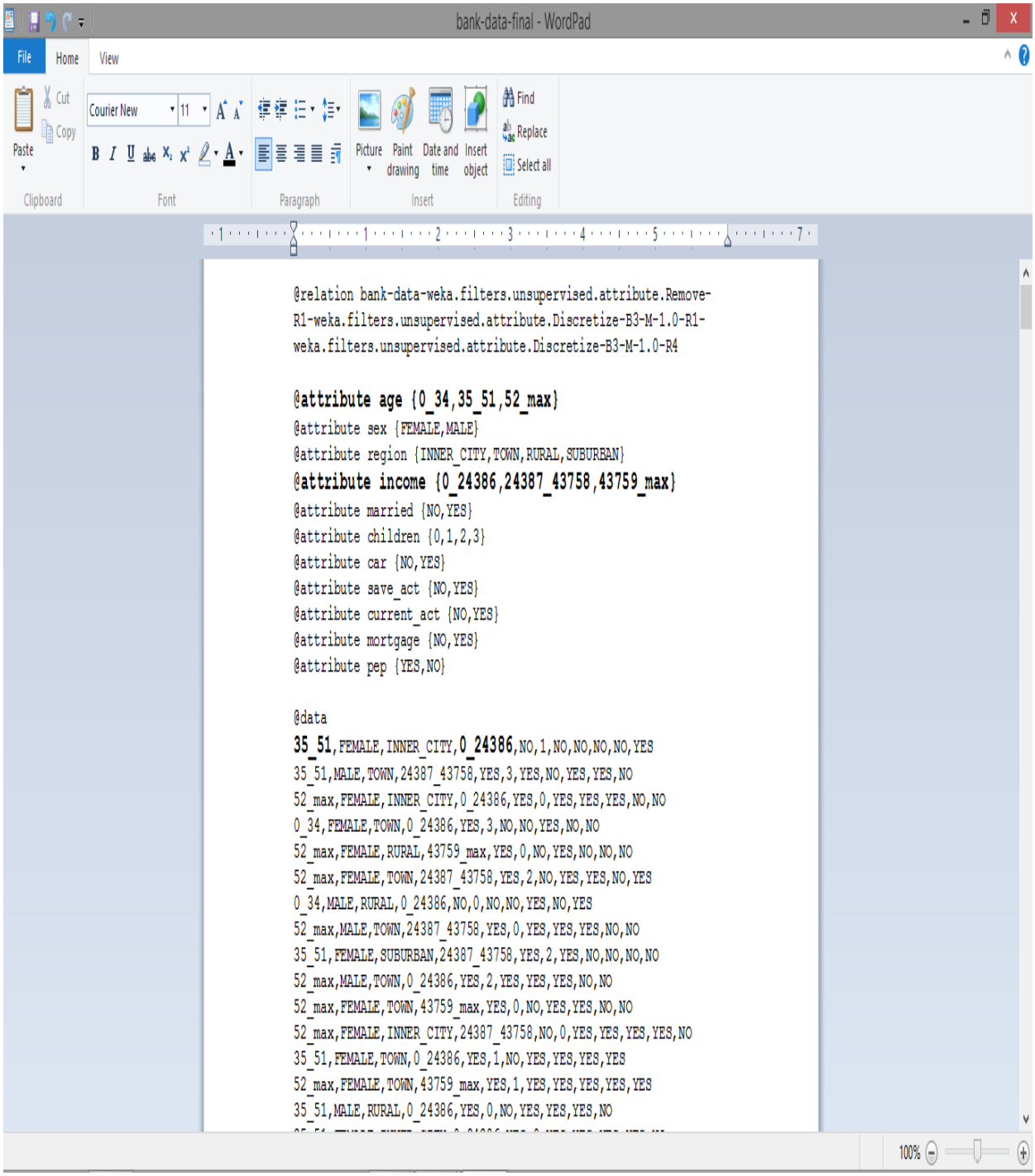


Fig.13 Discretization in Income attributes

The following Fig.14 depicts the labels assigned for the instances (one instance highlighted) of



the attributes age and income. Fig.15 shows the customer dataset after discretization. Fig.14 Labels assigned for the instances of the attributes age and income.

