Early Detection of ovarian cancer (*BARCA 1 & BARCA 2* MUTATION) risk prediction for low income country using Data mining technology: Bangladesh

[1]Md. Shariful Islam, [5]M. Salahuddin, [4]Selina Khatun, [2]Sayed Asaduzzaman, [2]Kawsar Ahmed, [3]Sojib Paul, [1]Masum Parvez, [1]Hasibul Haque Rakib, [1*]Abu Zaffar Shibly
[1] Department of Biotechnology and Genetic Engineering, Faculty of Life Science, Mawlana Bhashani Science and Technology University, Tangail-1902, Bangladesh
[2]Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Santosh, Tangail-1902, Bangladesh
[3]Indoor Medical Officer (Surgery) Dhaka Medical College Hospital
[4]Department of Obstetrics and Gynecology, Bangabandhu Sheikh Mujib Medical University (Ex PG hospital, Dhaka, Bangladesh)
[5]Li Ka Shing Faculty of Medicine, University of Hong Kong

Presenter: Md. Shariful Islam, sharifbge@gmail.com
***Corresponding Author**: Abu Zaffar Shibly, zaffarshibly1987@gmail.com

ABSTRACT

Background: Ovarian cancer is the most lethal gynecological cancer which incidence increasing day by day in developing countries. More than 80% ovarian cancers occur in women over the age of 50. Therefore, identification of genetic factors including mutations in the *BRCA1* and *BRCA2* gene (breast cancer gene)as well as others factors is very crucial in developing novel strategy of ovarian cancer prevention. **Methodology:** This study was carried out in 521 cancer and non-cancer patients' data is collected from different diagnostic centre and data pre-processed. Then a structured questionnaire was used containing details of ovarian cancer risk factors including age, menopause end age, problem during pregnancy, first sex age, any infection in genital area, affected by ovarian cancer, abortion, pregnancy, BMI, menopause age after 50, food habit, obesity, excessive alcohol, late Menopause, early Menopause, hormone therapy, exercise, previous exposure to other sexually transmitted infections (STIs), marital status, genetic risk, outdoor activities and affected any cancer before based on the previous studies. **Results:** After pre-processing data is clustered using K-means clustering algorithm for identifying relevant and non-relevant data to ovarian Cancer. Next significant frequent patterns are discovered using AprioriTid shown in Table 1 and Decision Tree algorithm shown in Table 2. This ovarian cancer risk prediction system will be helpful in detection of a patient's predisposition to ovarian cancer. Specifically there were no work of ovarian cancer risk prediction system using data mining or Statistical approaches. **Conclusions:** The majority of cases are diagnosed at late stages when cure is impossible. Therefore early prediction of ovarian cancer should play a pivotal role in the diagnosis process and for an effective preventive strategy.

Keywords: Ovarian cancer, STIs, Data mining, Risk prediction

Corresponding Author: *Abu Zaffar Shibly

zaffarshibly1987@gmail.com

Poster Design & Printing by Genigraphics® - 800.790.4001

INTRODUCTION

The most common type of ovarian cancer is called ovarian epithelial cancer. It begins in the tissue that covers the ovaries. Cancer sometimes begins at the end of the fallopian tube near the ovary and spreads to the ovary. The most common hereditary cancer syndrome associated with ovarian cancer involves mutations in the *BRCA* gene (breast cancer gene). People who carry harmful mutations in this tumour-suppressor gene are at increased risk of breast and ovarian cancer. The lifetime risk of ovarian cancer for carriers of *BRCA1* is 35-46 per cent, and for *BRCA2* mutation carriers it is 13-23 per cent. According to CDC (Centers for disease control and prevention) it has been reported that most women get it without being at high risk. These cancers are often found at advanced stages. This is partly because they may not cause early signs or symptoms and there are no good screening tests for them. There are lots of works to detect the risk factors of ovarian cancer using population based case control study several databases, and algorithm and induction techniques. Clustering is a process of separating dataset into subgroups according to the unique feature. Clustering separated the dataset into relevant and non-relevant dataset to ovarian Cancer. AprioriTid (Ilias and Quan) and Decision Tree algorithm[8] are mainly used to find out frequent patterns of dataset. [1-5] Those algorithms are very easy and effective to find out frequent patterns. Significant frequent pattern, the set of data are mostly responsible to ovarian Cancer. Using this significant pattern we implemented a prediction system for ovarian Cancer. The main goal of our research is to develop a system that can be used by women for testing her ovarian Cancer risk level.

Methodology

Data collection

Total data of 521 participant (women) was collected from different Diagnostic center and Medical college hospital. Both the patients (400) and non-patients (121) were women whose ages were about 35 to 65. The whole data collection process was performed based on a questionere. About 27 risk factors were considered for ovarian cancer assessment in Bangladeshi population.

Data mining approaches using WEKA

Highly significant 10 factors have been exploited from the analysis of the statistical approaches and depending on the results of SPSS. Then those factors are ranked by ranking algorithm with attribute evaluator Correlation AttributeEval. Incomplete data hampers the analysis which has been eliminated or leveled. A little bit of data were changes to avoid collision of the data analysis.