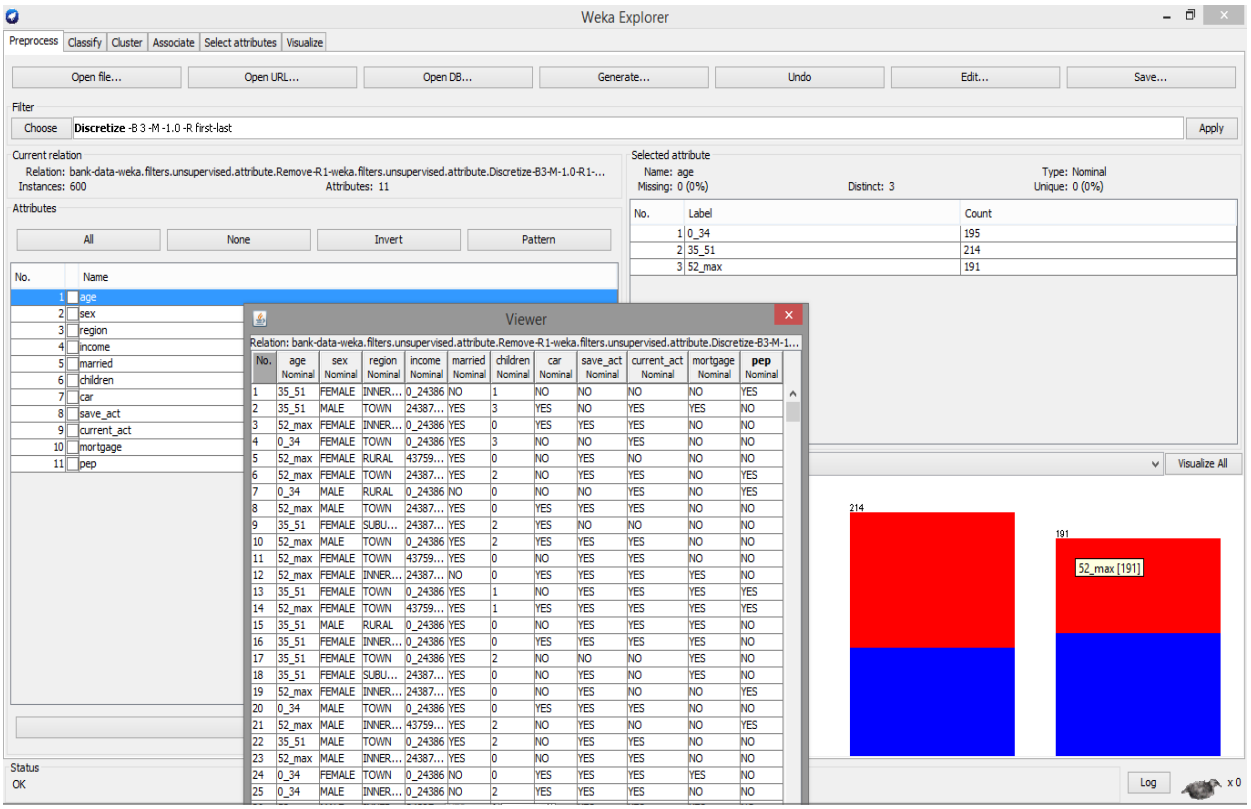


Fig.15 Customer Dataset after discretization

Apriori Algorithm Implementation in Weka:

The preprocessed data file is used for Association rule mining (Apriori Algorithm) and the following rules are generated by setting the necessary measures such as support and confidence is shown in Fig.16 and Fig.17.