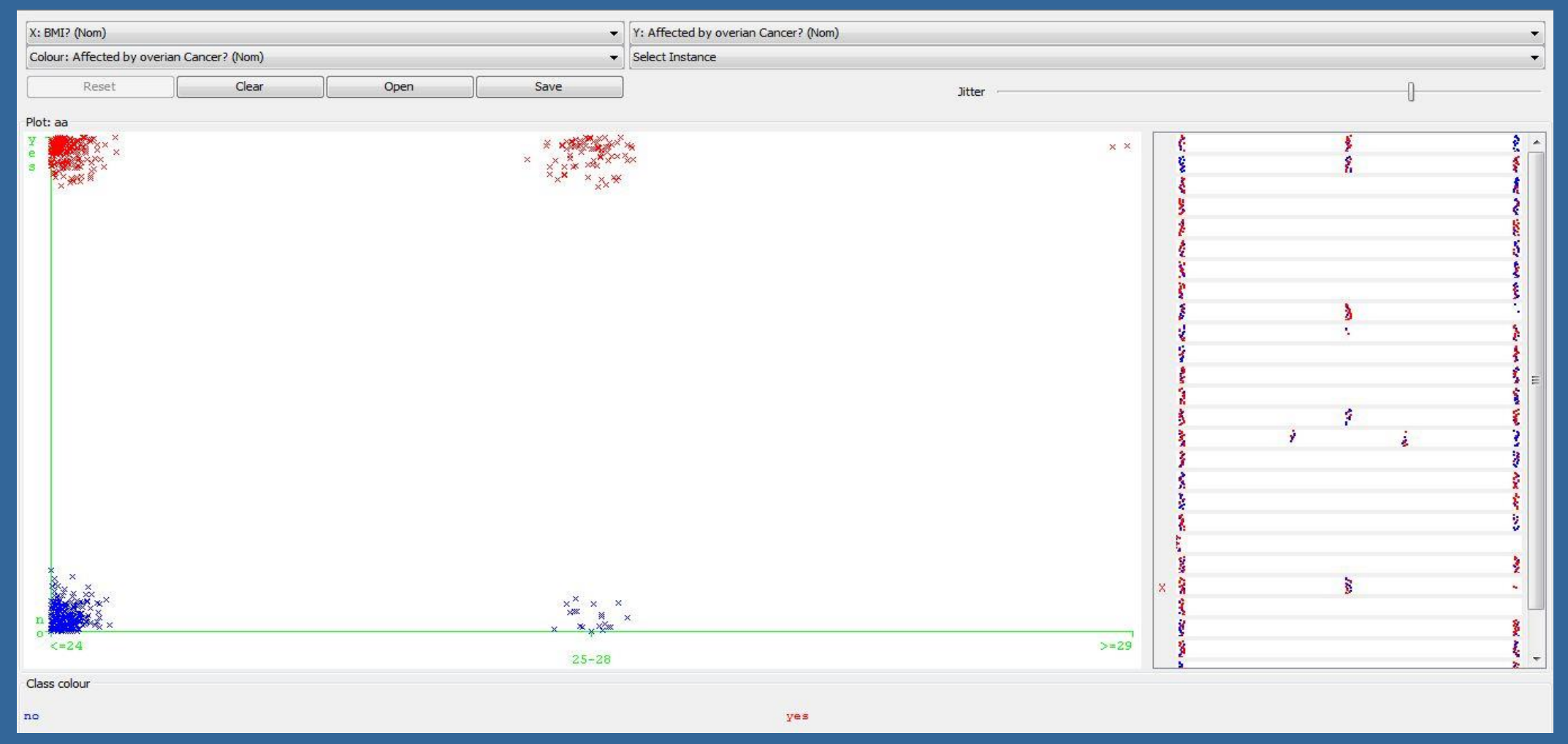

X: BMI? (Nom)
Y: Affected by overian Cancer? (Nom)
Colour: Affected by overian Cancer? (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot: aa
yes

Figure: 4

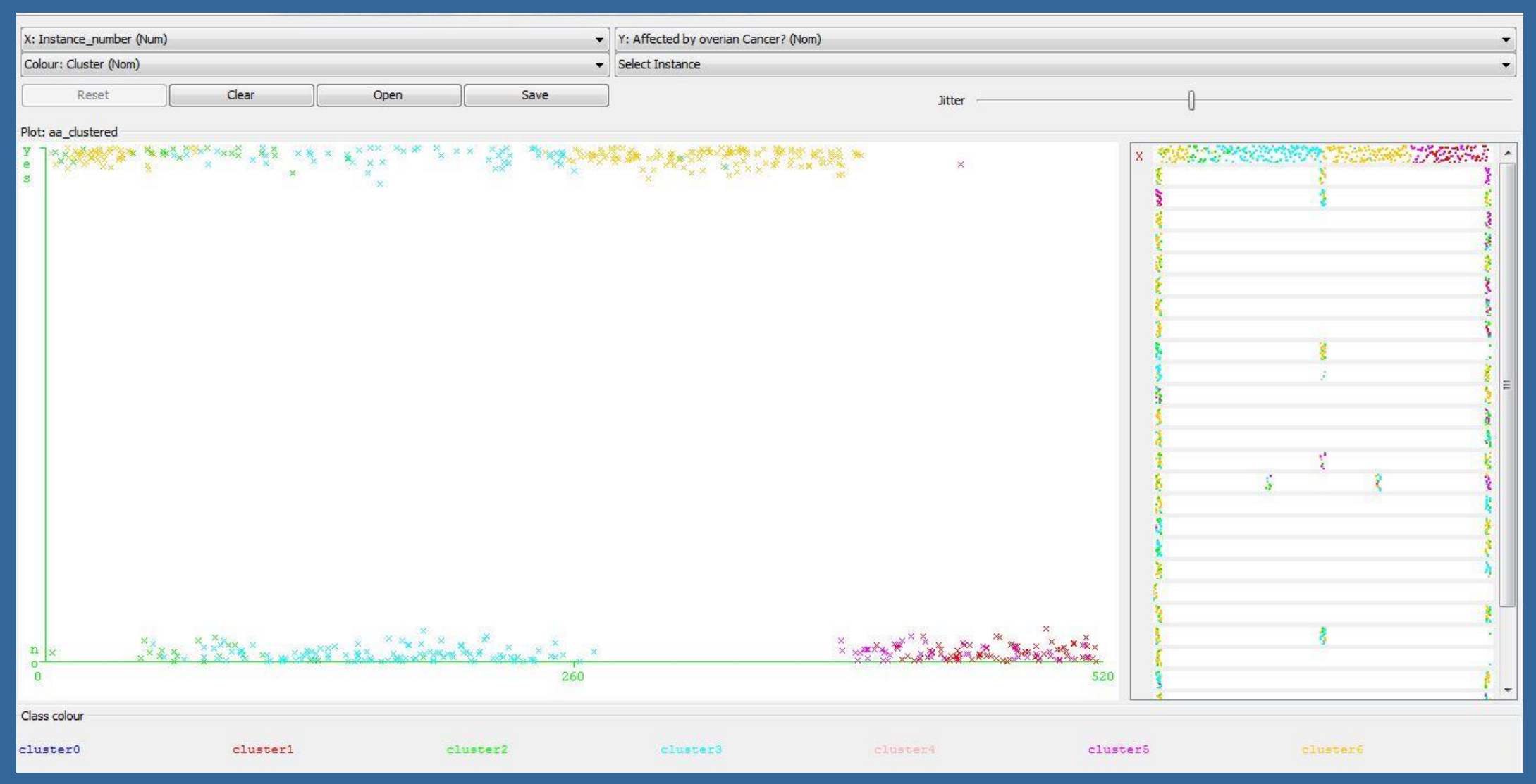


Figure: 5

RESULTS & Analysis

The frequency table was contrived by comparing the results of SPSS and WEKA. Both statistical and data mining approaches shows same frequency. Age range between 35 and 65 where the mean age was 50, approximately 521 Bangladeshi women's data were analyzed. Here 121 women were not affected (control group) and 421 women were affected by ovarian cancer (case group).We performed the data visualization analysis and clustering using data mining technology. The main approach of this study to finalize the data analysis with decision tree algorithm based tree by which we can predict if a person is affected by ovarian cancer. Likely if one have "problem during pregnancy?" then if had "abortion?" then if yes she has the possibility to have ovarian cancer here the tree is based on the highly significant 10 factors not all the factors. We also analyzed the algorithm based output among the figures which causes ovarian cancer combined. (CorrelationAttributeEval) evaluators shows the higher significance and the table also gives some decisions like factor (Problem during pregnancy) got higher precedence according the table. So it can be said that the factor (Problem during pregnancy) is highly significant than the others factors including abortion, use of napkin, birth control pill and so on.

Figure 1 describes the visualization of the output of the factors based on their frequency level Association, among the significant factors. **Figure** 2 describes the association/correlation among the significant factors. This is also called the significant Binary pattern. **Figure 3** shows the algorithm based output among the figures which causes cancer combined. **Figure 4** and **Figure 6** are the data clustering, which describes the one process of data mining to find out the significance of the risk factors. **Figure 5** describes Clustered data between the factor "affected by ovarian cancer?" with the other factors. **Figure 7** describes a decision tree by which we can predict if a women is affected by ovarian cancer. Suppose if one have "problem during pregnancy?" then if had "abortion?" then if "yes" she has the possibility to have ovarian cancer here the tree is based on the highly significant 10 factors not all the factors.