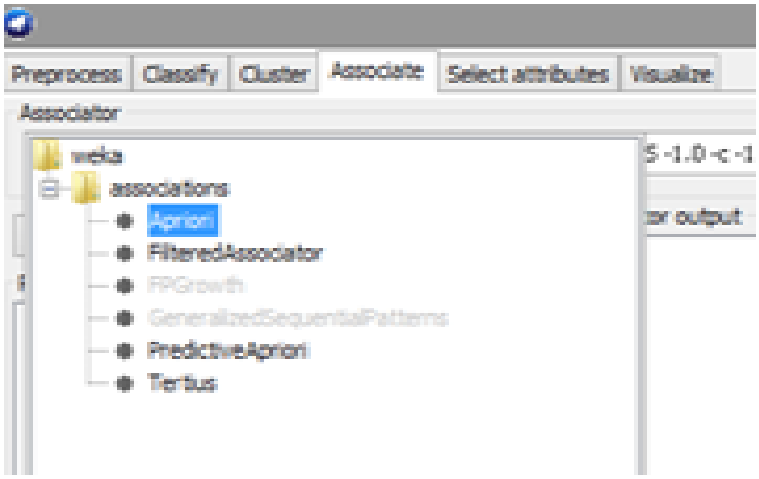
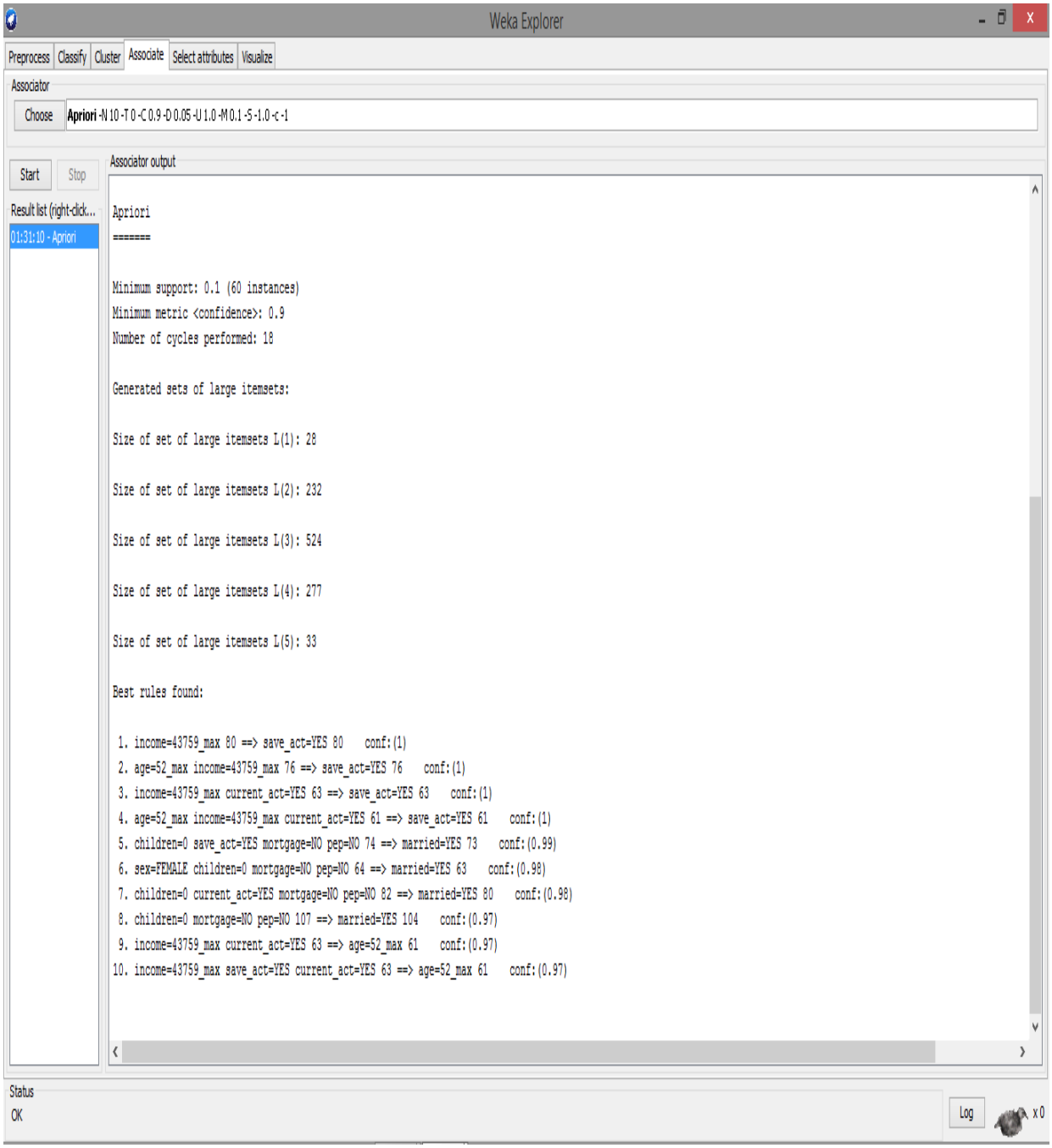