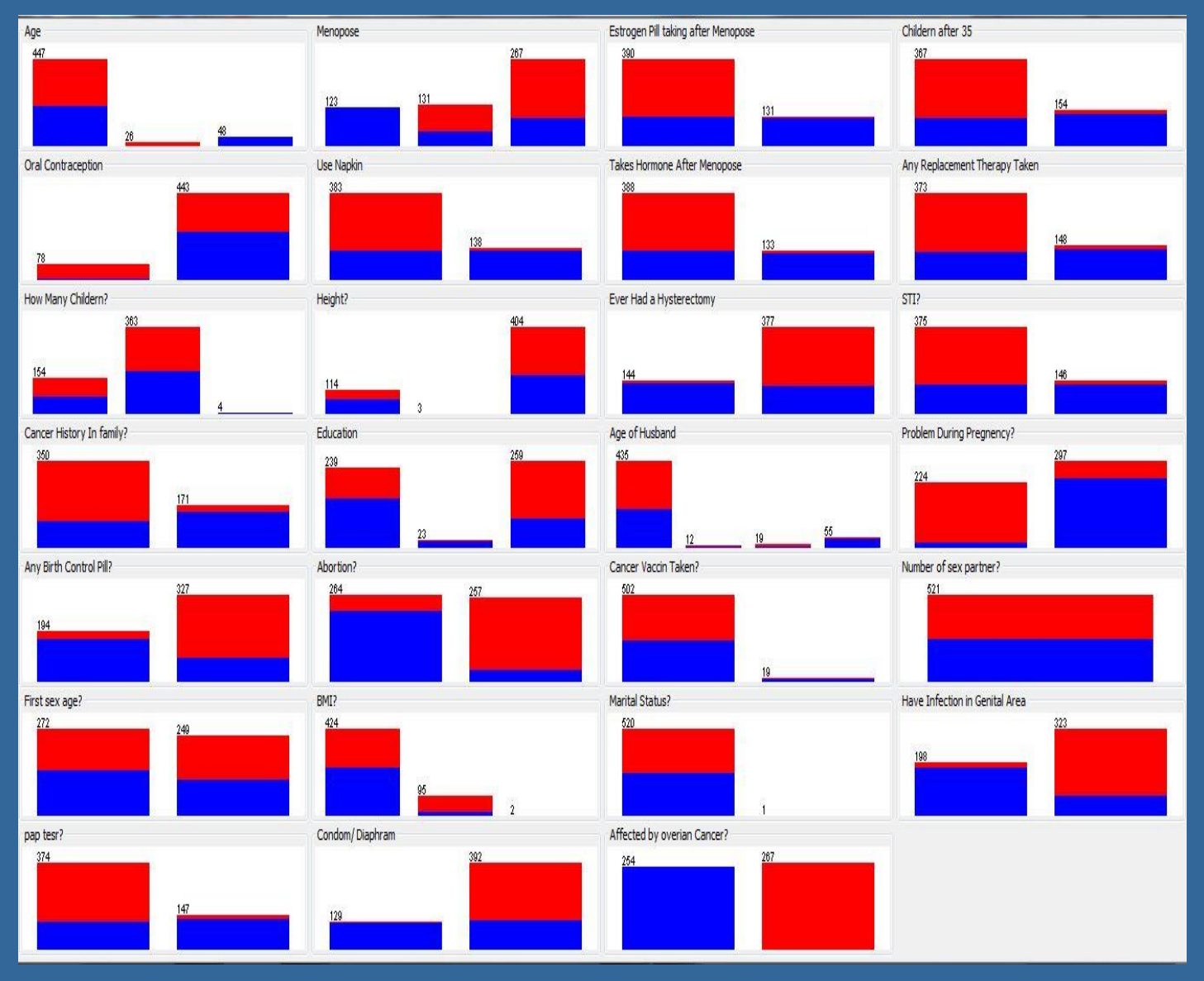
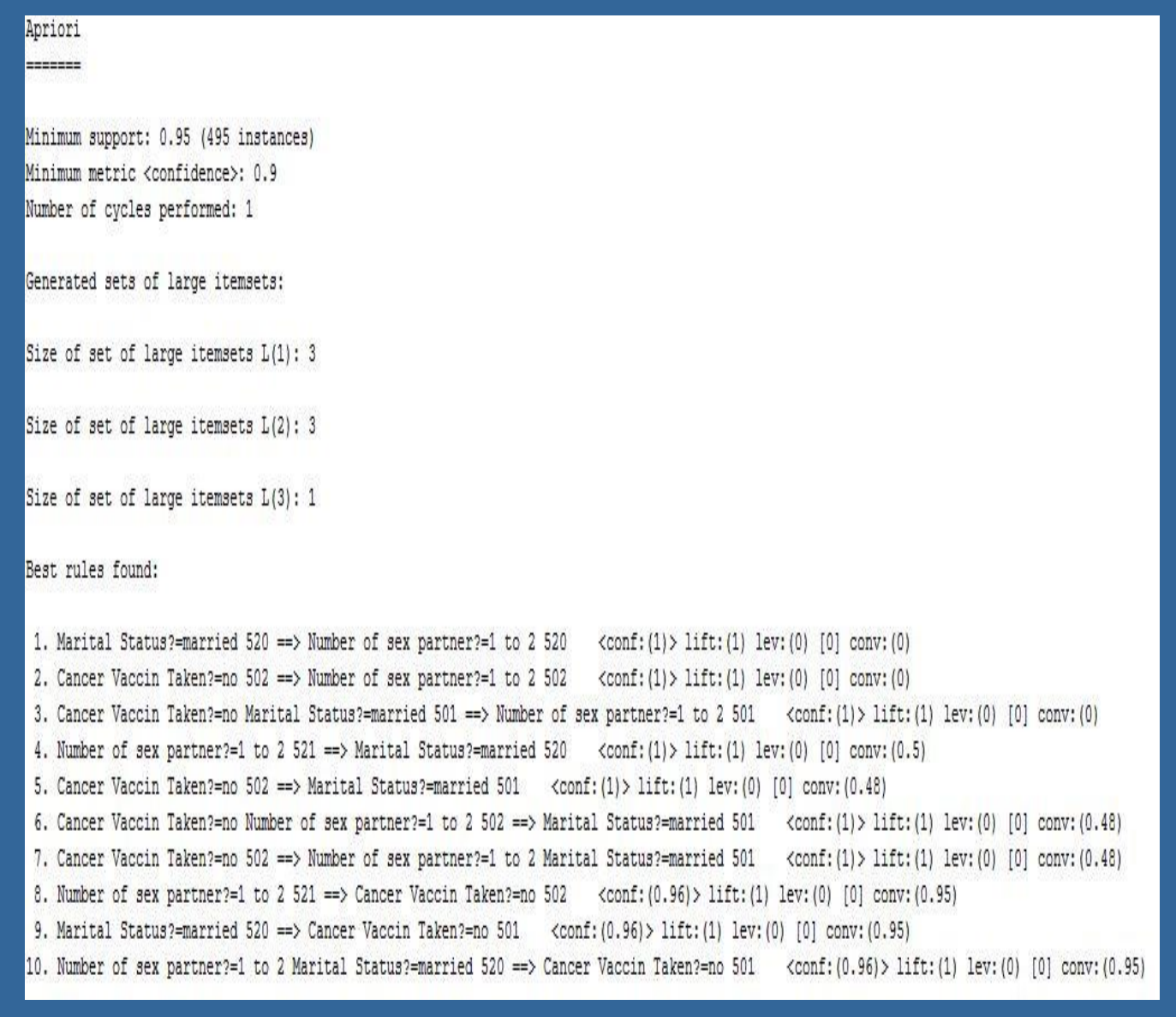


Figure: 1



```
Apriori
=======

Minimum support: 0.95 (495 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 1

Generated sets of large itemsets:

Size of set of large itemsets L(1): 3

Size of set of large itemsets L(2): 3

Size of set of large itemsets L(3): 1

Best rules found:

 1. Marital Status?=married 520 ==> Number of sex partner?=1 to 2 520    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 2. Cancer Vaccin Taken?=no 502 ==> Number of sex partner?=1 to 2 502    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 3. Cancer Vaccin Taken?=no Marital Status?=married 501 ==> Number of sex partner?=1 to 2 501    <conf:(1)> lift:(1) lev:(0) [0] conv:(0)
 4. Number of sex partner?=1 to 2 521 ==> Marital Status?=married 520    <conf:(1)> lift:(1) lev:(0) [0] conv:(0.5)
 5. Cancer Vaccin Taken?=no 502 ==> Marital Status?=married 501    <conf:(1)> lift:(1) lev:(0) [0] conv:(0.48)
 6. Cancer Vaccin Taken?=no Number of sex partner?=1 to 2 502 ==> Marital Status?=married 501    <conf:(1)> lift:(1) lev:(0) [0] conv:(0.48)
 7. Cancer Vaccin Taken?=no 502 ==> Number of sex partner?=1 to 2 Marital Status?=married 501    <conf:(1)> lift:(1) lev:(0) [0] conv:(0.48)
 8. Number of sex partner?=1 to 2 521 ==> Cancer Vaccin Taken?=no 502    <conf:(0.96)> lift:(1) lev:(0) [0] conv:(0.95)
 9. Marital Status?=married 520 ==> Cancer Vaccin Taken?=no 501    <conf:(0.96)> lift:(1) lev:(0) [0] conv:(0.95)
10. Number of sex partner?=1 to 2 Marital Status?=married 520 ==> Cancer Vaccin Taken?=no 501    <conf:(0.96)> lift:(1) lev:(0) [0] conv:(0.95)
```