Fig.16 Apriori Algorithm Implementation in Weka



Support and Confidence threshold:
The following Fig.17 shows the parameters set

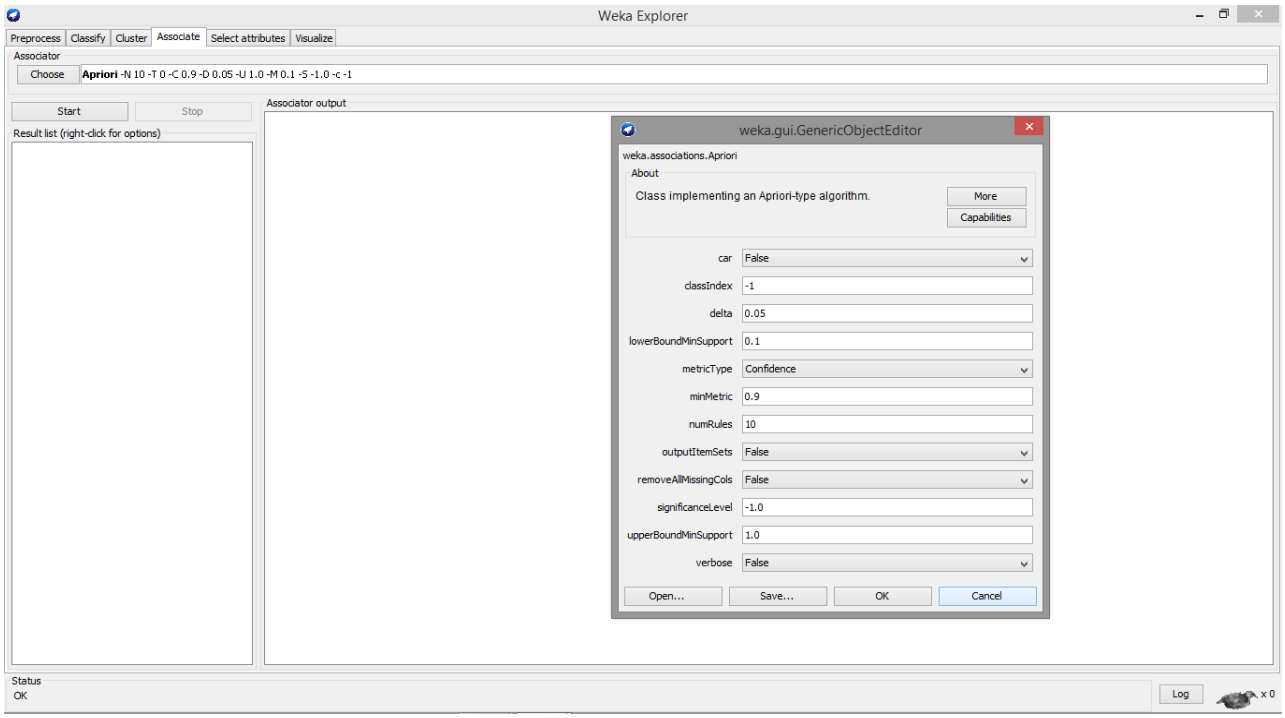


Fig. 17. Minimum Support and Confidence threshold

Output-Rules Generated:

The screen shot shows the rules generated by applying Apriori Algorithm for association rule mining is shown in Fig.18.

== Run information ===

Scheme: weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1

Relation: bank-data-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-R1-weka.filters.unsupervised.attribute.Discretize-B3-M-1.0-R4