Figure: 2

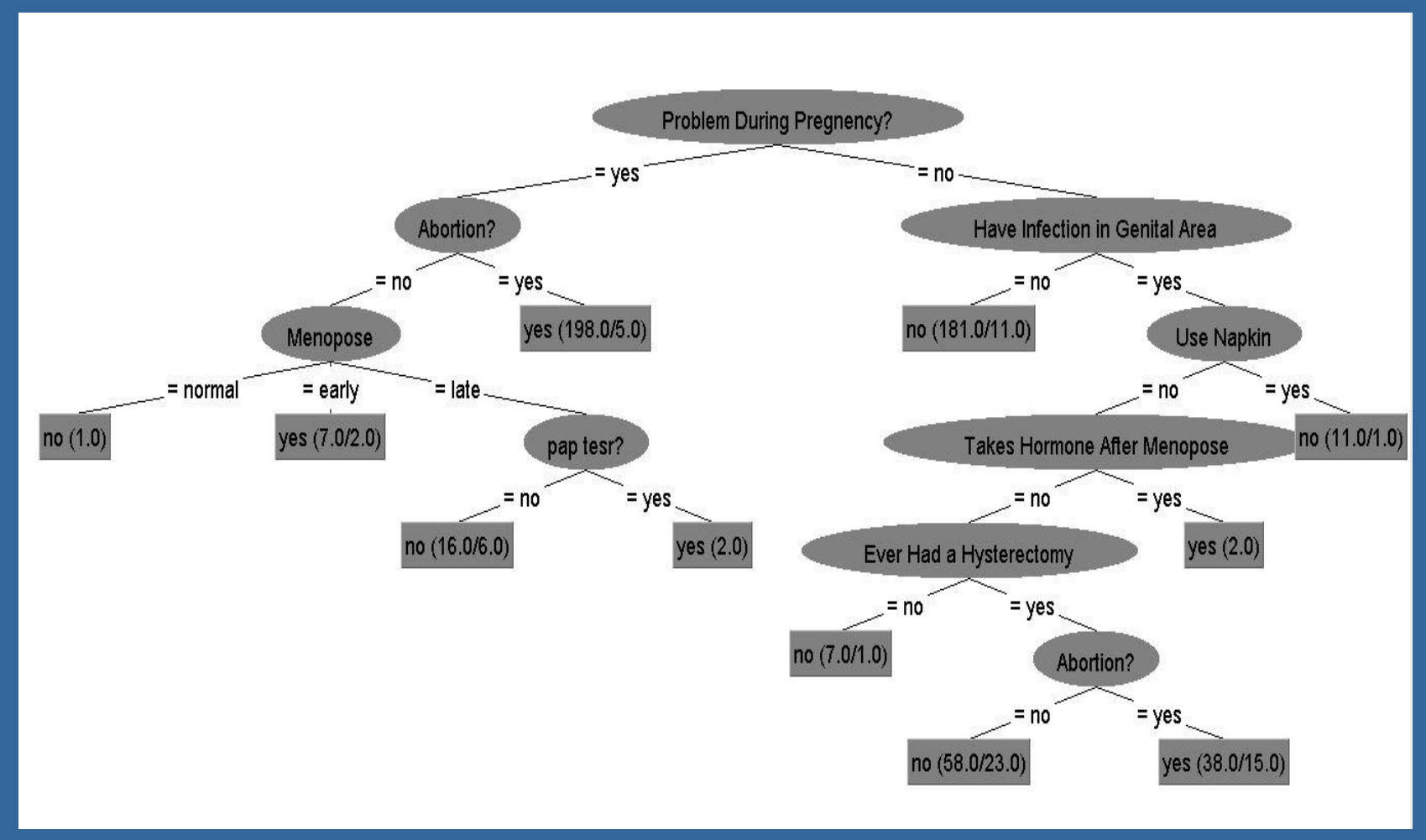
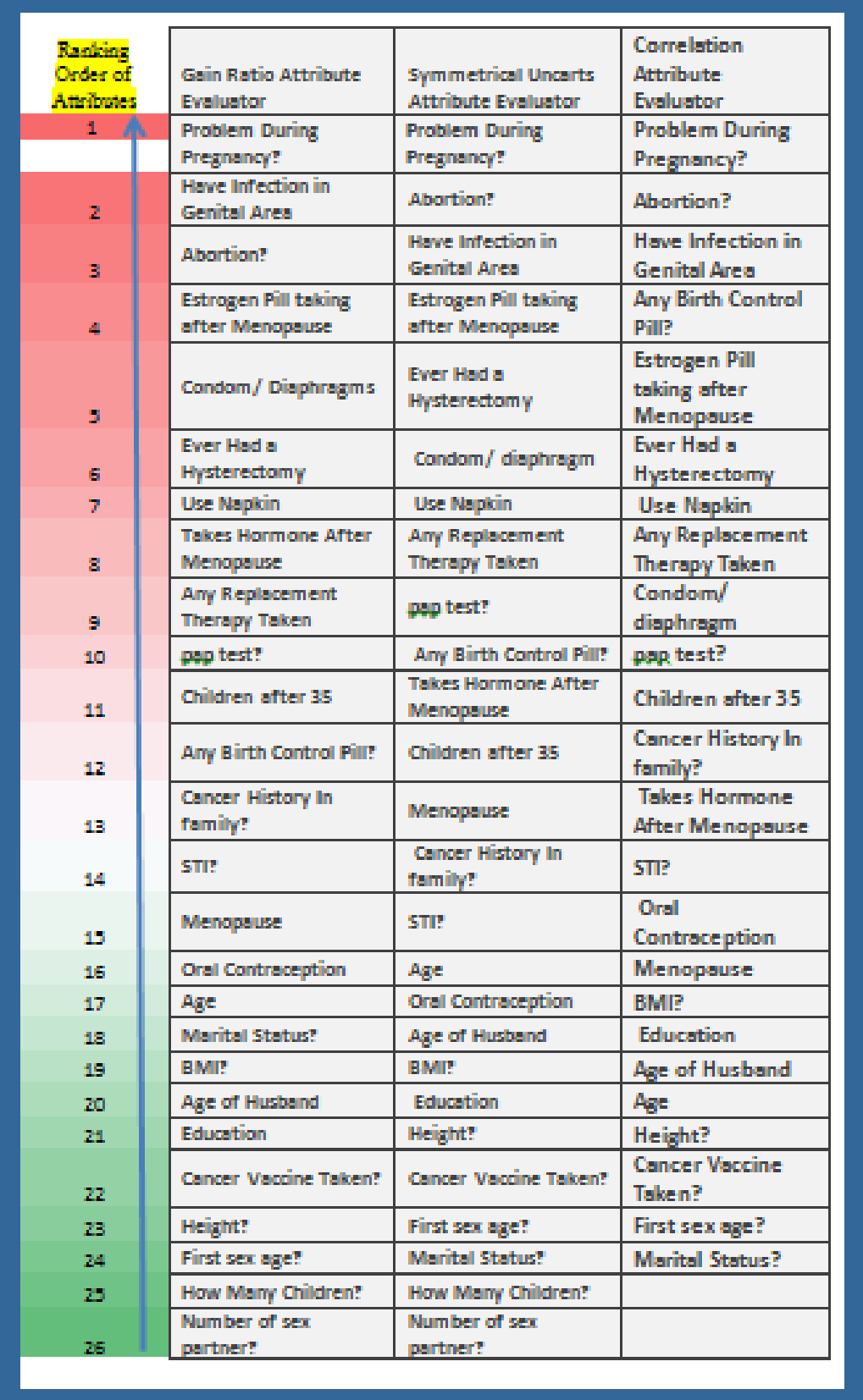
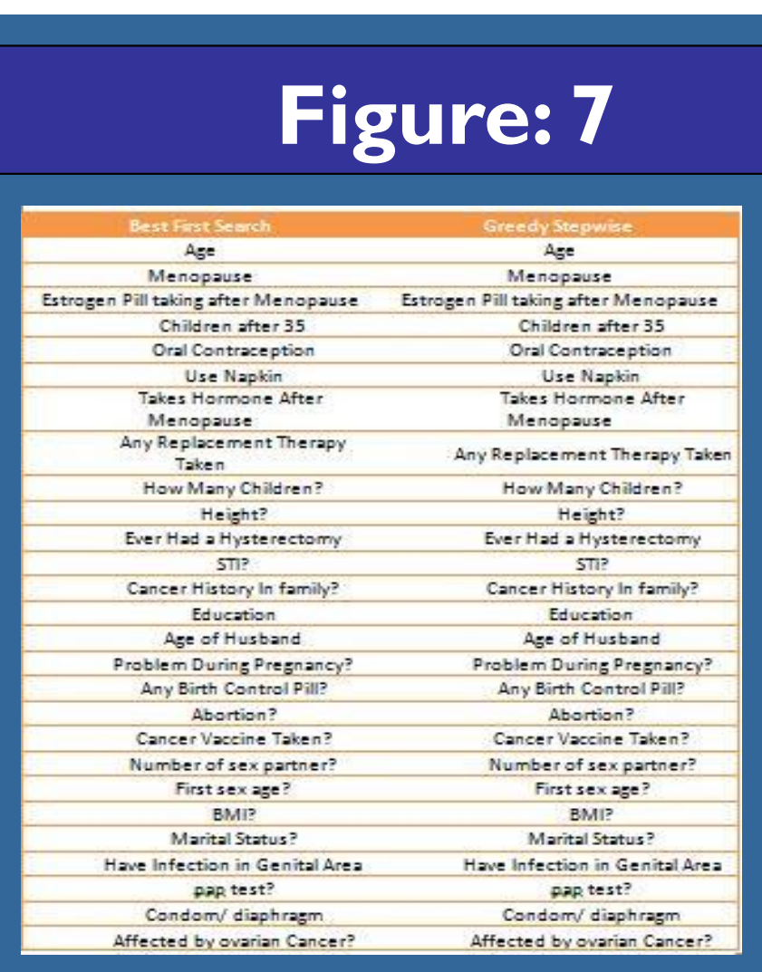


Figure: 7



Ranking Order of Attributes	Gain Ratio Attribute Evaluator	Symmetrical Uncerts Attribute Evaluator	Correlation Attribute Evaluator
1	Problem During Pregnancy?	Problem During Pregnancy?	Problem During Pregnancy?
2	Have Infection in Genital Area	Abortion?	Abortion?
3	Abortion?	Have Infection in Genital Area	Have Infection in Genital Area
4	Estrogen Pill taking after Menopause	Estrogen Pill taking after Menopause	Any Birth Control Pill?
5	Condom/ Diaphragms	Ever Had a Hysterectomy	Estrogen Pill taking after Menopause
6	Ever Had a Hysterectomy	Condom/ diaphragm	Ever Had a Hysterectomy
7	Use Napkin	Use Napkin	Use Napkin
8	Takes Hormone After Menopause	Any Replacement Therapy Taken	Any Replacement Therapy Taken
9	Any Replacement Therapy Taken	pap test?	Condom/ diaphragm
10	pap test?	Any Birth Control Pill?	pap test?
11	Children after 35	Takes Hormone After Menopause	Children after 35
12	Any Birth Control Pill?	Children after 35	Cancer History In family?
13	Cancer History In family?	Menopause	Takes Hormone After Menopause
14	STI?	Cancer History In family?	STI?
15	Menopause	STI?	Oral Contraception
16	Oral Contraception	Age	Menopause
17	Age	Oral Contraception	BMI?
18	Marital Status?	Age of Husband	Education
19	BMI?	BMI?	Age of Husband
20	Age of Husband	Education	Age
21	Education	Height?	Height?
22	Cancer Vaccine Taken?	Cancer Vaccine Taken?	Cancer Vaccine Taken?
23	Height?	First sex age?	First sex age?
24	First sex age?	Marital Status?	Marital Status?
25	How Many Children?	How Many Children?	
26	Number of sex partner?	Number of sex partner?	