Instances: 600

Attributes: 11

age
sex
region
income
married
children
car
save_act
current_act
mortgage
pep

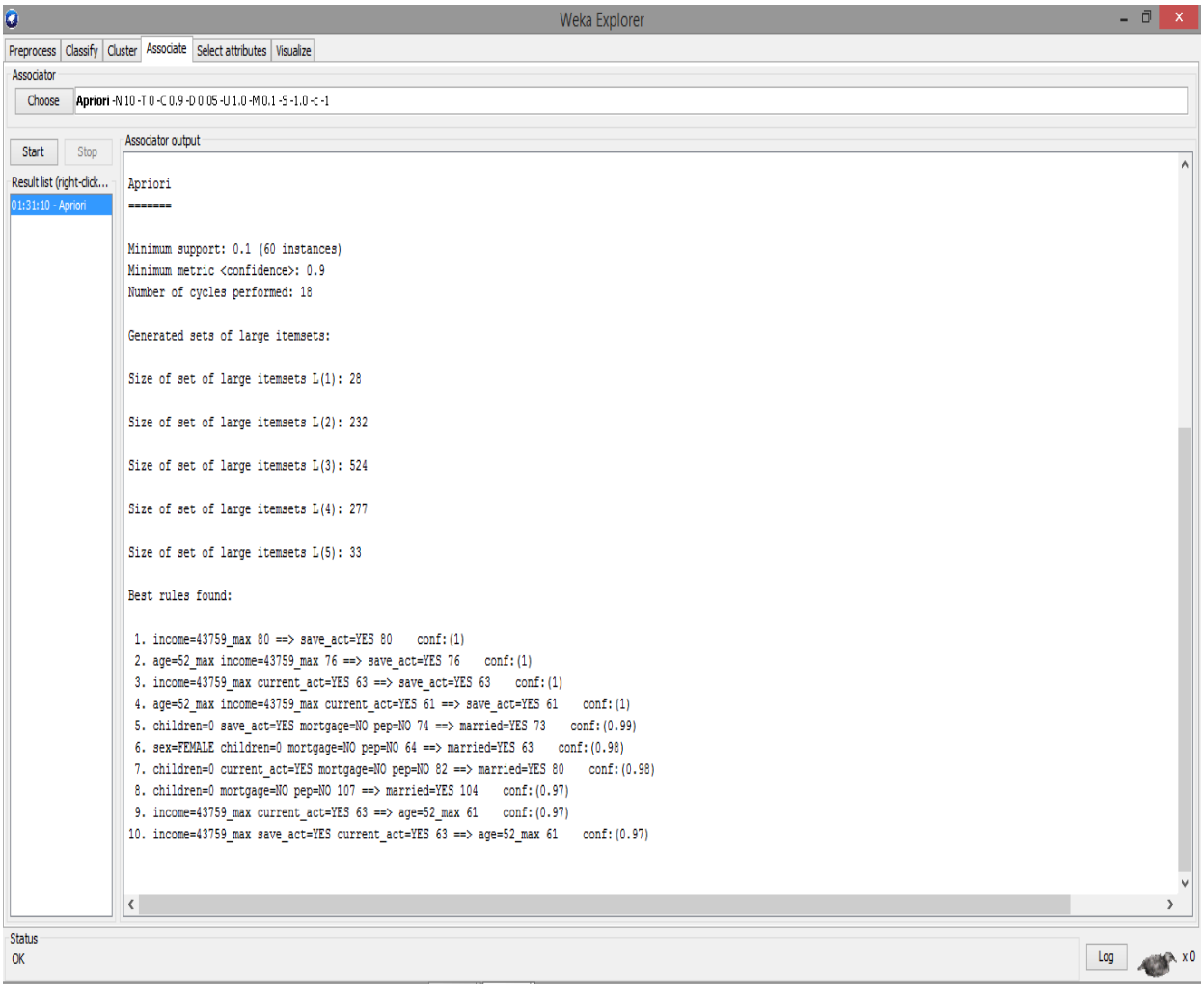


Fig.18. Output rule generated

=== Associator model (full training set) ===
Apriori
========
Minimum support: 0.1 (60 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18
Generated sets of large itemsets:
Size of set of large itemsets L(1): 28
Size of set of large itemsets L(2): 232
Size of set of large itemsets L(3): 524
Size of set of large itemsets L(4): 277
Size of set of large itemsets L(5): 33

Best rules found:
1. income=43759_max 80 ==> save_act=YES 80 conf:(1)
2. age=52_max income=43759_max 76 ==> save_act=YES 76 conf:(1)
3. income=43759_max current_act=YES 63 ==> save_act=YES 63 conf:(1)
4. age=52_max income=43759_max current_act=YES 61 ==> save_act=YES 61 conf:(1)
5. children=0 save_act=YES mortgage=NO pep=NO 74 ==> married=YES 73 conf:(0.99)
6. sex=FEMALE children=0 mortgage=NO pep=NO 64 ==> married=YES 63 conf:(0.98)
7. children=0 current_act=YES mortgage=NO pep=NO 82 ==> married=YES 80 conf:(0.98)
8. children=0 mortgage=NO pep=NO 107 ==> married=YES 104 conf:(0.97)
9. income=43759_max current_act=YES 63 ==> age=52_max 61 conf:(0.97)
10. income=43759_max save_act=YES current_act=YES 63 ==> age=52_max 61 conf:(0.97)

CONCLUSION:

The above rules infer that most of the customers whose age is above 52 and income greater than 43,000 have a saving account and current account. Most of the customers who has no children, no mortgage and no personal equity plan is married.