Figure: 6



Best First Search	Greedy Stepwise
Age	Age
Menopause	Menopause
Estrogen Pill taking after Menopause	Estrogen Pill taking after Menopause
Children after 35	Children after 35
Oral Contraception	Oral Contraception
Use Napkin	Use Napkin
Takes Hormone After Menopause	Takes Hormone After Menopause
Any Replacement Therapy Taken	Any Replacement Therapy Taken
How Many Children?	How Many Children?
Height?	Height?
Ever Had a Hysterectomy	Ever Had a Hysterectomy
STI?	STI?
Cancer History In family?	Cancer History In family?
Education	Education
Age of Husband	Age of Husband
Problem During Pregnancy?	Problem During Pregnancy?
Any Birth Control Pill?	Any Birth Control Pill?
Abortion?	Abortion?
Cancer Vaccine Taken?	Cancer Vaccine Taken?
Number of sex partner?	Number of sex partner?
First sex age?	First sex age?
BMI?	BMI?
Marital Status?	Marital Status?
Have Infection in Genital Area	Have Infection in Genital Area
pap test?	pap test?
Condom/ diaphragm	Condom/ diaphragm
Affected by ovarian Cancer?	Affected by ovarian Cancer?

Figure: 3

DISCUSSION

It can be noted by the analysis that the possibility to be affected by cancer is much higher whose ages are above 46. The largest part of the cancer affected women's age was among 46 to 60. Being a developing country most of the women of Bangladesh are uneducated.[7] According to our analysis, about two third of the total cancer affected patients were uneducated. The fact is that the largest amount of cancer affected women has no idea about cancer. There are some strongly correlated factors like "whose first sex age was below 16 "and "whose number of children was 5 or more" which were observed by the analysis. Among all the factors STI and problems arising during pregnancy was found as highly significant. However, the majority of Bangladesh citizens cannot afford healthcare and do not have access to the complex care our patient received. Development of universal health care insurance must be part of the strategy in Bangladesh for complex care such as for ovarian cancer. [6]

The proposed method is implemented using java. The proposed method can efficiently and successfully predict the risk of ovarian cancer. According to our analysis, among the 400 ovarian cancer patients about most of them were married which was shown in data visualization chart. We found most of our ovarian cancer patients illiterate and poor rural people who are not conscious about their health due to lack of knowledge and cannot afford for the proper treatment after being diseased or diagnosed, so they have a greater risk of suffering from ovarian cancer death than that of educated and rich people. These findings emphasize the need to develop health education programs that enhance ovarian cancer knowledge among women who are in low socioeconomic groups. So, the government and NGOs should gear up for a population based counseling program.

CONCLUSIONS

In conclusion, as Bangladesh is a low incoming and population country, most of the women are not aware of deathful ovarian cancer disease because of lacking education. Moreover, a vast number of hearsay cramps the society nastily, women conceive disgrace of discuss on ovarian cancer with others. The researchers had worked on ovarian cancer and different tools and techniques has been updated day-by-day. In the paper the associative relation of factors has been detected with ovarian cancer and the possibility of preferences among the obtained factors has been estimated. The results can be used to increase awareness among women about different factors. It will be helpful for early prevention and better than cure. Moreover the research would depict to future researchers to find out some new era in the meantime to save the women from this atrocious curse. Otherwise one day the victims would by us or any relatives or anyone of the society.

ACKNOWLEDGEMENTS

The authors are grateful to all of the subjects who participated in this research.

REFERENCES

1. Pepa, Chiara Della et al. "Ovarian Cancer Standard of Care: Are There Real Alternatives?" Chinese Journal of Cancer 34.1 (2015): 17–27.
2. National Cancer Institute Available at: http://www.cancer.gov/types/ovarian
3. Centers for Disease Control and prevention available at: http://www.cdc.gov/cancer/ovarian
4. World health organization. Available: http://www.who.int/mediacentre/factsheets/fs297/en/
5. Pervin S, Islam F, Hall T and Goodman A. The Management of Ovarian Cancer in Bangladesh: A Report of a Long-Term Survivor. Austin J Obstet Gynecol. 2015; 2(4): 1047.
6. Hussain SM Comprehensive update on cancer scenario of Bangladesh, South Asian J Cancer.2013 Oct; 2(4):279-84.
7. Cancer Registry Report, 2005-2007; National Institute of Cancer Research and Hospital, Dhaka, Bangladesh; Published on December (2009).
8. Jiang Su and Harry Zhang (2006). A Fast Decision Tree Learning Algorithm. Copyright ° c 2006, American Association for Artificial Intelligence. Available at: www.aaai.org