REFERENCES

1. Donn Morrison, Ruili Wang, Liyanage C. De Silva, Ensemble methods for spoken emotion recognition in call-centres, Speech Communication, Elsevier, Vol. 49, pp.98-112, 2007
2. Amira A. El Tayeb, Vikas Pareek, Abdelaziz Araar, Applying Association Rules Mining Algorithms for Traffic Accidents in Dubai , International Journal of Soft Computing and Engineering, Vol. 5, No.4, pp. 1-12, 2015
3. Roghayeh Saneifar and Mohammad Saniee Abadeh, Jan Zizka et al., Association Rule Discovery for Student Performance Prediction Using Metaheuristic Algorithms, pp. 115–123, 2015, DOI : 10.5121/csit.2015.51510
4. Amit Dipchandji Kasliwal, Dr. Girish S. Katkar,"Web Usage mining for Predicting User Access Behaviour", International Journal of Computer Science and Information Technologies, Vol.6, No.1, pp-201-204, 2015.
5. Rahul Neve 1, K.P Adhiya ", Comparative Study of Web Mining Algorithms for Web Page Prediction in Recommendation System" , international Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, No.1, pp.969-976, 2013
6. Dhruva Mistry, Kirti Sharma, Samip A.Patel, Recommend Websites through Weblog Files using Association Rule, International Journal of Computer Applications, Vol.126, No.2, pp.16-19, 2015.
7. Kashish Ara Shakil, Shadma Anis and Mansaf Alam, Dengue Disease Prediction Using Weka Data Mining Tool, Pp.1-26, 2015.

Author(s) Biography

Thangamani completed her B.E., from Government College of Technology, Coimbatore, India. She completed her M.E in Computer Science and Engineering from Anna University and PhD in Information and Communication Engineering from the renowned Anna University, Chennai, India in the year 2013.

Dr. M. Thangamani possesses nearly 23 years of experience in research, teaching, consulting and practical application development to solve real-world business problems using analytics. Her research expertise covers Medical data mining, machine learning, cloud computing, big data, fuzzy, soft computing, ontology development, web services and open source software. She has published nearly 70 articles in refereed and indexed journals, books and book chapters and presented over 67 papers in national and international conferences in above field. She has delivered more than 60 Guest Lectures in reputed engineering colleges and reputed industries on various topics. She has got best paper awards from various education related social activities in India and Abroad. She has organized many self-supporting and government sponsored national conference and Workshop in the field of data mining, big data and cloud computing. She continues to actively serve the academic and research communities and presently guiding Ph.D Scholars under Anna University. She is also seasonal reviewer in IEEE Transaction on Fuzzy System, international journal of advances in Fuzzy System and Applied mathematics and information journals. She has organizing chair and keynote speaker in international conferences in India and countries like California, Dubai, Malaysia, Singapore, Thailand and China. She has received many awards for academic activities. She is on the editorial board and reviewing committee of leading research journals, which includes her nomination as the Editor in chief to International Scientific Global journal for Engineering, Science and Applied Research (ISGJESAR) & International Research Journal in Global Engineering and Sciences (IRJGES) and on the program committee of top international data mining and soft computing conferences in various countries. She has Life Membership in ISTE, Member in CSI, International Association of Engineers and Computer Scientists in China, IAENG, IRES, Athens Institute for Education and Research and Life member in Analytical Society of India. She is currently working as Assistant Professor at Kongu Engineering College at Perundurai, Erode District.

Mr. N.Suresh Kumar is currently working as Assistant Professor in the Department of Information Technology at Sri Ramakrishna Engineering College, Coimbatore. He is a part time research scholar under Anna University, Chennai. He has a Master's degree in Software Engineering (2010) at Anna University of Technology, Coimbatore. He has nearly 13 years of academic experience and 2 Years of Industry experience. He is proficient in Data Mining, Software Engineering, Database Management System, Web services and ontology development. He has published the papers in International and National Journals and presented the papers in National conferences. He is a life member of ISTE, IAENG and IACSIT.